RESEARCH REPORT
Statistical Reporting Service
U.S. Department of Agriculture

A COMPARISON OF SEVERAL REGRESSION MODELS FOR FORECASTING PECAN YIELDS

by Chapman P. Gleason
Research Division
Research and Development Branch

NOVEMBER 1974

CONTENTS

SUMMARY

Several different regression models are compared to determine which are best for forecasting average yield per tree. Criteria are proposed to determine which variables and ultimately which regression models are better than others. Using the proposed criteria, a simple linear regression model using the number of nuts counted on photographs was found to be "best". Recommendations are made for further research. A new method of expanding the number of nuts counted on photographs to the tree level is also presented. The study was based upon data collected in 1972 in central and southern Mississippi.

A COMPARISON OF SEVERAL REGRESSION MODELS FOR FORECASTING PECAN YIELDS

BY
CHAPMAN P. GLEASON

INTRODUCTION

Research studies have shown that limb sampling and photographic data collection procedures are promising methods of providing data with which to forecast the average yield (number or weight of nuts) per tree.

Two different approaches to forecasting the average yield per tree were proposed in 1971. The first involved using two simple linear regressions---yield versus the number of nuts on sample limbs; and yield versus the number of nuts on photographs. The second involves a multiple regression approach to the forecasting problem---yield versus the number of nuts on limbs and photographs. The research was aimed at answering two questions:

1. Which of the above approaches is better?
2. If simple linear regression is as good as multiple regression, which regression gives the best estimate of average yield per tree?

From a cost standpoint, one variable may be easier and cheaper to collect and provide more precise forecasts. Tests of statistical hypothesis will be formulated to answer the first question. Standard errors, R^2, and C.V.'s will be compared to answer the second.

DATA COLLECTION PROCEDURES

Block Selection:

Five block of Stuart variety pecans were subjectively selected in two separate geographic areas of Mississippi. Three blocks were located in central Mississippi (Hinds County), all managed by one operator. Two additional blocks were located in the southwestern corner of the State (Wilkinson County), each managed by different individuals.

Sample Tree Selection:

For each of the three blocks in Hinds County, the trees used in a 1971 research project were used again (Wood (8)). In Wilkinson County it was necessary to select four trees in each of the newly selected blocks. A two-stage procedure was used to select the trees.

1. Two rows were randomly selected with equal probability of selection for each row.
2. Within each selected row, two trees were randomly selected with equal probability. In this approach, if rows are varying lengths trees in short rows have a greater probability of selection than those in long rows.

Sample Limb Selection:

For each selected tree, the total number of accessible (reachable by a six-foot ladder) sample limbs was enumerated and a 50 percent simple random sample with equal probabilities of selection was taken. Sample limbs were defined as those with cross-sectional area between 1.8 and 5.5 square inches. For each tree, the total number of sample limbs (both accessible and inaccessible) was estimated using either bare tree mappings of limbs or bare tree stero photographs. N_j will denote the total estimated number of sample limbs for the i-th tree. The trees in Hinds County had stero photographs

taken in early April 1971. (Huddleston (4) describes the uses of photography to estimate the total number of sample limbs.) Bare tree mappings of limbs were made and used to estimate the total number of sample limbs for the trees selected in Wilkinson County.

Sample Limb Counts:

For each tree, once the sample limbs were selected all nuts on the limb were counted by tagging each cluster of fruit and counting the number of nuts in each tagged cluster. This prevented counting errors and gave an indication of monthly fruit droppage from the clusters.

Photography Procedures:

To avoid the poor quality photography that plagued the research efforts in 1970 and 1971, the following techniques were used:

1. The tripod which held the camera was located 50 feet from the base of the tree with the sun at the back of the photographer.
2. A florist stake was placed directly below the tripod.
3. The metal photography frame was placed two feet in front of the camera lens.
4. The angle of the camera from the tree was recorded.

The photographs were taken up a vertical column of the tree through a metal frame. A Miranda Sensorex camera with an in-lens light meter improved the photograph significantly over those with a camera which has no in-lens light meter.

Counts of Nuts from Photographs:

Each slide was projected on a grid. The number of nuts in each cell was counted by a photo interpreter. A certain subset of the slides was recounted for computation of photo adjustments factors. (See Wood (7,p.19) for a discussion of methods used to compute photo adjustment factors.)

Nut Droppage Prior to Harvest:

On the first photography visit, two square plots, each two feet by two feet, were randomly located on the ground beneath the canopy of each tree. The identified area was then gleaned for nuts. On each subsequent field visit, the amount of droppage (number of nuts) in the plot was counted and removed.

Harvest Data:

At harvest, the each tree was shaken and all "good" nuts were collected. The nuts that remained on the ground were deemed "bad". Each tree was visited three times to collect harvest data. Three one-pound samples of nuts were selected from each collection of nuts for a tree. However, it was apparent that for several of the trees that these collection of nuts were mixed due to classification errors by the laborers who harvested the trees. For this reason and the fact that a good nut cannot be distinguished from a bad nut on a photograph the biological yield was used as the dependent variable in the analysis that follows. The term LBNUTS, will denote the collection of all nuts. The total harvest data for each tree are given in Table 1.

Table 1 - Harvest Data, Mississippi Pecans, 1972

BLOCK	TREE	Pounds of "good" nuts harvested	Number of "good" nuts per pound 1/	Pounds of "bad" nuts harvested	Number of "bad" nuts per pound 1/	LBNUTS
A	1	7.1	68.7	2.1	66.5	9.2
A	2	3.5	49.3	0.5	49.3	4.0
A	3	6.2	56.0	2.1	56.0	8.3
A	4	14.5	51.3	7.5	50.3	22.0
B	1	60.8	57.0	8.2	83.3	69.0
B	2	38.0	92.7	10.3	90.0	48.3
B	3	19.1	40.0	6.2	68.7	25.3
B	4	37.5	68.3	8.0	73.0	45.5
D	1	13.6	58.7	25.4	54.3	39.0
D	2	52.6	57.7	35.7	83.0	88.3
D	3	0.3	97.0	1.9	97.0	2.2
D	4	7.1	61.7	5.0	107.7	12.1
E	1	2/	2/	2/	2/	191.5
E	2	2/	2/	2/	2/	237.5
E	3	2/	2/	2/	2/	62.8
E	4	2/	2/	2/	2/	45.1
F	1	2/	2/	2/	2/	85.5
F	2	2/	2/	2/	2/	93.2
F	3	2/	2/	2/	2/	65.2
F	4	2/	2/	2/	2/	146.2

1/ This was estimated by averaging the three one-pound samples from each collection of nuts.

2/ Determination of pounds of good/bad nuts could not be made.

DATA EXPANSIONS

Limb Expansions:

The expanded number of nuts from sample limbs (NNSL) for each tree was computed as follows:

$$(1) \quad NNSL = \frac{N_i}{n_i} \sum_{j=1}^{n_i} X_{ij}$$

where for the i-th tree, N_i is the estimated total number of sample limbs, n_i is the number of sample limbs selected, X_{ij} is the total number of fruit counted on the j-th selected sample limb. It is noted that we are sampling only those limbs which are accessible, whereas (1) is an expansion to the total tree based on the fallacious assumption that each of the N_i sample limbs had a non-zero chance of selection. Horticulturists contend that the lower limbs produce fewer nuts than the higher limbs. Hence, an under-estimate of the total number of fruit will be realized from (1).

Photography Expansions:

The counts of nuts using ground photographs were expanded to a tree level by two methods. The first expansion assumed that the shape of a tree is a sphere. (Wood (7,p.20) gives a discussion of the method.) The independent variable using this assumption is NNPS (number of nuts counted from photographs, sphere assumption.) The second expansion to a tree level assumes that the shape of the tree is a parabolid. For this expansion two parameters for every tree must be estimated, the height (h) and the radius (r). It can be proved (See Strout (5)) that estimated bearing surface area of the tree assuming the tree is shaped as a parabolid is:

$$SAP = \frac{\Pi r}{6h^2} \left((4h^2 + r^2)^{3/2} - r^3\right).$$

Thus, the number of nuts counted on photographs using the parabolid assumption (NNPP) is,

$$(2) \quad NNPP = \frac{SAP}{TAMF} \left(\sum_{j=1}^{n_i} X_{ij}\right)$$

Where n_i is the number of photograph taken on the i-th tree, X_{ij} is the number of nuts on the j-th photograph, TAMF is the total area of the middle frame. (See Wood (7, p.22)).

The number of nuts counted on photographs were adjusted to reflect the fact that each photo interpreter counts a different number of nuts for any given slide. To minimize interpreter differences and to measure such deviation from the "norm" a balanced incomplete block design was used in counting the slides. (Wood (7,p.19)) gives a discussion of methods used to estimate interpreter adjustment factors.) The count of fruit on each slide was adjusted for interpreter differences by multiplying the interpreter adjustment factor times the number of fruit counted. When two interpreters counted the same slide these adjusted counts were averaged.

The radius (r) was estimated by averaging the longest and the shortest distance from the trunk to the edge of the tree canopy. The height (h) was roughly estimated by using the number of photographs n_i taken and knowing the distance from the trunk to the camera. As with limb counts, these methods of expansion to the tree level will under-estimate the true number of nuts on the tree since all nuts do not grow on the periphery of the tree. However, since flower buds develop on new growth that tends to occur on the periphery of the tree, most of the fruit is produced near the surface.

Drop Expansion:

The nut droppage from the i-th tree was estimated as follows:

$$(3) \quad DROP = \frac{\pi r^2}{8} \left(\sum_{j=1}^{2} X_{ij} \right)$$

where (r) is the estimated radius, and X_{ij} is the number of nuts in the j-th drop count unit for the i-th tree. Observe that πr^2 is the area of a circle and 8 is the total area sampled using both 2' x 2' drop units, so the ratio $\pi r^2/8$ is an area expansion factor.

RESULTS

General:

Previous investigations (by Wood (7,8)) found that both NNSL and NNPS to be significantly correlated with the estimated number of good nuts at harvest. In this investigation the biological yield, or total weight of harvested nuts --- LBNUTS, was the dependent variable. In addition to the reasons mentioned previously, the number of good nuts is a variable that is influenced by marketing and other economic conditions which were uncontrolled or immeasurable in the research project.

Two data sets were used in the analysis. The first were unadjusted counts from color transparancies, the second were counts adjusted for interpreter differences.

Correlation Coefficients:

Tables 2 through 7 gives the product moment correlation coefficient for each pairwise combination of variables, its significance probability and the number of observations. The significance probability of a correlation coefficient is the probability that a larger (in absolute value) correlation coefficient should arise by chance of the true population parameter $\rho=0$. The pairwise correlation was computed based on the assumption that the random variables have bivariate normal distributions.

Table 2: Correlation Matrix, unadjusted photography data, Mississippi Pecans, July 1972
Correlation Coefficients/ Prob > |R| under H_0:ρ=0/ number of observations

	LBNUTS	NNPS	NNPP	LIMB 1/
LBNUTS	1.000000 0.0000 20	0.589406 0.0063 20	0.692073 0.0010 20	0.449205 0.0512 19
NNPS		1.000000 0.0000 20	0.973957 0.0001 20	0.853912 0.0001 19
NNPP			1.000000 0.0000 20	0.814804 0.0001 19
LIMB				1.000000 0.0000 19

Table 3: Correlation Matrix, unadjusted photography data, Mississippi Pecans, August 1972
Correlation Coefficients/ Prob > |R| under H_0:ρ=0/ number of observations

	LBNUTS	NNPS	NNPP	LIMB 1/	DROP
LBNUTS	1.000000 0.0000 20	0.835472 0.0001 20	0.909419 0.0000 20	0.439499 0.0571 19	-0.022093 0.9235 20
NNPS		1.000000 0.0000 20	0.970281 0.0001 20	0.565588 0.0112 19	-0.038837 0.8651 20
NNPP			1.000000 0.0000 20	0.473863 0.0384 19	-0.051213 0.8245 20
LIMB				1.000000 0.0000 19	0.392730 0.0931 19
DROP					1.000000 0.0000 20

1/ There were no accessible sample limbs on tree F4

Table 4: Correlation Matrix, unadjusted photography data, Mississippi Pecans, September, 1972
Correlation Coefficients/ Prob > |R| under H_0:ρ=0/ number of observations

	LBNUTS	NNPS	NNPP	LIMB 1/	DROP
LBNUTS	1.000000 0.0000 20	0.750568 0.0003 20	0.821828 0.0001 20	0.427667 0.0650 19	-0.066414 0.7771 20
NNPS		1.000000 0.0000 20	0.977678 0.0001 20	0.638995 0.0035 19	-0.093471 0.6967 20
NNPP			1.000000 0.0000 20	0.568388 0.0108 19	-0.142778 0.5545 20
LIMB				1.000000 0.0000 19	0.335533 0.1573 19
DROP					1.000000 0.0000 20

Table 5: Correlation Matrix, adjusted photography data Mississippi Pecans, July 1972
Correlation Coefficients/ Prob > |R| under H_0:ρ=0/ number of observations

	LBNUTS	NNPS	NNPP	LIMB 1/
LBNUTS	1.000000 0.0000 20	0.71448 0.0001 20	0.805644 0.0001 20	0.449205 0.0512 19
NNPS		1.000000 0.0000 20	0.972430 0.0001 20	0.745579 0.0004 19
NNPP			1.000000 0.0000 20	0.664007 0.0022 19
LIMB				1.000000 0.0000 19

1/ There are no accessible sample limbs on tree F4

Table 6: Correlation Matrix, adjusted photography data Mississippi Pecans, August 1972
Correlation Coefficients/ Prob > |R| under H_0:ρ=0/ number of observations

	LBNUTS	NNPS	NNPP	LIMB 1/	DROP
LBNUTS	1.000000 0.0000 20	0.899645 0.0001 20	0.914822 0.0001 20	0.439499 0.0571 19	-0.022093 0.9235 20
NNPS		1.000000 0.0000 20	0.971275 0.0001 20	0.462843 0.0438 19	-0.057109 0.8058 20
NNPP			1.000000 0.0000 20	0.332710 0.1611 19	-0.069774 0.7669 20
LIMB				1.000000 0.0000 19	0.392730 0.0931 19
DROP					1.000000 0.0000 20

Table 7: Correlation Matrix, adjusted photography data Mississippi Pecans, September 1972
Correlation Coefficients/ Prob > |R| under H_0:ρ=0/ number of observations

	LBNUTS	NNPS	NNPP	LIMB 1/	DROP
LBNUTS	1.000000 0.0000 20	0.846902 0.0001 20	0.874279 0.0001 20	0.427667 0.0650 19	-0.066414 0.7777 20
NNPS		1.000000 0.0000 20	0.973825 0.0001 20	0.538335 0.0166 19	-0.108465 0.6530 20
NNPP			1.000000 0.0000 20	0.418659 0.0715 19	-0.163419 0.5024 20
LIMB				1.000000 0.0000 19	0.335533 0.1573 19
DROP					1.000000 0.0000

1/ There were no accessible sample limbs on tree F4

Analysis of Regression Models:

Consider the linear regression model:

$$(1) \quad Y = \beta_o + \beta_1 X_1 \ldots + \beta_k X_k + \varepsilon$$

In the classical linear regression, Y is an observable random variable, the X_i's are fixed observable variables, and the error term is an unobserved random disturbance. However, in our situation the regressor variables are observable stochastic variables which are assumed to be distributed independent of the disturbance. All the classical tests and estimation procedures are valid when this assumption can be justified. (Goldberger (3) discusses stochastic regression.)

In the analysis presented, the following models of the above form were considered:

$$(M1) \quad Y = \beta_o + \beta_1 (NNPS)$$

$$(M2) \quad Y = \beta_o + \beta_1 (NNPP)$$

$$(M3) \quad Y = \beta_o + \beta_1 (LIMB)$$

$$(M4) \quad Y = \beta_o + \beta_1 (NNPS) + \beta_2 (LIMB)$$

$$(M5) \quad Y = \beta_o + \beta_1 (NNPP) + \beta_2 (LIMB)$$

$$(M6) \quad Y = \beta_o + \beta_1 (NNPS) + \beta_2 (LIMB) + \beta_3 (DROP)$$

$$(M7) \quad Y = \beta_o + \beta_1 (NNPP) + \beta_2 (LIMB) + \beta_3 (DROP)$$

Y is the dependent variable LBNUTS. Note that M6 and M7 have more terms than M1 and M2 because of the inclusion of two independent variables. The difference between M1 and M2 (or between M4 and M5) is just the different methods of expansion of the photography variable.

Several questions arise about the models M1 through M7. Are all the independent variables necessary in models M6 and M7? Which, if any, of the seven models is the "best" regression model for forecasting average

weight of nuts per tree? These questions were considered and criteria formulated to answer them.

Seven criteria will be proposed to answer the above questions. The list of criteria is certainly not exhaustive but was chosen to evaluate certain disirable properties. The criteria are as follows:

(C1) The square of the multiple correlation coefficient, R^2. The R^2 value should increase by the inclusion of another or possibly several independent variables into the model. The larger the R^2 value, the better the model explains the variation in the data. A substantial increase in the R^2 value for any model over M1 (or M2) by including the variable LIMB into the regression would indicate that the LIMB variable is explaining some additional variation in the data.

(C2) The standard error of estimate, $s=\sqrt{s^2}$, the residual mean square estimates the variance about regression $\sigma^2 Y \cdot X$. The smaller the value of s the more precise will be the predictions.

(C3) The coefficent of variation, **CV** , The $CV = s/\bar{Y}$ should decrease if increased precision is obtained by the inclusion of another variable.

(C4) The sequential F-test. This criterion accesses the contribution of another variable added to an equation in stages. In (1) let $SS(b_o,\ldots,b_k)$ be the sums of square due to regression.

Now for $j=1,2,\ldots,k$ let $SS(b_j|b_o,b_1\ldots,b_{j-1})$ be the sequential sums of squares for the j-th beta parameter. $SS(b_j|b_o,b_1,\ldots,b_{j-1})$ is the difference between the sums of squares due to the regression of Y on $X_1,\ldots,X_j$ and the sums of square due to the regression of Y on $X_1,\ldots,X_{j-1}$. This is denoted by $SS(b_o,b,\ldots,b_j)$ and $SS(b_o,b,,\ldots,b_{j-1})$, respectively. The j-th sequential F-test $j=1,2,\ldots k$ is:

$$F(b_j|b_o,\ldots,b_{j-1}) = \frac{SS(b_j|b_o,\ldots,b_{j-1})}{ESS(b_o,b,\ldots,b_k)/N-(k+1)}$$

$ESS(b_o,b_1,\ldots,b_k)$ is the residual SS of general model (1), and N is the number of units in the sample. The above F has 1 and N-(k+1) degrees of freedom. Note that,

$$\sum_{j=1}^{k} SS(b_j|b_o,\ldots,b_{j-1}) = SS(b_o,\ldots,b_k).$$

Thus, the total sum of squares due to regression in the full model (1) is just partioned into single degrees of freedom SS's.

(C5) The partial F-test criteria. This criteria considers the order in which the variables enter into the model. This criteria accesses the value of a variable as if it were to enter the regression equation last. The effect of X_j may be larger when the regression equation includes only X_j. However, when the same variable entered into the equation after other variables, it may affect the response very little. The F-test is as follows. For j=1,...,k

$$F(b_j|b_o,b_1,\ldots,b_{j-1},b_{j+1},\ldots,b_k) = \frac{SS(b_j|b_o,b_1,\ldots,b_{j-1},b_{j+1},\ldots,b_k)}{ESS(b_o,b_1,\ldots,b_k)/N-(K+1)}$$

where,

$$SS(b_j|b_o,\ldots,b_{j-1},b_{j+1},\ldots,b_k) = SS(b_o,\ldots,b_k) - SS(b_o,\ldots,b_{j-1},b_{j+1},\ldots b_k)$$

$SS(b_o,\ldots,b_{j-1},b_{j+1},\ldots,b_k)$ is the sum of squares due to the regression of Y on $X_1,X_2,\ldots,X_{j-1},X_{j+1},\ldots,X_k$, i.e. the regression on all variables except the j-th. This F has 1 and N-(k+1) degrees of freedom. It is not noted that

$$T = \sqrt{F(b_j|b_o,\ldots,b_{j-1},b_{j+1},\ldots b_k)}$$

has the T distribution with N-(k+1) d.f., and this statistic is used to

test if $\beta_j=0$ in (1). Thus, the j-th partial F-test is equivalent to a T-test of $\beta_j=0$.

(C6) The extra sums of square criteria. This criteria accesses whether it was worth while to include certain terms in the general regression model (1). It is a joint test of the parameters $\beta_{j+1},\ldots,\beta_k$ in (1). Consider the reduced model

$$(2)\quad Y = \beta_o+\beta_1X_1+\ldots+\beta_qX_q \text{ where } q<k.$$

And let $SS(b_o,\ldots b_q)$ denotes the SS due to the regression (2). Then $SS(b_{q+1},..b_k|\, b_o,\ldots,b_q)=SS(b_o,\ldots,b_k)-SS(b_o,b_1,\ldots b_q)$ is the extra SS due to the inclusion of the terms $\beta_{q+1}X_{q+1}+\ldots+\beta_kX_k$ into the model (1). Now, the sum of squares $SS(b_o,\ldots,b_k)$ has k d.f. and $SS(b_o,\ldots,b_q)$ has q d.f., thus $SS(b_{q+1},\ldots,b_k|\, b_o,\ldots,b_q)$ has k-q d.f. So if $\beta_{q+1}=\beta_{q+2}=\ldots=\beta_k=0$ then $SS(b_{q+1},\ldots,b_k|\, b_o,\ldots,b_q) \sim\sigma^2\chi^2$ k-q, and is independnet of $ESS(b_o,b,\ldots b_k)$. Hence, $F(b_{q+1},,\ldots b_k|\, b_o,\ldots,b_q) = \dfrac{SS(b_{q+1},\ldots,b_k|\, b_o,\ldots,b_q)/k-q}{ESS(b_o,b_1,\ldots,b_k)/N-(k+1)}$

has the F distribution with k-q and N-(k+1) d.f.

(C7) Significance of regression. This criteria determines whether the regression of Y on $X_1,\ldots,X_k$ is significant. The test is

$$F = \frac{SS(b_o,\ldots,b_k)/k}{ESS(b_o,b_1,\ldots b_k)/N-(K+1)}$$

This is a test of the hypothesis $H:\beta_1=\beta_2=\ldots=\beta_k=0$, which is equivalent to testing that the true multiple correlation coefficient R is 0.

The seven criteria can be broken into two general areas. First, the R^2, s, and CV are measures of how well the linear regression model fitted the data. The other four criteria are statistical tests on certain parameters in the regression model. Tables 8 through 13 present the seven

criteria to determine the "best" regression model. The analysis was done using the STEPWISE procedure of the Statistical Analysis System (1). This program deleted records with missing observations. Since tree F-4 has no accessible sample limbs it was deleted in the analysis presented.

Table 8: Criteria to determine the "best" regression model, adjusted photography data, Mississippi Pecans, July 1972

MODEL	Criteria for "best" regression model									
	R^2	s	C.V.%	Sequential F-test for parameters 2/			Partial F-test for parameters 2/			F for significance of regression
				β_1	β_2	β_3 1/	β_1	β_2	β_3 1/	
M1	0.630	38.780	63.8	28.97*			28.97*			28.97*
M2	0.741	32.486	53.5	48.52*			48.53*			48.52*
M3	0.202	56.977	93.8	4.30**			4.30**			4.30**
M4	0.676	37.414	61.6	31.13*	2.26		23.43*	2.26		16.70*
M5	0.767	31.716	52.2	50.90*	1.84		38.87*	1.84		26.37*

** Indicates the F is significant $\alpha = .10$
* Indicates the F is significant $\alpha = .01$
1/ Drop was not observed the first month
2/ A blank indicates that the F test is not applicable with this model.

Table 9: Criteria to determine the "best" regression model, adjusted photography data, Mississippi Pecans, August 1972

MODEL	Criteria for "best" regression model											
	R^2	s	C.V.%	Sequential F-test for parameter 1/			Partial F-test for parameter 1/			F-test for extra SS Criteria models M6 and M7 $H_o: \beta_2 = \beta_3 = 0$	F for significance of regression
				β_1	β_2	β_3	β_1	β_2	β_3		
M1	0.901	20.089	33.1	154.32*			154.32*				154.32*
M2	0.869	23.065	38.0	112.96*			112.96*				112.96*
M3	0.193	57.284	94.3	4.07**			4.07*				4.07**
M4	0.901	20.707	34.1	145.25*	0.00		114.10*	0.00			72.62*
M5	0.888	21.999	36.2	124.17*	2.68		99.26*	1.68			68.43*
M6	0.904	20.997	34.6	141.27*	0.44	0.12	105.85*	0.56	0.12	0.281	47.28*
M7	0.888	22.720	37.4	116.42*	2.52	0.00	88.21*	2.09	0.00	1.26	39.65*

** Indicates the F is significant $\alpha = .10$
* Indicates the F is significant $\alpha = .01$
1/ A blank indicates that the F-test is not applicable with this model.

Table 10: Criteria to determine the "best" regression model, adjusted photography, Mississippi Pecans, September 1972

MODEL	Criteria for "best" regression model											
	R^2	s	C.V.%	Sequential F-test for parameter 1/			Partial F-test parameter 1/			F-test for extra SS Criteria models M6 and M7	F for significance of regression
				β_1	β_2	β_3	β_1	β_2	β_3	$H_o: \beta_2=\beta_3=0$	
M1	0.818	27.200	44.8	76.45*			76.45*				76.45*
M2	0.812	27.655	45.5	73.41*			73.41*				73.45*
M3	0.183	57.647	94.9	8.81**			3.81**				3.81**
M4	0.823	27.653	45.5	73.96*	0.45		57.87*	0.44			37.20*
M5	0.815	28.271	46.5	70.24*	0.27		54.68*	0.27			35.26*
M6	0.846	26.683	43.9	79.45*	0.94	1.73	61.75*	2.19	1.73	1.33	27.371*
M7	0.829	28.050	46.2	71.35*	1.52	0.00	54.45*	1.25	0.00	0.76	24.28*

** Indicates the F is significant. $\alpha = .10$
* Indicates the F is significant. $\alpha = .01$
1/ A blank indicates that the F-test is not applicable with this model.

Table 11: Criteria to determine the "best" regression model, unadjusted photography Mississippi Pecans, July 1972

MODEL	Criteria for "best" regression model										F for significance of regression
	R^2	s	C.V.%	Sequential F-test for parameter 2/			Partial F-test for parameter 2/				
				β_1	β_2	β_3 1/	β_1	β_2	β_3 1/		
M1	0.421	48.527	79.9	12.36*			12.36*				12.36*
M2	0.518	44.258	72.9	18.30*			18.30*				18.30*
M3	0.202	56.977	93.8	4.30**			4.30**				4.30**
M4	0.462	48.235	79.4	12.51*	1.20		7.72**	1.21			6.86*
M5	0.575	42.878	70.6	19.49*	2.11		14.02*	2.11			10.80*

** Indicates the F is significant $\alpha = .10$
* Indicates the F is significant $\alpha = .01$
1/ Drop was not observed the first visit
2/ A blank indicates that the F-test is not applicable with this model.

Table 12: Criteria to determine the "best" regression model, unadjusted photography data, Mississippi Pecans, August 1972

MODEL	Criteria for "best" regression model												
	R^2	s	C.V.%	Sequential F-test for parameter 1/			Partial F-test for parameter 1/			F-test for extra SS Criteria models M6 and M7: $H_o: \beta_2=\beta_3=0$	F for significance of regression
				β_1	β_2	β_3	β_1	β_2	β_3		
M1	0.793	29.041	47.8	64.98*			64.98*				64.98*
M2	0.874	22.663	37.3	117.62*			117.62*				117.62*
M3	0.193	57.284	94.3	4.07**			4.07**				4.07*
M4	0.799	29.496	48.6	62.99*	0.48		48.12*	0.48			31.73*
M5	0.874	23.358	38.5	110.71*	0.00		86.24*	0.00			55.36*
M6	0.806	29.906	49.2	61.27*	0.47	0.56	44.56*	0.94	0.56	0.546	20.77*
M7	0.876	23.894	39.3	105.81*	0.20	0.09	78.32*	0.29	0.09	0.147	35.37*

** Indicates the F is significant $\alpha = .10$
* Indicates the F is significant $\alpha = .01$
1/A blank indicates that the F-test is not applicable with this model.

Table 13: Criteria to determine the "best" regression model, unadjusted photography, Mississippi Pecans, September 1972

MODEL	Criteria for "best" regression model												
	R^2	s	C.V.%	Sequential F-test for parameter 1/			Partial F-test for parameter 1/			F-test for extra SS Criteria models M6 and M7		F for significance of regression
				β_1	β_2	β_3	β_1	β_2	β_3	$H_o: \beta_2=\beta_3=0$		
M1	0.668	36.724	60.5	34.27*			34.27*					34.27*
M2	0.749	31.915	52.6	50.88*			50.88*					50.88*
M3	0.183	57.647	94.9	3.81**			3.81**					3.81*
M4	0.688	36.978	60.9	33.80*	0.77		25.32*	0.77				17.28*
M5	0.756	32.492	53.5	49.09*	0.40		37.51*	0.40				24.74*
M6	0.714	36.328	59.8	35.02*	0.79	1.58	26.40*	2.02	1.58	1.19		12.46*
M7	0.790	31.114	51.2	53.53*	1.00	1.88	41.44*	2.45	1.88	1.44		18.81*

** Indicates the F is sifnificant $\alpha = .10$
* Indicates the F is significant $\alpha = .01$
1/ A blank indicates that the F-test is not applicable under this particular model.

DISCUSSION OF RESULTS

Inspecting the correlation matrices (Tables 2-7) indicate that most of the variables are correlated with yield (LBNUTS). However, DROP is not significantly correlated with yield for any month, nor was DROP significantly correlated with any of the other independent variables. This confirms earlier findings of Wood (7,p.,15;6,p.12).

However, DROP was included in all the regression analysis presented since even if the correlation between two variables is small it may influence the multiple correlation coefficient (r) a great deal when several variables are in a regression model simultaneously. In this particular case, it was not true that DROP had a substantial effect on the coefficient of determination, which is the square of the multiple correlation coefficient. This can be seen by comparing Models M4 and M6 and Models M5 and M7 in Tables 8 through 13. Based on these findings and previous results (by Wood (6,7)) drop counts should not be included in any further work in developing a pecan forecast model.

There also appears to be a stronger relationship between final harvested weight of nuts with adjusted photography variates, than with the unadjusted photography. This indicates that interpreter adjustment factors are necessary. The sample correlation coefficient of the photo variable with LBNUTS is always greater than the limb count variable LIMB. Also, in general, the photo count variable NNPP has larger sample correlation coefficient than does the photo count variable NNPS. This could possibly be attributed to a more precise estimate of the bearing surface of a tree by assuming the tree is a parabolid rather than a sphere.

Tables 8 through 13 show that:

1. Each regression is significant at the .01 level.
2. The F-test is the extra SS criteria (where applicable) is insignificant. Thus, β_2 and β_3 are simultaneously zero in Models M6 and M7.
2. The partial F-test indicates that the LIMB and DROP variables contributed very little when they were included in the last stage. However, the contribution of the photo count variables is important even when the LIMB and/or DROP variables are introduced in the equation first. This is indicated by the significant Partial-F of the β_1 parameter.
4. Once the photo variable was in the model the contribution of additonal variables were significant. This is indicated by the sequential F-test.
5. Comparing Models M1 through M3 indicates that in each case M1 and M2 have larger R^2's and smaller standard errors than Model M3.

What the seven criteria indicate is that only one variable needs to be considered (and collected); it is the photographic variable. Further, M1 or M2 is the "best" regression model to use to forecast yield per tree. Table 14 and 15 give the estimated regression parameters for Models M1 and M2. When fitting these models, plots of residuals were examined for any departure from any of the underlying assumptions. (Draper and Smith (2) describe methods for examining residuals.) None was found.

Table 14: Estimate of regression parameters, adjusted data, by month and model, Mississippi Pecans, 1972.

MODEL	MONTH					
	JULY		AUGUST		SEPTEMBER	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
M1	15.457	0.023	13.541	0.025	15.154	0.016
M2	11.567	0.026	18.779	0.022	18.972	0.148

Table 15: Estimate of regression parameters, unadjusted data, by month and model, Mississippi Pecans, 1972.

MODEL	MONTH					
	JULY		AUGUST		SEPTEMBER	
	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\beta}_1$
M1	24.200	0.019	18.516	0.021	21.317	0.013
M2	18.435	0.024	16.687	0.023	18.521	0.015

CONCLUSIONS

Based upon the analysis performed on data collected in 1972, only photographs need to be collected in any further pecan research for forecasting yield per tree (LBNUTS) until improvements in the limb sampling procedure can be achieved which will make the accessible limbs representative of a larger portion of the tree. This will require the use of a type of mechanical lift equipment which was not available for this study. The variable DROP failed to be significantly correlated with yield nor was it useful in model building. The variable LIMB is not needed in any forecasting model once any type of photographic variable is in the model.

RECOMMENDATIONS

Future research studies should focus attention on photographic data collection and improving this technique for this particular nut crop. Different expansions to a tree level using photography may produce even better results. A more refined estimate of the height may also improve the expansions on a per tree basis. However, other characteristics of the tree and its immediate environment must not be overlooked. For example, how do differing management techniques influence yield? Possibly this answer is "greatly", indicating that stratification based on management practices might be necessary.

A random selection of blocks of different varieties and ages is needed to determine if different forecast models are needed. This will necessitate a complete sampling frame of operations for the population. Accurate estimates of tree numbers by individual blocks must be secured for each operation.

Future investigation should also consider whether monthly models

are necessary for forecasting yield per tree. Possibly the regression parameters would be stable over months during the growing season indicating that just the development and maintenance of only one forecasting equation would be necessary. Thus, monthly change in average yield per tree would be reflected in the change in average photo counts per tree and not in the change in beta parameters.

The under-estimation of total number of nuts counted on photographs needs further analysis and investigation. If the magnitude of under-estimation is consistent year-to-year, the use of a relative change forecast of production could be utilized based on either photo counts or accessible limb counts. This method of estimation is used in Florida on citrus. For example, an estimation of the following form for a particular variety and age class might be

$$(1)\quad P_t = \frac{N_t}{N_{t-1}} \times \frac{T_t}{T_{t-1}} \times P_{t-1}, \text{ where}$$

P_t is the forecast of production in year t,

P_{t-1} is the actual production for the previous year,

N_t is the forecasted average weight (or number) of nuts per tree using the photo expansion for year t,

N_{t-1} is the average weight (or number) of nuts per tree using the photo expansion for year t-1,

T_t is the number of bearing trees of a particular age and variety for year t,

T_{t-1} is the number of bearing trees of a particular age and variety for year t-1.

Another ratio (H_t/H_{t-1}) could be included in (1) to indicate the proportion of nuts intended for commercial harvest. This ratio would

probably be very volatile since price and the tendency of the trees to be cyclic in yield usually determine whether a noncommercial operator will harvest his pecan crop.

It should be noted that for this forecast the actual production is needed for a particular region by variety and age of trees. Also, accurate estimates of the relative change in number of bearing trees must be secured. Observe also that N_t/N_{t-1} is the relative change in estimated weight of nuts per tree, so that if the method of expansion and estimation consistently under-estimates the true weight of nuts per tree, this effect will cancel out in the ratio. A more detailed discussion of this forecast method can be found in Stout (5) and Williams (6).

Finally, counting the nuts on slides is tedius difficult, and very time consuming. Automated fruit counting procedures would be extremely desirable for any operational level study.

REFERENCES

1. Barr, Anthony J., and Goodnight, James H., "A User's Guide to the Statistical Analysis System", Raleigh: Student Store, North Carolina State University, 1971.
2. Draper, Norman and Smith, Harry, "Applied Regression Analysis", New York: John Wiley and Sons, 1966.
3. Goldberger, Arthur S., "Econometric Theory", New York: John Wiley and Sons, 1964.
4. Huddleston, Harold F., "The Use of Photography in Sampling for Number of Fruit Per Tree", Agriculture Economic Research, July 1971, Vol. 23, No. 3.
5. Stout, Roy G., "Estimating Citrus Production by Use of Frame Count Survey", Journal of Farm Economics, November 1962, Vol. XLIV, No. 4. pp. 1037-1049.
6. William, S. R., "Forecasting Florida Citrus Production Methodology & Development", January 1971, "Florida Crop and Livestock Reporting Service", Orlando, Florida.
7. Wood, Ronald A., "A study of the Characteristics of the Pecan Tree for Use in Objective Yield Forecasting", Research and Development Branch, Standards and Research Division. Statistical Reporting Service, U.S. Department of Agriculture, Washington, D.C.
8. Wood, Ronald A., "The Development of Objective Procedures to Estimate Yield for Pecan Trees", Research and Development Branch, Research Division, Statistical Reporting Service, U. S. Department of Agriculture, Washington, D.C.