1018

A Foundation for Objective Forecasts of Cotton Yields

By Walter A. Hendricks and Harold F. Huddleston

As part of the expanded research program in the Agricultural Estimates Division, Agricultural Marketing Service, United States Department of Agriculture, extensive plant observations were made over the old Cotton Belt during the 1954 season. The data were studied in relation to final yields so that similar observations, taken before harvest in the 1955 season, may be used to make experimental forecasts of yields. This paper summarizes the findings of the 1954 work and indicates how the results may be used in 1955. The results are in terms of average relationships for the region as a whole and should not be expected to apply to any one locality within the region.

FOR THIS STUDY, a sample of about 200 cotton fields was selected, with probabilities proportional to size of fields, from a list of about 3,000 farms that were enumerated in a probability area sample in June 1954. Each sample field was visited as of August 1 and September 1 to get counts of bolls, blooms, and squares, together with data on weevil and other damage. On the second visit samples of open cotton were weighed and small portions were taken to the office for determinations of moisture loss. A third visit to the sample fields was made at the end of the season to get farmers' reported production for the entire farm and for each sample field, and to check the amount of open cotton and the number of unopened bolls left in the sample fields after harvest.

The sampling units used for plant observations within fields consisted of two adjacent 10-foot row segments; two such double-row units were selected in each sample field. The August 1 and September 1 plant observations were taken on the same units; new units were chosen for the post-harvest observations.

All hills or plants, and the burrs, open bolls, and large unopened bolls in the sampling units were counted. The fruit counts were by separate categories in the September 1 survey, but they were lumped together in the August 1 survey. In both surveys, however, detailed counts by categories were made on 2 hills or plants for each 10-foot row section. These detailed counts also included a count of squares, blooms, and small bolls. Fruit was stripped from half of these plants and counted again, to verify the original counts. The picked fruit was examined for weevil damage.

These data were studied from the viewpoint of developing an objective forecasting procedure in regard to yields.

The Multiple Regression Approach

The multiple regression approach ordinarily comes to mind first in such problems. When data on final yields are available, together with counts of squares, blooms, small bolls, large bolls, and open bolls, as of a given date, a multiple regression equation presumably can be developed to describe the relationship of fruit counts as of that date to final yield. But two difficulties arise in attempting to evaluate the net regression coefficients in such an equation. Net regression coefficients estimated from observed data often have large sampling errors. But there is a still more serious objection to this approach. If the equation is to be used to forecast yields in future years, it should describe the relationship between fruit counts and yields over time. In this case, that means a "between-year" regression. But when data are at hand for only one year it is impossible to compute the "between-year" regression.

As stated previously, growers were requested to report final yields on the sample fields at the end of the season. But those reported yields were apparently at too low a level, as compared with Bureau of Census ginnings data for the region as a whole. For this reason, no attempt was made to relate fruit counts on the sample fields to the yields reported for these fields. Instead, the fruit counts as of August 1 were related to the September 1 count of large bolls plus open bolls. Data from 4 hills, 2 per sampling unit in each sample field on which fruit was counted both months, were

used for this purpose. The resulting equation is

$$Y=1.503+0.855X_1+0.0773X_2+0.2540X_3+0.3705X_4$$

In this equation, Y is the number of large bolls plus open bolls counted on 2 hills in a sampling unit as of September 1 and the independent variables are corresponding August 1 fruit counts on the same hills as follows:

X_1=number of large bolls
X_2=number of small bolls
X_3=number of blooms
X_4=number of squares

If this equation is interpreted literally, it says that the squares and blooms present on August 1 had a greater probability of reaching the large-boll stage by September 1 than did the small bolls. Such a conclusion would hardly be in accord with fact. The most reasonable interpretation that can be placed upon these results is that they arose from the varying degrees of maturity of plants in different parts of the Cotton Belt and that the relationship implied by the equation is spurious. It was decided not to pursue this approach further.

A "Probability of Survival" Model

As the standard multiple-regression approach is subject to the limitations outlined above, it was decided to attempt to deduce the numerical values of the net regression coefficients instead of attempting to evaluate them from the observed data. This involves setting up some sort of realistic hypothesis about the probability of survival for each category of fruit, counted as of August 1, during the period August 1 to September 1. One of the simplest hypotheses that might be proposed is that this probability is equal to the age of the fruit on August 1 divided by the age at which it is "mature."

About 21 days are normally required for a new square to become a bloom. Hence, the average age of the squares counted as of August 1 may be taken as approximately 10.5 days.

Blooms normally exist for only about 2 days before they become small bolls. A small boll becomes a "large" boll 21 days thereafter. Hence, the average age of fruit in the bloom stage can be taken as 22 days and that in the small-boll stage as 33.5 days.

From this discussion it appears that the total time required for a new square to reach the large-boll stage is 21+2+21=44 days. The probabilities of survival may thus be estimated as shown below.

Large bolls: 44/44=1.000
Small bolls: 33.5/44=.761
Blooms: 22/44=.500
Squares: 10.5/44=.239

The equation for translating August 1 fruit counts into an estimate of large bolls present on the same plants as of September 1 would thus take the form,

$$Y=X_1+0.761X_2+0.500X_3+0.239X_4$$

Applying this equation to the August 1 fruit counts gives an estimated average of 58.1 large bolls per 10 feet of cotton row as of September 1. This compares with an average of 56.6 large bolls actually counted per 10 feet of row on that date. This suggests that a satisfactory model can be devised by some such approach as an alternative to the usual multiple regression approach.

The simple hypothesis upon which the equation is based could doubtless be refined much further but such attempted refinements would be meaningless unless they were accompanied by more detailed objective data with which these hypotheses could be tested.

The research program for the present crop year makes provision for tagging fruit in the various categories on sample plants early in the season and tracing the development of each class of fruit throughout the season. This should be of considerable help in arriving at a valid forecasting equation. Meanwhile, several alternative hypotheses to the simple one described above have been tried on an exploratory basis. These all lead to equations with coefficients approximately equal to those obtained above.

An Empirical Approach

Until some of the questions raised by the studies outlined above can be answered, an approach that compares the fruit counts in the various categories made on August 1 with those made on September 1, and with the situation at harvest, can be used to determine these probabilities empirically. For convenience, all counts are expressed in terms of counts per 10 feet of cotton row.

As of August 1 these counts are 78.5 squares, 28.7 blooms plus small bolls, and 22.6 large bolls.

Bloom and small-boll counts were combined because the life of a bloom is so short that it did not seem necessary to treat blooms separately.

As of September 1 the counts are 12.1 blooms plus small bolls and 56.6 large bolls. Squares were not counted because it was believed that squares present on September 1 would not be likely to mature by harvesttime.

To complete the picture, a count of bolls picked at harvest and a count of fruit still on the plants after harvest are needed. It was intended to derive an estimate of bolls picked by dividing the farmer's reported yield for each sample field after harvest by the weight of cotton per boll, derived from weighings of open cotton made as of September 1. But, as stated earlier, farmers' reports on yields for the sample fields appeared to be at too low a level when compared with Census ginnings data at the end of the year. For that reason it seemed preferable to base the estimate of the number of bolls picked in the sample fields on the official yield estimate for the entire region.

The weight of seed cotton per boll, found by weighing cotton picked from open bolls as of September 1, was only slightly higher than the weight customarily assumed by cotton growers—1 pound of seed cotton per 100 bolls. Therefore, the standard factor was used. Assuming that 1 pound of seed cotton is equivalent to 100 bolls, and that 100 pounds of seed cotton are equivalent to 37 pounds of lint, it was possible to estimate the number of bolls per 10 feet of row picked by farmers.

The numbers of open and unopened bolls remaining on the plants after harvest were counted when the post-harvest observations were taken. Adding these counts to the estimate of bolls picked by the farmer gave a total estimate of 68.8 bolls per 10 feet of row at harvesttime. Of this total, 91 percent represents fruit picked by the farmer and 9 percent represents fruit still on the plants after the farmers finished harvesting. About half of this 9 percent represents open bolls that were missed in the harvesting operation or that opened after harvest was completed. The remaining half represents bolls that failed to mature, including those that were killed by drought.

Several features of these figures are worthy of note. First, the sum of small bolls and large bolls counted as of September 1, $12.1+56.6=68.7$, agrees almost perfectly with the total boll "count" of 68.8 at the end of the season. This suggests that a count of both small and large bolls is all that is needed as of September 1 to estimate the total boll count at the end of the season. An additional observation is that the count of 56.6 large bolls as of September 1 is larger than the sum of the small and large bolls counted as of August 1; some of these large bolls developed from squares counted as of August 1.

To formulate a mathematical expression of these relationships, let X_1, Y_1, and Z_1 represent August 1 counts of squares, blooms plus small bolls, and large bolls: Y_2 and Z_2 the September 1 counts of blooms plus small and large bolls, and Z_3 the total boll count at the end of the season.

The September 1 count of blooms plus small bolls may be regarded as the August 1 count, Y_1, plus an unknown fraction of the August 1 square count, minus an unknown fraction of Y_1 which developed into large bolls between August 1 and September 1:

$$Y_1+a_1X_1-bY_1=Y_2 \quad . \quad . \quad . \qquad (1)$$

The September 1 count of large bolls contains the large bolls counted as of August 1, plus an unknown fraction of the August 1 square count, plus an unknown fraction of the August 1 blooms, plus small bolls. This last component is the same quantity, bY_1, that appears in the preceding equation. The relationship is

$$Z_1+a_2X_1+bY_1=Z_2 \quad . \quad . \quad . \qquad (2)$$

It was pointed out earlier in this article that the total boll "count" at the end of the season is almost exactly equal to Y_2+Z_2. But to complete the picture, let that count be represented by the large bolls counted September 1, plus an unknown fraction of blooms and small bolls counted September 1. It is also assumed that the fraction of blooms and small bolls maturing to large bolls between September 1 and harvest is equal to the fraction maturing between August 1 and September 1. That is,

$$Z_2+bY_2=Z_3 \quad . \quad . \quad . \qquad (3)$$

Substituting the observed data for the variables in equations (1), (2), and (3):

$$28.7+78.5\ a_1-28.7\ b=12.1$$
$$22.6+78.5\ a_2+28.7\ b=56.6$$
$$56.6+12.1\ b=68.8$$

The fractions, a_1, a_2, and b, can be evaluated from these equations. But as it is clear from the third equation that b=1.0 almost exactly, there is little point in making an exact solution. This value of b could also be deduced on logical grounds alone because less than a month is required for blooms and small bolls to reach the large-boll stage. Taking b=1.0 gives

$$a_1=0.154$$
$$a_2=0.068$$

This means that 15.4 percent of the August 1 squares become blooms or small bolls by September 1 and another 6.8 percent of the August 1 squares become large bolls by September 1. Furthermore, all of the blooms, and the small and large bolls counted as of September 1 appear to be in the picture as mature cotton or unopened bolls at the end of the season.

These relationships permit experimental objective yield forecasts to be made from August 1 and September 1 fruit counts during the 1955 crop season. On August 1 the following equation may be used:

$$Z_3=0.222X_1+Y_1+Z_1 \quad . \quad . \quad . \qquad (4)$$

This provides a forecast of total bolls per 10 feet of row at the end of the season. In terms of pounds of lint per acre, assuming 37 pounds of lint per 100 pounds of seed cotton, and assuming 1 pound of seed cotton per 100 bolls, the yield per acre, unadjusted for normal losses, would be $4.67Z_3$.

On September 1 the forecast of Z_3 is simply

$$Z_3=Y_2+Z_2 \quad . \quad . \quad . \qquad (5)$$

This forecast is also in terms of number of bolls per 10 feet of row; it must be multiplied by 4.67 to convert it into pounds of lint per acre.

A Basis for Forecasting Yields in 1955

Three distinct approaches that utilize fruit counts on August 1 and September 1 have been described. Each provides a basis for forecasting cotton yields. All the models are similar in that they estimate or predict the number of mature bolls to be produced as the first step; this number is multiplied by an average weight of seed cotton per boll to give the yield for the sample plot or a given fraction of an acre. As mentioned earlier, the multiple regression approach may not provide very stable estimates of the net regression coefficients or a basis for determining between-year coefficients. For this reason little reliance will be placed on this approach in 1955.

The other two models are preferred as a basis for predicting total mature fruit because they conform more closely to the known behavior of the fruiting habits of the cotton plant. Any forecast of yield based on fruit counted as of a given date, however, will require an allowance for harvesting loss and for failure of bolls to open. During the 1954 season, losses from these combined sources amounted to 9 percent.

The behavior of this deduction from year to year is not known—at present there is no basis for assuming that the 1954 deduction represents the usual situation or that it is either larger or smaller than usual. In absolute terms, such losses have been found, in general, to be related to the level of yield. Therefore, it is hoped that the assumption of a constant fraction or a proportional allowance for harvesting losses and unopened bolls may serve as a good first approximation. The results so far suggest that detailed plant observations show much promise as a tool for making forecasts of yields.