A STUDY OF SAMPLING AND ESTIMATING PROCEDURES FOR CALIFORNIA CLING PEACHES

by

Ronald A. Wood

and

Fred B. Warren

Research and Development Branch
Standards and Research Division
Statistical Reporting Service

January 1972

Table of Contents

A STUDY OF SAMPLING AND ESTIMATING PROCEDURES FOR CALIFORNIA CLING PEACHES

by

Ronald A. Wood

and

Fred B. Warren

Introduction

This report presents the result of four years of experimentation aimed at improving the forecasting model for the California cling peach crop. Accurate forecasts are imperative for the economic stability of the industry. The present forecast helps determine production in conjunction with market demand analysis. High or low forecasts cause instability in the market mechanism; hence, the need for improvement of forecasting methods to keep economic stability at a maximum.

Present objective yield estimates of peach production are obtained from the expansion of counts from sample limbs (terminals). Sample limbs are selected by a random path method using probabilities proportional to size (PPS). [1] The two objectives of this study were to evaluate (1) the possible use of different variables, such as tree size and fruit counts from photographs, as covariates in double-sampling, and (2) several different methods of selecting sample limbs.

[1] For a detailed explanation of this method refer to Jessen, Raymond J., "Determining the Fruit Count on a Tree by Randomized Branch Sampling," Biometrics, March 1955, pp. 99-109.

This report includes two study periods. First, 1967-68 was a period of experimentation. During these two years many different sampling and estimation techniques were tried. The best procedures developed in 1967-68 were used in a pilot project in 1969-70 to test for operational efficiency and feasibility.

Summary

Findings 1967-68

Four sample limb selection procedures were tested in 1967 and 1968:

1. Multiple stage random path selection with equal probabilities at each stage until terminal limbs were selected.

2. Multiple stage random path selection with probabilities proportional to cross sectional area (CSA) at each stage.

3. Single stage random selection of sample units with equal probabilities to each possible sample unit.

4. Single stage random selection of sample units with probabilities proportional to CSA of possible sample units.

In both 1967 and 1968, single stage sampling with equal probability of selection for each sample unit had the minimum variance.

Coefficients of correlation were computed between each of four different variables--counts of fruit from color slides (photo counts), sum of primary cross section areas (CSA's), number of terminals and sum of terminal CSA's, and the total number of fruit on the same trees. The most consistent relationship was between photo counts and actual fruit counts where the r values were .855 and .742 in 1967 and 1968, respectively. Both correlations are significant at the .05 level. The number of terminal limbs vs. total number of fruit had r values of .520 and .708 in 1967 and 1968, respectively. These correlations are also significant at the .05 level. The sum of primary CSA's was significantly correlated with total fruit at the one percent level (α = .01) in 1967 but was below the five percent level (α = .05) in 1968. The sum of terminal CSA's was significantly correlated with total fruit at α = .05 in 1967 but not in 1968.

A series of tests relating to photography gave the following results:

1. The coefficient of correlation between photo counts and actual fruit was significant at α = .01 in 1967 but was significant at only α = .05 in 1968.

2. The regression coefficients for the simple linear regression of photo counts with actual fruit in 1967 and in 1968 were not significantly different.

3. Tests for significant differences in the number of fruit counted from photographs between sides of a tree, diagonals on one side of a tree, and quadrants within a diagonal were made. The tests showed no significant differences between diagonals within sides for either 1967 or 1968. The sides of a tree did not behave as nicely. There was a significant difference between the two sides photographed in 1967 but no significant difference in 1968.

Findings 1969-1970

The coefficients of correlation computed in 1969 for number of terminals and sum of primary CSA's with expanded limb counts were .609 and .659, respectively. Both of these figures are significant at the .01 level. The coefficient of correlation for photo counts with expanded limb counts (.168) was not significant at the .05 level.

In 1970 the coefficient of correlation for photo counts with expanded limb counts was .566. This value is significant at $\alpha = .01$. Neither the number of terminals nor the sum of CSA's of the primary limbs was significantly correlated with the expanded fruit counts at $\alpha = .05$.

An optimum allocation based on 1970 data, where the terminal limbs were selected to have a CSA of 0.8 to 2.0 square inches, showed that eight sample limbs should be selected from one tree per block.

The sum of CSA's of the primary branches is an economically feasible covariate in a double sampling model. The optimum allocation under double sampling would be to measure the primary CSA's on four trees and count peaches on sample limbs on one tree in each block. Based on the pilot results of 1969-70 and the techniques employed in this study, there is no economic advantage in using either photo counts or the number of terminals per tree as a covariate in a double sampling design.

Research 1967 and 1968

Photographic Procedures--Field and Office

The initial stage of operation was the selection of trees. In 1967 16 trees within one block were selected so that a wide variety of trunk CSA's were observed.

Eleven trees were photographed in June. The photographs were taken from opposite sides of the selected trees. One position was selected as a random compass direction from the tree. The second position was 180 degrees around the circle from the first (Figure 1). A vertical pole and a crossbar were used to divide each side of the tree into quadrants (Figure 2). Each quadrant was photographed separately.

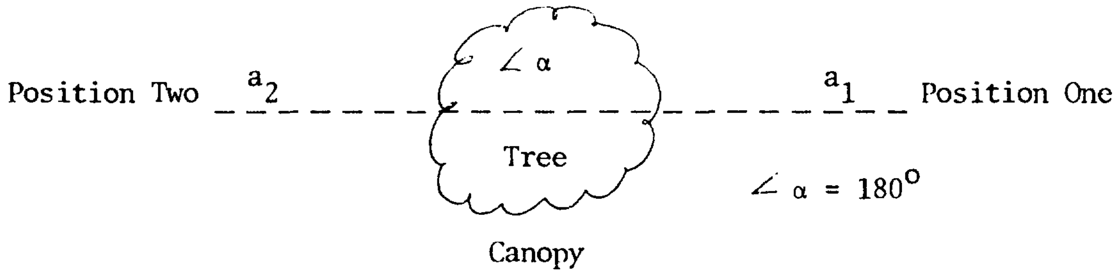
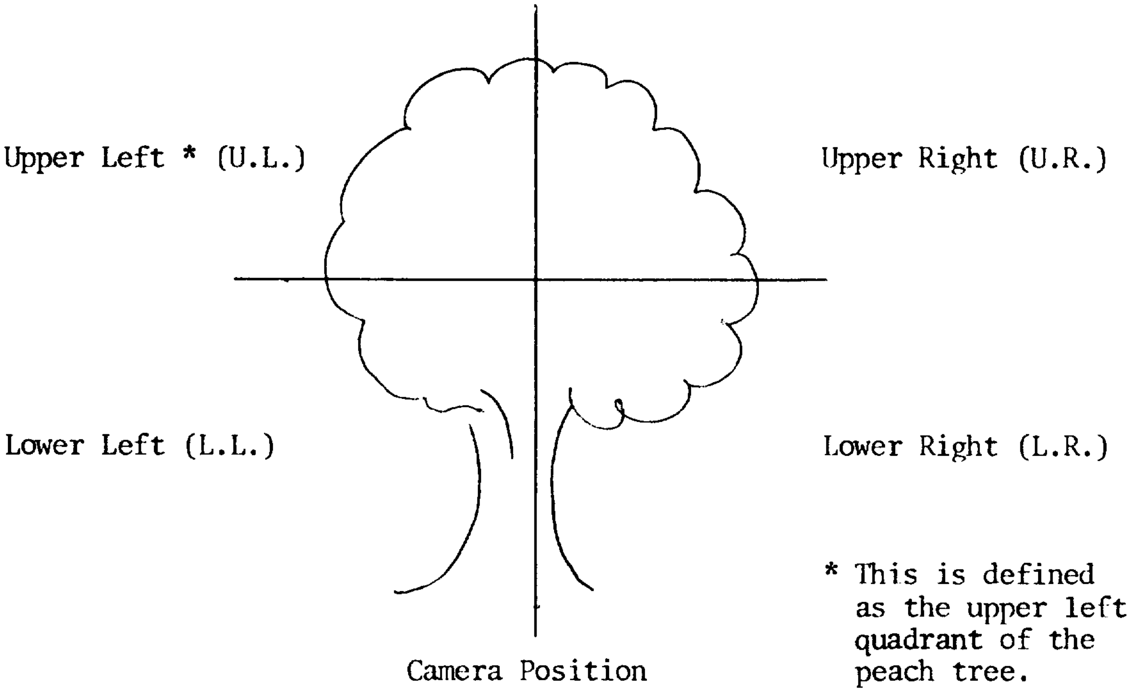


Figure 1.- Selection of camera location



* This is defined as the upper left quadrant of the peach tree.

Figure 2.- Division of tree by a vertical pole and cross bar for photography

The counting of peaches visible on the slides was done in the Research and Development Branch. Each slide was projected on a white screen divided into quadrants and counted by two persons. The first person would mark a dot for each peach observed. Then, the second would circle each dot he agreed with and place an X next to any other peach observed. Upon completion, differences were reconciled by consultation.

In 1968 the project was expanded to six blocks with four trees in each block. The selected trees were photographed in the same manner as in 1967.

In 1968 the slides were projected onto a screen divided into a grid pattern (Figure 3) and the consultation between interpreters was eliminated. The number of visible fruit in each cell of the grid was recorded on a photo record sheet which was a reduced image of the screen grid. Four counters, of approximately equal counting ability, were chosen from eight people available for counting. Each slide was counted by two different people. Each counter was paired with every other counter the same number of times.

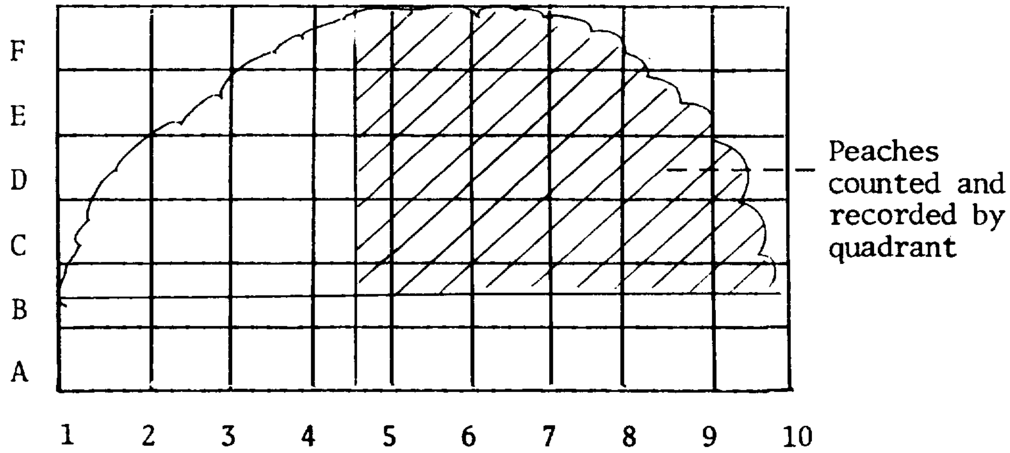


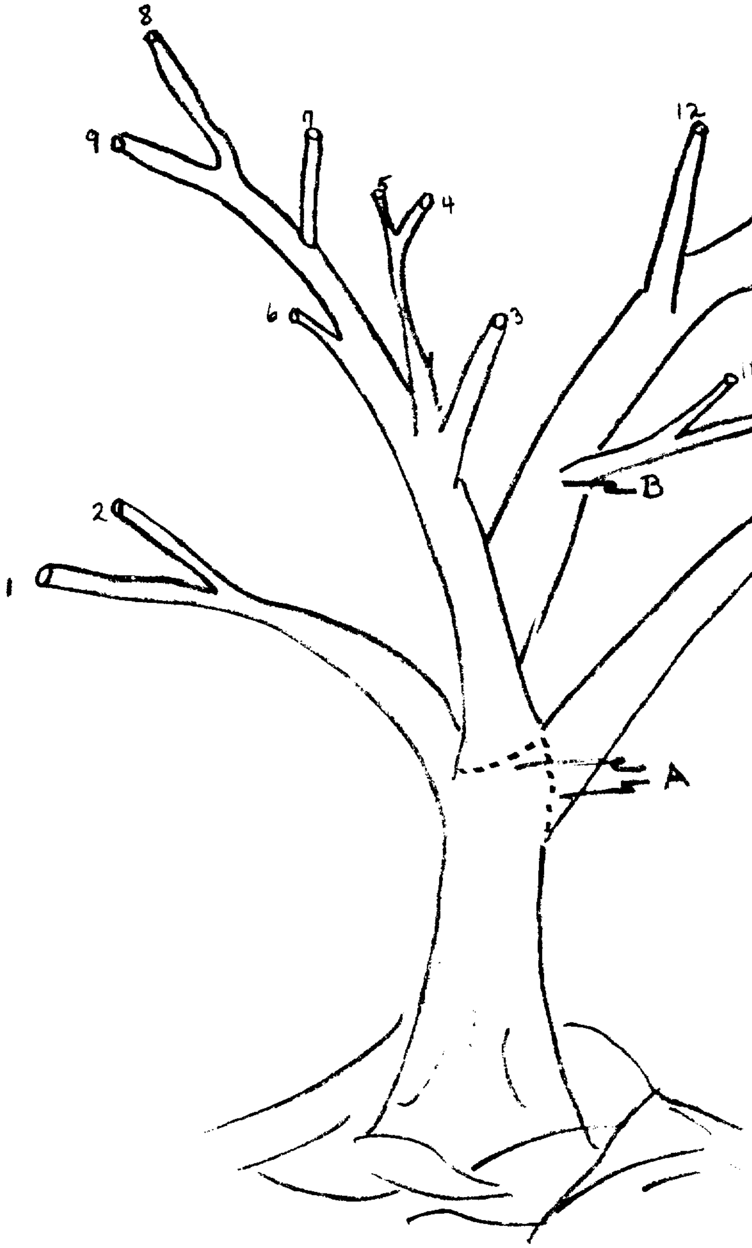
Figure 3.- Division of photo projection screen into cells for counting purposes. One quadrant of a California peach tree is simulated on the grid above.

Tree Measuring and Fruit Counting Procedures

In 1967 the 16 sample trees were completely counted as follows. Field sketches were made of each tree (Figure 4). Each limb was divided into terminal and nonterminal (path) branches. 2/ (A nonterminal branch is one which has a CSA greater than 2.0 square inches and which has at least two terminal or nonterminal branches emerging from it.) CSA measurements were taken at the base of all primaries, all nonterminal limbs, and all terminals. Each terminal was numbered and every nonterminal (path) section was labeled with a letter. Counts of peaches on these trees were then identified by their location in the tree, either path section or terminal.

In 1968, stereo photographs of the trees were taken during the winter (dormant) season. Sketches (maps) of the sample trees were prepared in the office from these photographs before the fieldwork started. The original plan was to count fruit on all trees. However, not enough time was available to complete all counts. In five blocks, two trees were not fully counted and in the sixth block none were fully counted. Where trees were not fully counted, fruit on randomly selected sample limbs were counted and the counts expanded to an estimated total for the tree. The limbs were selected from the sketches made from the stereo slides.

2/ See Appendix I for definitions.



CSA Measurements are taken at base of

A - Primaries

B - Nonterminal secondaries

C - Terminals

Figure 4.- Sample field sketch showing division of tree into primaries, nonterminal secondaries, and terminal branches, and location of points where CSA measurements are taken.

Tree Fruit Count Analysis

The 1967-68 total tree counts served a twofold purpose. First, they were used in finding the best linear unbiased estimator (BLUE) for the number of fruit per trees for California cling peach trees. Counts of fruit for the total tree eliminated the problem of sampling errors in the dependent variable found when direct expansion of limb counts are used for **total** number of fruit per tree estimators. Although total fruit counts are more time consuming, the elimination of sampling error in the dependent variable provided a sounder foundation for further research.

The total fruit counts for trees studied were used to simulate four different sampling and estimating procedures. These procedures were:

1. Single stage sampling with equal probabilities.
2. Random path (multiple stage) with equal probabilities.
3. Single stage sampling with probabilities proportionate to size.
4. Random path with probabilities proportionate to size.

For each tree, which had been totally counted, all possible combinations of sample limbs could be used for making direct expansion estimates for total fruit on the tree. The variance of the estimates from these four sampling procedures were compared to determine which gave the minimum variance. The computational procedure for the variances follow: 3/

3/ Ibid., pp. 104-105.

I. The expanded number of fruit per tree based on individual limbs:

$$\hat{X}_{ij} = X_{ij}/P_{ij}$$

Where: $\hat{X}_{ij}$ = estimated number of fruit on the ith tree computed from the jth sample limb.

X_{ij} = number of peaches on the jth sample limb in the ith tree.

P_{ij} = the probability of selecting the jth sample limb in the ith tree.

II. The variance of the estimate for a tree is:

$$V(X) = \sum_{j=1}^{n} P_{ij} \ (X_{ij}-X_i)^2$$

Where: X_i = the actual count of peaches of the ith tree.

The within tree variances for 1967-68 are summarized in Table 1. The magnitude of the variances for the two years vary because of the differences in fruit counts between years.

This data indicates that selecting terminal limbs with equal probabilities as a single stage sample would produce estimates with slightly lower variances than the other methods considered. Other studies 4/ have shown that single state sampling with probabilities of selection proportionate to the CSA of terminal limbs had lowest variances. The improved efficiency of single stage equal probability sampling in this case may have resulted from:

4/ Huddleston, Harold, "The Use of Photography in Sampling for Number of Fruit Per Tree," Agri. Econ. Res., July 1971, Vol. 23, No. 3.

(a) the heavy pruning practiced on cling peaches negating the effect of the size (CSA) of the terminal, and (b) the comparatively close restriction on limb size allowed for terminal limbs.

Under an equal probability model each terminal was given an equal chance of selection from the tree. The efficiency of this model depends on the ability to define each terminal within a narrow CSA range; e.g., a branch with a CSA between 0.8 - 2.0 inches.

Table 1.- Average within tree variance by four different sampling procedures, California cling peaches, 1967-68

Estimating procedure	1967	1968
Multiple-stage random path		
Equal probability..................	207,532	3,451,222
Probability proportional to size....	89,142	1,101,935
Single-stage random selection of terminals		
Equal probability..................	76,538	915,946
Probability proportional to size....	110,827	961,319

Correlation and Regression Analyses

In 1967-68 simple regression and correlation coefficients were computed between the number of peaches counted on each tree and:

(1) The number of peaches counted on photographs (photo counts)

(2) Sum of primary limb CSA's

(3) Number of terminal limbs

(4) Sum of terminal limb CSA's

These tests were used to determine if any of these variables could be employed as supplementary covariates in a regression type estimation model. Results of these tests are in Table 2.

Table 2.- Correlation and regression coefficients for cling peaches, California, 1967-68

Year and parameter	Average number of peaches per tree	Total peach count vs.			
		Photo count	Sum of primary CSA's	Number of terminals	Sum or terminal CSA's
1967.......	474				
r.........		.855	.685	.520	.562
b.........		2.802	12.233	20.327	10.673
$\bar{X}$.........		162.0	38.4	20.0	36.8
1968.......	1570				
r.........		.742	.443	.708	.268
b.........		3.5	14.32	75.69	10.09
$\bar{X}$.........		344.5	80.39	23.8	70.5

A t-test was used to test the significance of the relationships mentioned above.

Let: $t = r\sqrt{n-2} \;/\; \sqrt{1-r^2}$

and if $t_{.01}<t$ then $B \neq 0$ (B = true regression coefficient)

if $t_{.01}>t>t_{.05}$ then B is marginally $\neq 0$

if $t_{.05}>t$ then assume $B = 0$ where B = the regreesion coefficient of the population.

Table 3.- Test of significance of regression coefficients for California cling peaches, 1967

Fruit vs.	Degrees of freedom	t	$t_{.05}$	$t_{.01}$
Photo count............	9	4.94	2.26	3.25
Sum of primary CSA's...	14	4.56	2.14	2.97
Number of terminals....	14	2.27	2.14	2.97
Sum of terminal CSA's..	14	2.54	2.14	2.97

Table 4.- Test of significance of regression coefficients for California cling peaches, 1968

Fruit vs.	Degrees of freedom	t	$t_{.05}$	$t_{.01}$
Photo count............	8	3.14	2.31	3.35
Sum of primary CSA's...	8	1.67	2.31	3.35
Number of terminals....	8	2.84	2.31	3.35
Sum of terminal CSA's..	8	1.04	2.31	3.35

The only cases for which $B \neq 0$ were for the photo counts and sum or primary CSA's in 1967. The t values for number of terminals and photo counts in 1968 were between the $t_{.01}$ and $t_{.05}$ level. The test results indicate that further research should be conducted on relationship of total fruit count and:

(1) Photo counts

(2) Number of terminals

(3) Sum of primary CSA

Analysis of Photo Counts

The consistency of the b's over the two years, 1967 and 1968, will help to evaluate the photo counting procedure. The question is, "What effect did the change in photo counting methodology have on the final result?"

The test for equality between coefficients in two relations gives the desired results. 5/ In this case the two simple least square regressions are:

$Y_1 = B_1X_1 + e_1$ e has a normal $(0, \sigma^2)$ distribution

$Y_2 = B_2X_2 + e_2$

where Y_1 or Y_2 = actual fruit for respective year

X_1 or X_2 = photo count for respective year

To test if $B_1 = B_2 = B$, use the F test

$$F = (Q_3/k)/(Q_2/m + n - 2k)$$

with (k, m + n - 2k) d.f.

where k = number of variables

m = number of observations in 1967

n = number of observations in 1968

Q_1 = sum of squared residuals for the pooled observations

$= (\Sigma y^2)\ (1 - r^2)$

$= (11{,}543{,}742.6)\ (1 - .735389) = 3{,}054{,}586.3$

5/ Johnston, J., Econometric Methods, McGraw-Hill Book Company, Inc., N.Y., 1960, pp. 136-138.

Q_2 = sum of square of each set added

$= (\Sigma y_1^2)\ (1 - r_1^2) + (\Sigma y_2^2)\ (1 - r_2^2)$

$= (703{,}828.7)\ (1 - .73101) + (4{,}913{,}082.5)\ (1 - .55124)$

$= 2{,}367{,}605.7$

$Q_3 = Q_1 - Q_2 = 686{,}975.6$

Then $F = 2.47$

$F_{.05} = 3.59 > 2.47 = F$

$F_{.01} = 6.11 > 2.47 = F$

$\therefore B_1 = B_2 = B$

Since the F-value is not significant, one must accept the hypothesis that there is no significant difference between the regression coefficient in 1968 and that of 1967. Therefore, we can assume that the change in photo counting procedures had little or no effect upon the regression coefficients.

Further analysis of the photo counts was made in an attempt to reduce the number of counts required. A nested analysis of variance was calculated for the photo counts. In 1967, tree, diagonal and quadrant (slide) were the three levels of sampling since only one block was sampled, but in 1968, block, tree, diagonal and quadrant constituted four levels of sampling (Tables 5 and 6).

The data contained in Table 5 represents one block of comparatively young trees. Therefore, this block would represent a small portion of the universe of California cling peach trees. The information found in Table 6 pertains to a larger portion of the universe with respect to age and number of trees. Therefore, the information in Table 6 will be given more weight in this analysis.

Table 5.- Analysis of variance of photo counts, cling peaches, California, 1967

Source	Degrees of freedom	Mean square	F	$F_{.05}$	$F_{.01}$
Trees.........	10	679.26	1.97	2.85	4.54
Side..........	11	344.24	5.51	2.27	3.20
Diagonal......	22	62.46	.67	1.79	2.29
Quadrant......	44	93.37			

Table 6.- Analysis of variance of photo counts, cling peaches, California, 1968

Source	Degrees of freedom	Mean square	F	$F_{.05}$	$F_{.01}$
Block.........	5	6,803.228	8.71	2.77	4.25
Tree..........	18	780.962	2.80	2.06	2.80
Side..........	24	279.027	.97	1.74	2.20
Diagonal......	48	286.676	.89	1.48	1.73
Slide.........	96	321.616			

In 1967 and 1968, differences between diagonals were found to be insignificant at the .05 levels. This implies that under an optimum allocation, it would probably be necessary to count photographs of only one diagonal of each side of a tree.

Markedly different results were obtained with respect to differences between sides of trees for the two years. There was a highly significant difference in 1967 but no significant difference (at $\alpha = .05$) in 1968.

The two years are again in contrast with respect to significant differences between trees. In 1967 there is no significant difference while 1968 shows a significant difference at $\alpha = .01$. The comparatively young trees used in 1967 were more uniform than were the six blocks of different ages used in 1968.

Data between blocks was only compiled for 1968. The information shows a significant difference between the blocks at $\alpha = .01$. However, further study is necessary to reach a final conclusion.

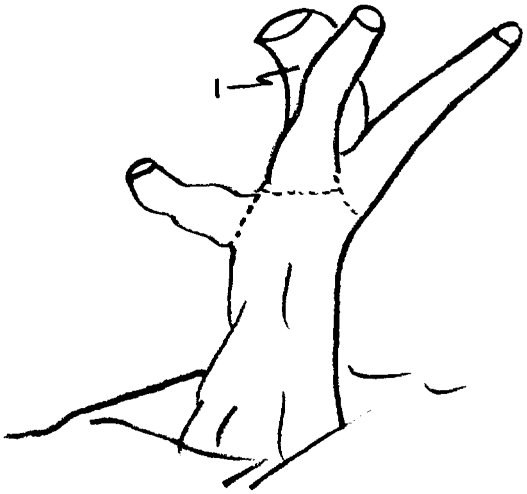
Research 1969 and 1970

Sample Tree Mapping Procedures

In 1969, 22 blocks in the Yuba City - Marysville area of California were randomly selected for observation. A systematic sample of three trees was chosen in each of these 22 blocks. Stereo pictures of these trees were taken in the spring before they leafed out. Measurements of CSA's for the primary limbs of these trees were made at the same time.

Photo enlargements of pictures of the bare trees were used to divide (map) the tree into terminal units of approximately equal size (CSA) and path sections. The path sections were then combined with designated terminals to form sampling units.

The trees were mapped starting at the trunk. The first divisions going from the trunk are the primaries (Figure 5). Each primary is completely mapped, one at a time. This involved marking every major split until a terminal limb was reached (Figure 6). The goal was to select terminal limbs which would have



Primary -----------

1.....This primary would have been defined from a second Itek print since it was not clearly distinguishable. At least two shots are taken of each tree from different angles.

Figure 5.- Primary limbs

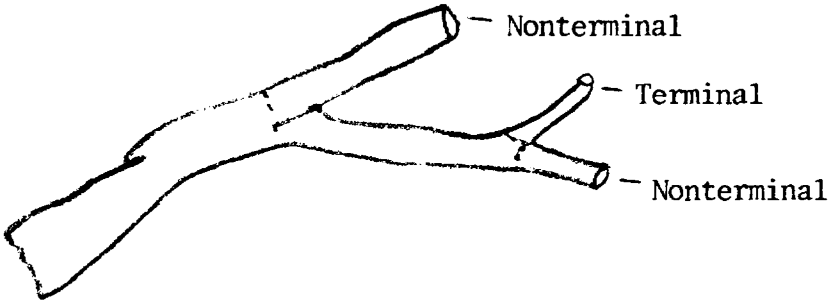
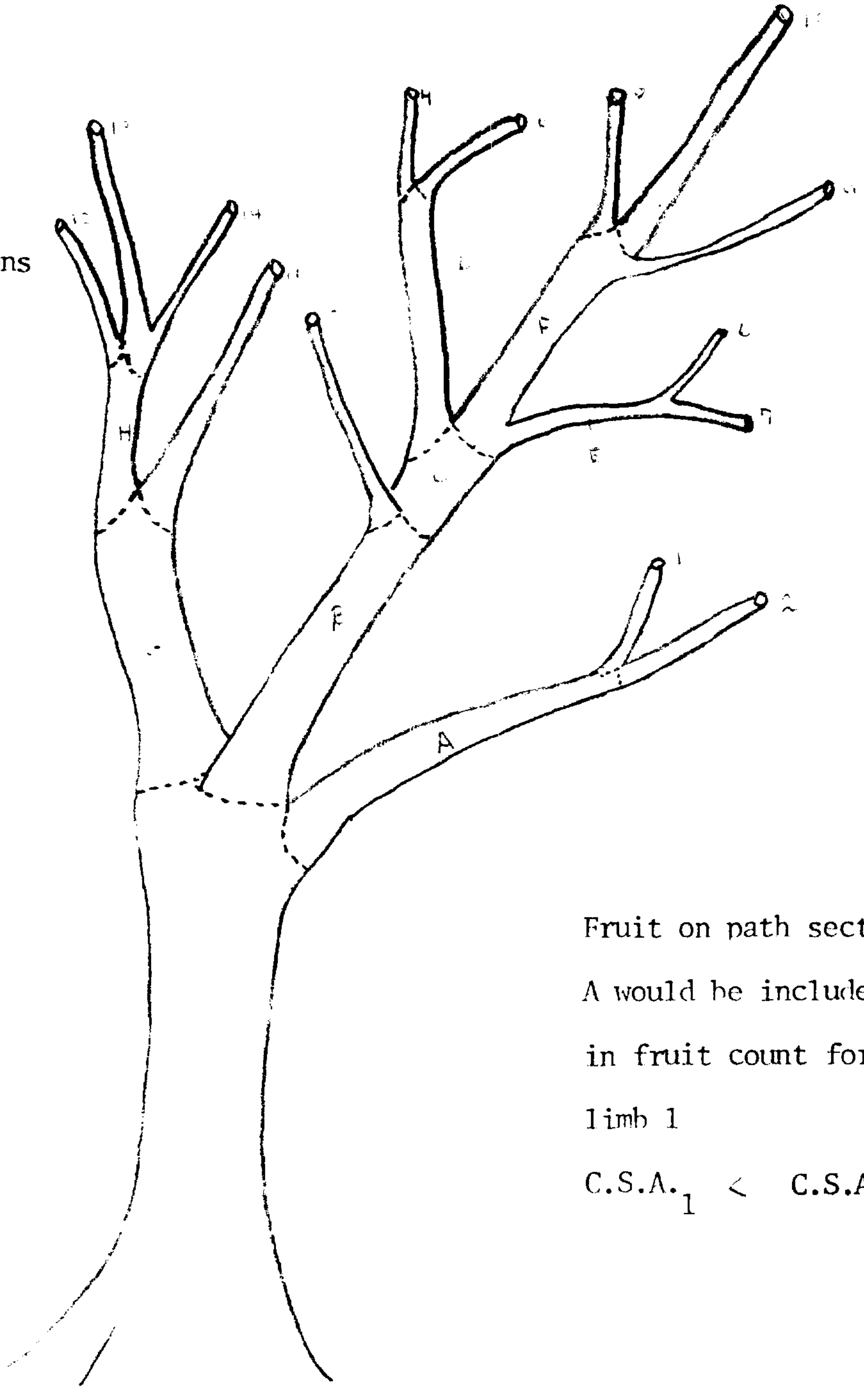


Figure 6.- Marking of major split

a cross-sectional area of 0.6 to 2.0 square inches (a diameter of 1/16 to 1/8 inch on the enlarged photograph). If a limb appeared to be too large but could not be divided into at least two terminal size limbs the whole limb was considered a terminal. When limbs appeared to be too small several nearby limbs were combined into one "terminal." After all terminals on the primary were identified, they were numbered, starting with the terminal closest to the trunk (Figure 7). When two or more limbs branched from the same location the smallest limb was given the lowest number. Each path section was assigned a letter. In counting, each path section was assigned to the lowest numbered terminal limb which was above it. When two or more terminals emerged from the same point, the path section was assigned to the terminal with smallest number. For example, in Figure 7, path section A would be assigned to terminal 1, path section B to terminal 3, path section C and D to terminal 4, etc. After the first primary had been mapped, the next primary to the left was mapped in the same manner, and so on around the tree. Lettering and numbering was continuous from one primary to the next. After the trees were mapped, a systematic sample of two clusters of three sample units each (terminal limbs) was selected from each tree.

Sample Limb Selection

The first step of sample terminal selection for each tree was to compute a sampling interval by dividing the total number of sampling units (terminals) by the appropriate number of limbs to be counted, two in 1969 and four in 1970. Then a random number between .1 and the sampling interval was drawn. This



Letters--
Path Sections

Numbers--
Terminals

Fruit on path section
A would be included
in fruit count for
limb 1

C.S.A.$_1$ < C.S.A.$_2$

Figure 7.--Tree mapping, terminal-path section number identified the first limb in the first sample cluster. Then the sampling interval was added to the random start to get the number of the first sample limb in the second cluster. This process was continued as many times as needed. The clusters of sample limbs were composed of the selected limb and the next two higher numbered terminals. Since the terminals were numbered continuously it was possible for the cluster to be located on two different primaries.

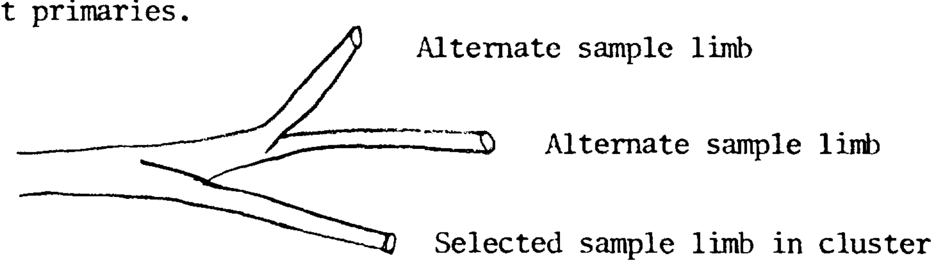
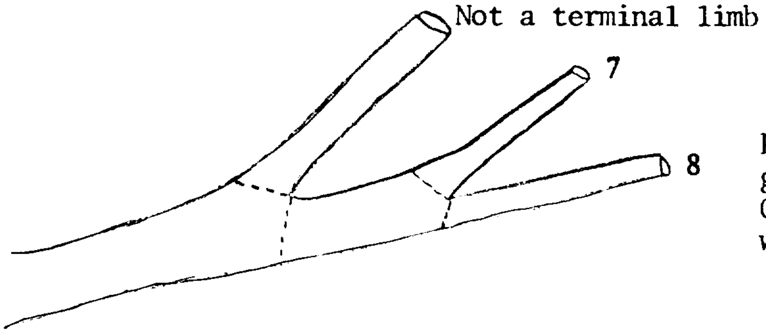


Figure 8.- Cluster selection

Limb Counting Procedure

All terminal limbs to be counted were marked in the field with plastic ribbon (flagging tape). Terminals were designated with blue flagging tape and the corresponding path sections by red (Figure 9). Each sample unit (terminal and corresponding path sections) in the cluster was checked for the presence of fruit. The first sample unit in the cluster to have any peaches was used as the count unit. In some cases, no fruit were present on any of the three sample units. The number of sample units in the sample cluster which had any fruit was recorded. Finally, fruit on the selected terminal and its corresponding path sections were counted and recorded.



Path sections A and C go with terminal 7. Counts for A, C and 7 were recorded separately.

Figure 9.- Path count

Field and Office Photography Procedures

The procedures for field photography in 1969 and 1970 were the same as described for 1967. The only change from 1968 in photo counting was to incorporate an adjustment factor for interpreter effects into the counting scheme. One-third of the June slides were recounted such that each interpreter was recounted an equal number of times by each of the other interpreters. This system avoids the time consuming total recount. Procedures for computing the adjustment factors are given in a later section.

Coefficients of Correlation

Photo counts

The correlation between the 1969 photo counts and the expanded tree counts was much lower than in 1967 and 1968. The correlation between expanded limb counts and photo counts (adjusted for interpreters) was .14. This is not significant at $\alpha = .05$. A possible explanation for the low correlations in 1969 was the poor quality of photography relative to 1967 and 1968. Several slides which were counted were underexposed.

The correlation between photo counts vs. limb counts improved in 1970. The correlation coefficient was .512 which is significant at $\alpha = .01$.

These r values represent an underestimate of the true correlation between the number of peaches per tree and the number counted from photographs since the variance of the expanded limb counts used in computing r includes a within tree variance component. Computations of corrected r values are found in Appendix II. In 1969 the corrected r = .168. This is not significant at $\alpha = .05$. The 1970 corrected value equals .566, which is significant at $\alpha = .01$.

Terminals per tree

The correlation between the expanded number of fruit from limb counts and the number of terminals in 1969 was .486. This is significant at $\alpha = .01$. The correlations between number of terminals vs. expanded limb counts in 1970 was .306. This is not significant at $\alpha = .05$. These values are not corrected for sampling error in the expanded limb counts.

The corrected r value of 1969 was .609 which is significant at the .01 level. The corrected r value for 1970 was .344 which still is not significant at $\alpha = .05$. The low correlation in 1970 may be because the count of number of terminals was a year old. The high correlation in 1969 gives support to further analysis in the optimum allocation section.

Sum of primary CSA's

In 1969 the uncorrected coefficient of correlation between sum or primary CSA and expanded limb counts for fruit numbers was .53. The corrected correlation was .636. Both the corrected and uncorrected coefficients of correlation are significant at α = .01.

In 1970 the uncorrected r value of .33 between sum of primary CSA's and expanded limb counts for fruit numbers is not significant at α = .05. The corrected r value of .366 is not significant at α = .05. Again, this drop in correlation may be due to the CSA measurements being one year old.

The 1969 results point to sum of primary CSA's as the covariate best suited for use with expanded limb counts. The feasibility of its use is discussed in the section on optimum allocation.

Direct Expansion of Sample Limb Counts

In 1969 and 1970, the estimated number of peaches per tree was obtained from direct expansion of sample limb counts. Field procedures in the two years differed only in the numbers of trees and limbs per tree selected (Table 7). The method for expanding limb counts was the same for both years.

Table 7.- Sample design within block, cling peaches, California, 1969-70

Year	Number of trees	Number of sample limb clusters per tree
1969	3	2
1970	2	4

The basic count unit (terminal plus associated path) was part of a cluster of three sample units. The number of units with any fruit in each cluster was recorded along with the number of peaches counted on the selected count unit. The ratio of limbs bearing fruit to the total number of limbs in the cluster was used as a correction factor in estimating the number of peaches on a tree.

Let X_{ij} = the number of peaches on the jth limb in the ith tree

$\hat{X}_{ij}$ = the estimated number of peaches on the ith tree using the jth limb.

$\hat{\bar{X}}_i$ = the mean of the direct expansions of number of peaches on the ith tree.

R_{ij} = the ratio of limbs containing peaches to total number of limbs in the jth cluster (jth limb found in the jth cluster) of the ith tree.

E_i = reciprocal of the probability of the jth terminal limb occurring in the ith tree.

Where $\hat{X}_{ij} = R_{ij} X_{ij} E_i$

$$\hat{X}_{i.} = \sum_{j=1} \hat{X}_{ij}/n$$

Where n = number of sample limbs in the ith tree.

$\hat{X}_{i.}$ then is the estimated peach count on the ith tree.

An analysis of variance of the direct expansion estimates under study indicates there is no significant difference between trees within blocks (Table 8).

Table 8.- Nested analysis of variance of direct expansion estimates, cling peaches, California, 1969-70

Source	1969			1970		
	Degrees of freedom	M.S.	F 1/	Degrees of freedom	M.S.	F 2/
Blocks.....	21	2,229,279	3.419	17	1,074,960	1.430
Trees......	44	652,099	0.796	3/ 16	751,444	1.075
Limbs......	66	819,162		102	689,891	

1/ $F_{.05} = 1.56$.
2/ $F_{.05} = 1.74$.
3/ In two blocks one tree was not counted.

With no significant difference between trees within blocks the logical conclusion is to sample fewer trees within a block and more limbs per tree. The section on optimum allocation will focus on this point.

There is one further point for discussion in relation to the use of direct expansion estimates as the dependent variable in double sampling. Since only an estimate of the peaches per tree is used, the variance of the estimated number of peaches per tree will include a within tree error component (measurement error). Therefore, the variance of the estimated peaches per tree will be larger than the variance of the actual numbers. That is, if $\hat{X}_i = X_i + e_i$, where the $e \rightarrow N(0, \sigma_e^2)$ comes from sampling inside the tree, the sampling variance of X will be $S_{\hat{x}}^2 = S_x^2 + S_e^2$.

Whenever variance (X) is an overestimate of the true between tree variance, the computed r will underestimate the true correlation.

Let $r = \dfrac{\text{Cov}(XY)}{(\text{Var }(X))^{1/2}\ (\text{Var }(Y))^{1/2}}$

$y_{i.}$ = photo count, number of terminals, or sum of primary CSA

$x_{i.}$ = estimated number of fruit for the ith tree

Correction for Sampling Error in Estimated Tree Totals

In 1969 and 1970, the tree totals used included within tree sampling errors. These errors can be estimated and any downward bias in r values made trivial.

Defined are the following relationships.

(1) $r^2 = (Sxy)^2/Sx^2Sy^2$ where r^2 is without measurement error

(2) $r_1{}^2 = (Sx_1y)^2/Sx_1{}^2Sy^2$ where $r_1{}^2$ is with measurement error

where $X_1 = S + e$, and X and e are independent

Let (3) $Sx^2 = Sx_1{}^2 - Se^2$

Find the difference between r^2 and $r_1{}^2$

(4) $r^2 - r_1{}^2 = ((Sxy)^2/Sx^2Sy^2) - ((Sxy)^2/Sx_1{}^2Sy^2)$

Substituting (3) and (4) it can be shown that

(5) $r^2 - r_1{}^2 = r_1{}^2\ (Se^2/Sx_1{}^2 - Se^2)$

Let $\lambda = Se^2/(Sx_1{}^2 - Se^2)$

Simplifying, (5) produces

(6) $r^2 = r_1{}^2\ (1 + \lambda)$, and

(7) $r = r_1\ (1 + \lambda)^{1/2}$

Upon inspection of λ, Se^2 (the measurement error) is unknown and must be estimated. The sampling error of the California cling peach estimates arose from the within tree variance component of the estimated between tree variance. In both 1969 and 1970, estimates of the within tree variance component are available. These values were divided by the number of sample limbs used to estimate the tree total. The tree estimates were based on two limbs in 1969 and four limbs in 1970. This information leads to the following estimates of λ : λ = .428 for 1969*

λ .235 for 1970*

(*See Appendix II for computation.)

Table 9.- Coefficients of correlation of fruit counts with selected covariates, California peaches, 1969-1970

Variables correlated with direct expansion of fruit on sample limbs	Coefficient of correlation			
	1969		1970	
	Uncorrected	Corrected	Uncorrected	Corrected
Adjusted photo counts...	.14	.17	.51	.566
Number of terminals.....	.49	.61	.31	.344
Sum of primary CSA's....	.53	.66	.33	.366

Analysis of Photo Counts

In 1969, one diagonal of one side of each peach tree was photo counted. The decision to observe only one side and one diagonal of one side of the tree was based on the analysis of variance of photo counts in 1968 (Table 6). This analysis showed no significant difference between sid or between diagonals within a side of a tree. Since fewer trees were photographed in 1970, counts were made from photos of both diagonals of one side of each tree. A sample of slides was recounted each year. The recounts were used to estimate the differences between counters. These differences were then used to adjust the photo counts. For analysis, the balanced incomplete block model is appropriate. 7/ Only minor changes were made. Instead of using the additive treatment constant (T) computed for each photo interpreter, a multiplier was computed (T') (Table 10).

$$T_i' = \frac{U_i - T_i}{U_i}$$

Where T_i = additive treatment constant of the ith interpreter

T_i' = multiplier treatment constant of the ith interpreter

U_i = mean of all photo counts made by the ith interpreter

Thus Y_{ijk} = adjusted photo count

$$Y_{ijk}' = T_i' (Y_{ijk})$$

i = interpreter

j = slide

k = tree

7/ Graybill, Franklin A., An Introduction to Linear Statistical Models, McGraw Hill Book Company, Inc., N.Y., 1961, Vol. I, pp. 308-311.

Table 10.- Photo count adjustment coefficients 1/, cling peaches, California, 1969-70

Year	Counter			
	1	2	3	4
1969.........	.851	1.149		
1970.........	.809	.766	1.971	1.052

1/ A value less than one means that the interpreter on the average counted higher than other interpreters, vice versa for greater than one.

Cost Analysis

Cost is a major factor in determining the operational feasibility of the preceding research. The following tables show time in field, man-hours spent, cost for these men, mileage cost, and film cost.

Table 11.- Time study, average of the difference between starting and finishing times on field forms

Item	Average time per block minutes	Total trees	Total blocks	Sample trees per block
Initial tree selection - 1969......	71.5	348	24	3
Bare tree photography - 1969.......	35.6	72	24	3
Photography for photo counts - 1970	23.5	34	17	2

Table 12.- Average time required to count fruit on sample limbs, California peaches, 1969-70

Item	Average time per block minutes	Number of limbs per tree	Number of trees per block
Fruit counts on sample units - 1969	38.9	2	3
Fruit counts on sample units - 1970	60.4	4	2

The information in Tables 11 and 12 are actual times taken from field forms. Travel and office times were not recorded for summarization. The initial tree selection and bare tree photography in 1969 applied to both 1969 and 1970. This information will be converted to cost per tree to compute optimum sampling design.

Under operational conditions the peach objective yield data collection crews would consist of two-man teams. For the initial tree selection and bare tree photography the cost would be based on the work of a GS-9 and a GS-5. This portion of data collection would be repeated every four years so the costs should be amortized over a four year period. The values in parenthesis in Tables 13, 14, and 15 are the amortized costs.

During the regular peach objective yield, the cost would be based on the work of two GS-3's. The mileage cost is computed on the basis of $.10 per mile. The time and mileage data used to compute the cost were obtained from field experience. The average time for travel between trees was derived from other data collected during the research project.

Table 13.- Estimated travel costs, California cling peaches, 1969-70

Travel time	Man-hours		Mileage	
	Number	Cost	Number	Cost
	Hours	Dollars	Miles	Dollars
Costs per block				
Block location and tree selection	1.00	3.96 (.99) 1/	20	2.00 (.50)
Annual survey	1.00	2.15	20	
Costs per tree				
Tree selection and bare tree photography	.08	.33 (.08)		
Annual survey	.08	.18		

1/ All figures in parentheses are costs amortized over a four year period.

Table 14.- Average cost and average time for photo counts, California cling peaches, 1970

Item	Man-hours per tree	Cost per tree	Film cost
		Dollars	Dollars
Photography			
GS-9	.20	.95	--
GS-5	.20	.63	--
Photo counts (one diagonal)			
GS-4 (office)	.17	.53	--
Tree selection			
GS-9	.09	.40 (.10) 1/	--
GS-5	.09	.28 (.07)	--
Film	--	--	.30

1/ All figures in parentheses are costs amortized over a four year period.

Table 15.- Average cost and average time per tree to acquire limb counts, California cling peaches, 1969-70

Item	Man-hours per tree	Cost per tree	Film cost
		Dollars	Dollars
Tree selection			
GS-9	.40	1.90 (.48) 1/	--
GS-5	.40	1.26 (.32)	--
Bare tree photography			
GS-9	.20	.96 (.24)	--
GS-5	.20	.64 (.16)	--
Bare tree mapping and limb selection (office)			
GS-4	1.50	4.64 (1.16)	--
Itek prints	--	--	1.41 (.37)
Film	--	--	.68 (.17)
Limb count (1969) - two clusters per tree			
GS-3	.65	1.63	--
GS-3	.65	1.63	--
Limb count (1970) - four clusters per tree			
GS-3	1.01	2.17	--
GS-3	1.01	2.17	--

1/ All figures in parentheses are costs amortized over a four year period.

The average times applied to the above were taken from field form time information in Table 12. These times apply to two-man teams. The only office time (estimated) in Table 15 is that of mapping photographs of bare trees and limb selection by a statistical clerk. The costs were obtained by using the GS hourly wages; i.e., multiplying appropriate hourly wage by man-hours worked.

Table 14 shows the cost of photo counting. The times were derived from Table 12 except for office time (estimate). The cost was computed using GS hourly wages (1969).

The information in Tables 13, 14 and 15 will be used in the optimum allocation section.

Double Sampling

Estimates of peach numbers obtained from direct expansions of counts from sample limbs are good but expensive. One purpose of this project was to find variables which would be both relatively inexpensive and sufficiently correlated with the expanded limb counts so that they could be used in a double sampling model. The 1967-68 experimentation showed that counts from photographs, the number of terminals per tree, and sum of primary CSA's provided a group of such variables. The methodology of the double sampling design is explained in Hansen, Hurwitz and Madow. [8] The potential usefulness of such an estimation model can be tested as follows:

If $\rho^2 > 4C_1C_2/(C_1 + C_2)^2$, then the double sampling model will be more efficient than single stage sample; i.e., a sample estimate coming only from the limb expansions.

[8] Hansen, M., Hurwitz, W., Madow, W., Sample Survey Methods and Theory, John Wiley and Sons, Inc., N.Y., 1953, Vol. II, pp. 464-467.

Photo counts and direct expansion of limbs in a double sampling regression model:

C_1 includes the cost of photography for photo counts, photo counts (one diagonal), tree selection, and film.

$C_1 = 2.78$

C_2 includes the cost of bare tree photography, bare tree mapping, limb count, and film.

$C_2 = 6.07$

The best estimate of ρ^2 for photo count vs. expanded limb count over the two year period is $r^2 = .32$ for 1970. In the cost function, all costs of C_1 were recorded at their minimums, specifically only one diagonal and one side was assumed for photo counting of the selected trees. Even so, the right hand side of the inequality $4C_1C_2/(C_1 + C_2)^2 = 4(1.78)(3.72)/(1.78 + 3.72)^2 = .876$.

Therefore, r^2 larger than .876 would be required for the use of photography to be economically feasible as a covariate in a double sampling design. Since the observed coefficient of determination was much lower, this portion of the analysis was not continued.

Number of terminals and direct expansions of limbs in a double sampling regression model:

C_1 includes the cost of bare tree photography, and bare tree mapping.

$C_1 = 2.10$

C_2 includes the same costs as listed above for C_2, 6.07.

The highest estimate of ρ^2 for number of terminals vs. expanded limb count over the two year period is $r^2 = .346$ for 1969. But $4C_1C_2/(C_1 + C_2)^2 = 4(2.10)(3.72)/(2.10 + 3.22)^2 = .76$.

With an $r^2 = .76$ required for gains in a double sampling model, it is not necessary to carry this portion of the analysis any further.

Sum of primary CSA's and direct expansion of limbs in a double sampling regression model:

C_1 includes the cost of primary cross sectional area measurements

$C_1 = .20$

C_2 same as cost found for photo counts = 6.07

The highest estimate of ρ^2 for sum of primary CSA vs. expanded limb counts over the two year period is $r^2 = .44$ for 1969.

$4C_1C_2/(C_1 + C_2)^2 = 4(.20)(3.72)/(.20 + 3.72)^2 = .194$

The 1969 $r^2 = .44$ is greater than the ρ^2 necessary. Therefore, the 1969 data indicates that the sum of primary CSA's is economically feasible as a covariate in a double sampling model.

The double sampling model is:

Let X_i = estimated number of peaches from the ith tree in a sample of size m.

$\bar{X}'$ = average number of peaches per tree from the sample of size m.

$\bar{Y}$ = average value of covariate per tree from a sample size of n, where m is a subsample of n.

$\bar{Y}'$ = average of covariate per tree for the same trees used in computing $\bar{X}'$. Then the double sample estimate, $\bar{X}'$ is computed as $\bar{X} = \bar{X}' + B\ (\bar{Y} - \bar{Y}')$ or $\bar{X} = \bar{X}' + b\ (\bar{Y} - \bar{Y}')$

Where b is an unbiased estimate of B it is computed as

$$b = \frac{\sum^{m} (X_i - \bar{X}')\ (Y_i - \bar{Y}')}{\sum^{m} (Y_i - \bar{Y}')^2}$$

Optimum Allocation of Resources

For any sample design to be efficient it is necessary to find the combination of sampling units which minimizes the total cost for a fixed variance, or vice versa. Such a method is described by Snedecor and Cochran. 9/

Let n_1 = optimum number of blocks

n_2 = optimum number of trees per block

n_3 = optimum number of limbs per tree

c_1 = cost per block. This includes travel to and from blocks, and the time necessary to make observations on trees from which the sample is drawn.

c_2 = cost per tree. This includes bare tree photography and mapping of tree.

c_3 = cost per limb. This includes identification of limbs and limb counts.

9/ Snedecor, George, and Cochran, William, Statistical Methods, Iowa State University Press, Ames, Iowa, 1968, pp. 531-534.

$$(1)\ \text{Var} = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_1 n_2} + \frac{S_3^2}{n_1 n_2 n_3}$$

$$(2)\ \text{Cost} = n_1 c_1 + n_1 n_2 c_3 + n_1 n_2 n_3 c_3$$

$$(3)\ (\text{Var})(\text{Cost}) = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_1 n_2} + \frac{S_3^2}{n_1 n_2 n_3}\ (n_1 c_1 + n_1 n_2 c_2 + n_1 n_2 n_3 c_3)$$

It can be shown that (3) has its minimum value when

$$n_2 = \sqrt{\frac{c_1\ S_2^2}{c_2\ S_1^2}} \quad \text{and} \quad n_3 = \sqrt{\frac{c_2\ S_3^2}{c_3\ S_2^2}}$$

Then n_1 is found by solving either the cost or variance equation. To find the variance components, S_1^2, S_2^2, and S_3^2, it is necessary to compute a nested analysis of variance (Table 16).

The negative variance component computed for between trees in 1969 is not an acceptable solution. Some contributions of variance should have been attributed to the between tree component. The 1970 variance components were therefore used for optimum allocation calculations.

$$S_1^2 = 42{,}888$$

$$S_2^2 = 13{,}138$$

$$S_3^2 = 698{,}891$$

Table 17.- Estimated variance components for expanded limb counts of cling peaches, California, 1969-70

Source	1969			1970		
	Degrees of freedom	Mean square	Variance component	Degrees of freedom	Mean square	Variance component
Blocks....	21	2,229,279	262,863	17	1,074,960	42,888
Trees.....	44	652,098	-83,531	16	751,444	13,138
Limbs.....	66	819,161	819,161	102	698,891	698,891
Total.....	131	989,098		135	752,476	

For the two year period, the costs defined below remained relatively constant. The cost figures are derived from the cost section.

C_1 includes travel time between and to blocks, and tree selection

$C_1 = 7.99$

C_2 includes bare tree photography and bare tree mapping and limb selection and travel time between trees

$C_2 = 2.61$

C_3 includes limb count (annual survey) per limb

$C_3 = 1.62$

$$n_2 = \sqrt{\frac{C_1 \; S_2{}^2}{C_2 \; S_1{}^2}} \qquad = \sqrt{\frac{(5.90)\;(13,138)}{(2.61)\;(42,888)}}$$

n_2 = .832 = 1 tree per block

Since the total number of possible samplu units per tree was small, a finite population correction factor was incorporated into the formula for n_3.

$$n_3 = \sqrt{\frac{C_2\ S_3^2\ (1 - f)}{C_3\ \ S_2^2}}$$

where $1 - f = 1 - \frac{n_3}{\overline{N}_3}$,

with $\overline{N}_3$ = the average total number of terminals on the trees sampled. The value n_3 is found by squaring both sides of the equation and solving the quadratic.

$$n_3 = \left[- \frac{(C_2\ S_3^2)}{(C_3\ S_2^2\ N_3)} + \sqrt{\frac{(C_2\ S_3^2)^2}{(C_3\ S_2^2\ N_3)^2} + \frac{4(C_2\ S_3^2)}{(C_3\ S_2^2)}} \right] \Big/ 2$$

$$= \left[- \frac{(2.61)(698{,}891)}{(1.62)(13{,}138)(31.1)} + \sqrt{\left(\frac{(2.61)(698{,}891)}{(1.62)(13{,}138)(31.1)}\right)^2 + \frac{4(2.61)(698{,}891)}{(1.62)(13{,}138)}} \right] \Big/ 2$$

$$= 7.99$$

Solving the variance times cost formula for $n_2 = 1$ and $n_3 = 7$ or 8 it can be shown that the optimum combination is $n_2 = 1$, $n_3 = 8$.

To solve for the number of blocks it is necessary to solve either the variance or cost equation for a desired total variance or cost.

The second step is to study the various possible covariates, find which ones are suitable, and then determine the optimum allocation under double sampling. The double sampling section showed that the only economically feasible covariate was sum of primary CSA's.

The formula for determining the optimum ratio of number of trees with primary measurements (n) to number of trees (m) from n for counting fruit on sample limbs is: 10/

$$\frac{m}{n} = \sqrt{\frac{1 - \rho^2}{\rho^2} \quad \frac{C_1}{C_2}}$$

where m is the subsample of n

$$\frac{m}{n} = \sqrt{\frac{(1 - .40)}{(.40)} \quad \frac{.20}{3.72}}$$

$$\frac{m}{n} = \sqrt{.081}$$

$$\frac{m}{n} = 0.28$$

A ratio of .28 means limb counts should be made on about a 28 percent subsample of the trees selected for measuring the sum of primary CSA's. In this instance, the primary CSA's of four trees in each block would be measured and limb counts would be made on one of the four trees.

In 1970 the ρ^2 of .19 indicates no gains from a double sampling regression model. If this lack of consistency arose from using 1969 CSA measurements, then the measurements would need to be taken every year. In turn, this would increase the cost of C_1 to \$.80 and the ratio of $\frac{m}{n}$ to 0.57.

10/ Hansen, Hurwitz, and Madow, p. 466.

Conclusions

Many new facets of peach crop estimation have been studied. Some results were disappointing and others surprisingly successful.

It appears that there is a significant relationship between counts of peaches from photographs and actual tree counts. This relationship is overshadowed by a need for more advanced technology in photography, and for finding ways to reduce the high cost both of obtaining photography and of making the photo counts, relative to limb counts. In the field of photography, future projects should stress the need for consistency from one exposure to the next. Also, a continuing search should be conducted for photo equipment which gives maximum resolution with minimum training for field use.

Sum of primary cross sectional areas (CSA's) was the only variable which presently is economically feasible in a double sampling model when paired with direct expansion estimates. Further testing of the double sampling model using sum of primary CSA's should be carried out to compare its forecast results with present objective yield methods.

Information in this study indicated that the most efficient method of sampling,with respect to sampling error, was single-stage equal probability sampling. Other fruit studies had produced different results. This study did not attempt to compare the efficiency of the different procedures with respect to variance and cost. New studies should be started relating solely to further improvement of direct expansion methods.

Appendix I

Definitions

Block - a contiguous planting of trees of the same variety and age.

Cluster - a group of three consecutively numbered sampling units.

C.S.A. - cross sectional area, taken at the base of a section of a limb.

Expansion factor - the reciprocal of the probability of a limb being selected. Used to expand limb counts to estimate tree totals.

Green drop - when a crop is too large some trees are not harvested.

Itek print - a negative black and white enlargement (about 25x) made from a 35 mm color transparency.

Major split - fork in a tree forming either two path and/or terminal sections.

Path section - a nonterminal section of a limb. At least two terminal and/or nonterminal branches emerge from it at the next major split.

Primary - a major (at least 10 percent of the total C.S.A.) limb which emerges from the trunk of the tree.

Sample unit - a terminal limb and its associated path section (if any).

Terminal limb - a branch with a thickness generally between 1/16 and 1/8 inches as measured on the Itek prints or with a C.S.A. of 0.6 to 2.0 square inches, and from which no other terminal size branches emerge.

Appendix II

Correction for Sampling Error in Coefficient of Correlation

r_1 = underestimate of ρ (rho)

r = unbiased estimate of ρ

$r = (1 + \lambda)^{1/2} (r_1)$

$\lambda = Se^2/(Sx_1{}^2 - Se^2)$

1969

$Sx_1{}^2 = 1{,}161{,}650$

$Se^2 = 819{,}162/2 = 409{,}581$

$\lambda = (409{,}581/(1{,}161{,}650 - 409{,}581)) = .545$

Adjusted photo counts vs. direct expansion estimate:

$r = (1 + .545)^{1/2} (.14) \doteq .174$

Number of terminals vs. direct expansion estimate:

$r = (1 + .545)^{1/2} (.49) \doteq .609$

Sum of primary cross sectional areas vs. direct expansion estimate:

$r = (1 + .545)^{1/2} (.53) \doteq .659$

1970

$Sx_1{}^2 = 918{,}104$

$Se^2 = 698{,}891/4 = 174{,}723$

$\lambda = (174{,}723/(918{,}104 - 174{,}723)) = .235$

Adjusted photo counts vs. direct expansion estimate:

$r = (1 + .235)^{1/2} (.51) \doteq .566$

Number of terminals vs. direct expansion estimate:

$r = (1 + .235)^{1/2} (.31) \doteq .344$

Sum of primary cross sectional areas vs. direct expansion estimate;

$r = (1 + .235)^{1/2} (.33) \doteq .366$

Bibliography

Graybill, Franklin, An Introduction to Linear Statistical Models, New York, McGraw-Hill Book Co., Inc., 1961.

Hansen, Morris and others, Sample Survey Methods and Theory, Vol. I, New York, John Wiley and Sons, Inc., 1953.

Jessen, Raymond, "Determining the Fruit Count on a Tree by Randomized Branch Sampling," Biometrics, March 1955, pp. 99-109.

Snedecor, George and others, Statistical Methods, Ames, Iowa, Iowa State University Press, 1968.