35th SESSION OF THE INTERNATIONAL STATISTICAL INSTITUTE

FORECAST AND ESTIMATES OF CROP YIELDS FROM PLANT MEASUREMENTS

by

EARL E. HOUSEMAN *and* HAROLD F. HUDDLESTON

Statistical Reporting Service,
United States Department of Agriculture, Washington, D. C.

Beograd, 1965

14

FORECASTS AND ESTIMATES OF CROP YIELDS FROM PLANT MEASUREMENTS

by Earl E. Houseman and Harold F. Huddleston
Statistical Reporting Service,
United States Department of Agriculture
Washington, D. C.

There are two well-known sources of information for forecasting or estimating crop yields: (1) Farmer's reports on crop conditions or amounts harvested and (2) counts and measurements in sample plots within a sample of fields. This paper presents a discussion of progress and recent experience in the development of procedures and models for forecasting or estimating crop yields from plant counts and measurements. The principal crops on which work has been done by the Statistical Reporting Service include cotton, corn, soybeans, wheat, tobacco, oranges, lemons, peaches, pears, sour cherries, walnuts, pecans, filberts, and almonds.

In making plant measurements, there are three different periods of plant growth that need to be considered because each poses distinctly different problems. The first is the period of growth up to the time when all of the fruit[1/] has been set or the time when, if any additional fruit is set, the probability of it contributing to the yield is zero for practical purposes. The next period extends from the date all fruit is set to maturity or close to the harvest date. The third is a short period just prior to harvest when the problem is principally one of estimating the yield of the crop. Corresponding to the three time periods, the

[1/] In this paper "fruit" is used in a botanical sense and includes buds, blooms and other developing parts that have potential for contributing to the fruitage, that is, the product for harvest.

62

terms "early season forecasts," "late season forecasts," and "preharvest estimates" will be used.

Each June a general purpose agricultural survey, based on probability area sampling, is conducted. This survey provides information on acreages planted to various crops, livestock numbers, and other items. As farms are visited, all fields in the area sampling units are identified and the kind of crop and the acreage in each are ascertained. Hence, this survey provides a sampling frame for the work on crop yields. A subsample of the fields is then selected with probability proportional to acreage, giving a sample of fields in which plant measurements are taken. For some tree crops ,special tree censuses provide the frame for selecting a sample of fields. However, within each sample field two plots or two sample trees are selected at random, marked, and identified, then revisited periodically during the growing season, including harvesting of the plots when mature.

For the 1965 crop season, the Statistical Reporting Service expects to have in operation a program of preharvest sampling on four leading field crops as summarized in table 1. This program has been developing over a period of years.

For tree crops, preharvest sampling is not being done as the major interest is in forecasts several weeks prior to harvest. With reference to timing, estimates from preharvest sampling offer little advantage over estimates based on growers' reports on amounts harvested. The amount of some crops left unpicked as a result of selective harvesting may vary considerably from year to year.

Table 1.- Plans for Preharvest Sampling in 1965

Crop	Number of sample fields	Approximate size of plot in acres*	Approximate size of population		Standard error of estimated yield per acre
			Acres in millions	Percent of U.S. total	
Winter wheat	2,300	.0001	31.4	91	.25 bu.
Corn	3,300	.0023	54.5	95	.70 bu.
Soybeans	1,900	.0004	27.2	95	.30 bu.
Cotton	2,600	.0015	13.9	97	7.50 lb.

*Two plots are selected in each field.

Preharvest Estimates

As already indicated, the problem of preharvest estimates is essentially one of sampling and estimation, not forecasting. A minimum of about three years should be allowed to develop and implement an operating program of yield estimates for a crop from preharvest observations. In fact, if the goal is to have a successful preharvest survey on an operational basis during the third year, a well-planned, intensive effort by experienced mathematical statisticians in this line of work is needed.

Typically, the first year's effort would be limited to a very small number of fields to obtain preliminary measures of variability for establishing size of plots and other aspects of sample design, and to develop operating instructions for a pilot survey the next year. Alternative techniques of measuring the yield on small plots would be tried. This would include consideration of various means of locating sample plots objectively, and ascertaining the advantages of alternative instruments or equipment. Potential sources of error or bias would be identified and means of control considered. In addition, for developing means of estimating harvesting losses, sample plots should be gleaned after harvest. Thus the goal of the first year's effort is to develop, as fully as possible for trial the following year, sound, detailed operating specifications including training plans and a well-designed plan for measuring the quality of the work done.

The second year's effort could be regarded as an intensive and extensive pilot operation using a sample that might be one-fifth or one-fourth the size anticipated for a fully operational program. From the second year's experience much better information should become available on variance components and time requirements for various parts of the job so the sample design can be optimized. Quality checks on the field work should provide a basis for improvement of field procedures which must be rigorous and tightly controlled.

Experience has indicated that preharvest yield estimates (adjusted for harvesting losses) are likely to be on a different level than estimates derived from reports from farmers. Which is correct, if either? Since potential biases may be inherent in the procedures, it is important that provision be made for ascertaining the validity of the preharvest sampling and estimating techniques. The probability of selection of each plot is very small so an unusual amount of attention must be given to avoidance of non-random errors. Field workers may not be completely objective in the process of locating sample plots. Or, if plots are subsampled for certain characteristics there may be opportunity for bias in the techniques of subsampling. Also, instances have occurred where the definition of the fruit to be harvested has been replaced by a worker's own personal definition or interpretation, which resulted in fruit being harvested from plots having biased size or weight characteristics.

There are various ways of getting a valid independent check, depending upon the crop. Take corn as an example. Farmers generally do not have weight measurements of the amount harvested and often have only approximate measures on a volume basis. To obtain a good independent check, special arrangements might be made with selected farmers for getting the total weight and other relevant measurements, such as weight or size per fruit, for the entire crop harvested from particular fields. Sample plots in these fields should be selected and harvested using identical field procedures. The number of plots would need to be large enough to give estimates having low sampling error so any appreciable bias can be detected. Adjustments for such factors as differences in moisture percentage at the time of the preharvest sampling and the time of harvest may be necessary. Also, when comparing yield estimates and actual yield from the entire harvested field one should be on the alert for inconsistencies in concepts of acreage. One of the problems is making sure that the boundaries of the land from which sample plots are selected coincide with what a farmer regards as the acreage in a field.

For purposes of sampling, it is often useful to treat yield as the product of factors such as yield per plant and number of plants per acre, or in the case of cotton, for example, as the product of weight per boll, the number of bolls per plant, and the number of plants per acre. Some factors are simple and inexpensive to measure, while others may be time consuming or difficult to measure accurately. A good example is the counting of cotton plants in contrast to picking cotton or counting the fruit on a cotton plant. An optimum sampling plan considering time and variance components may call for counting of all plants in a two-row plot 20 feet long, whereas observations such as detailed fruit counts might be limited to only a few plants in a plot.

Matters of sampling design could be discussed at length; however, the importance of a balanced effort giving rigorous, tightly controlled procedures regarding all important sources of error is being stressed. Experience has indicated that inherent biases can be eliminated or controlled effectively by intensive training of the field staff, close supervision, quality checks, and providing clear, concise, well-defined field procedures, but astute observation is essential for the identification and control of factors affecting the quality of results.

Some advantages of preharvest sampling. Estimates derived from preharvest sampling are available earlier than estimates from postharvest farmers' reports. Prior to harvest, a farmer can report only his appraisal of the crop prospects. On the other hand, estimates based on preharvest sampling must be based on average harvesting losses or delayed until such time as harvesting loss can be determined from gleaning sample plots after harvest.

In addition to the time advantage just mentioned and the objectivity of the estimates owing to the techniques involved, preharvest sampling provides a means for getting much valuable information that cannot otherwise be easily obtained. Via laboratory analysis of samples taken from fields, information on various attributes of crop quality can be made available. Crop quality, components of yield, and harvesting losses can be related to varieties, cultural practices, weather, and other factors to get a good picture of the variables influencing yield and quality. Also, if deemed worthwhile, information on some types of insect damage, such as the number of ears of corn damaged by corn ear worms, can be readily obtained.

Late Season Forecasts

Forecasting the yield of a crop at periodic intervals during a growing season is obviously much more difficult than estimating yield at time of harvest. It is necessary to discover plant characteristics which may be used to predict components of yield. Forecast formulas must be based upon observable plant characteristics and a comprehensive knowledge of the fruiting behavior of the crop. The formulas must translate plant characteristics observed on any date into accurate forecasts. In contrast to the development of a program for preharvest sampling, any time schedule for developing and perfecting forecasting procedures is much more tenuous. A major reason for this is the necessity of having "between years experience" for the formulation and testing of models. In fact, one may continue to use more than one model for a particular crop after a forecasting program becomes operational in order to give the most promising alternatives a longer time test.

For purposes of this discussion, "late season" begins when all fruit has been set. Thus, the problem can be confined to estimating the number of fruit present and predicting the droppage and the sizing. In other words, the problem is that of predicting the survival of fruit in terms of number of fruit (ears of corn, cotton bolls, oranges, etc.) per acre and the average weight or size of fruit at time of harvest.

Prediction of number of fruit. It is known that the probability of survival is related to maturity of the fruit, which suggests a simple model as follows:

$$N = \Sigma p_i N_i$$

where N_i = number of fruit per plot in the ith maturity category

p_i = probability that a fruit in the ith maturity category will survive and contribute to the fruitage

N = estimated number of fruit that will be on the plants per plot at the time of harvest

The probability, p_i, is a function of time, that is, the probability of survival for a small cotton boll on the 15th of August, for example, is not the same as the probability for a small boll on the 15th of September. Incidentally, our general experience suggests, at least for some crops, that an index of the crop's stage of development may provide a better time reference than calendar date. More will be said on that point a little later.

The problem of defining maturity classes differs widely among the various crops. Cotton, for example, has clearly demarcated stages. A trained observer can accurately classify the fruit. On the other hand, the demarcation of maturity categories for ears of corn is more tenuous. Consequently, a major, skilled effort is required to establish standards, training, and supervisory procedures for achieving uniformity of classification among field observers and between years.

To obtain adequate information on the probabilities of survival, observations need to be taken at frequent intervals during the fruiting period for several years. This can be done by noting the disappearance of fruit that has been "tagged" by maturity categories provided the method of tagging does not affect the probability of survival. If rates of survival are found to differ substantially from year to year, a search for means of adjusting the rates from year to year on the basis of environmental observations, varietal changes, or other relevant factors may be called for.

Take cotton as an example. After several years of experience, good information by maturity categories became available on rates of survival, that is, the fraction of the fruit that would contribute to the fruitage. In Fig. 1, the average rate of survival in relation to the stage of development of the crop is shown for squares (buds), which is one of the fruit maturity categories used in the forecasting model. The more advanced a crop is at time of observation the lower the probability that a "square" will contribute to the yield. The stage of crop development is measured by an index which is the ratio of large bolls to all bolls in the sample plots. From similar relationships for other kinds of fruit categories, values for p_i in the forecasting model are obtained.

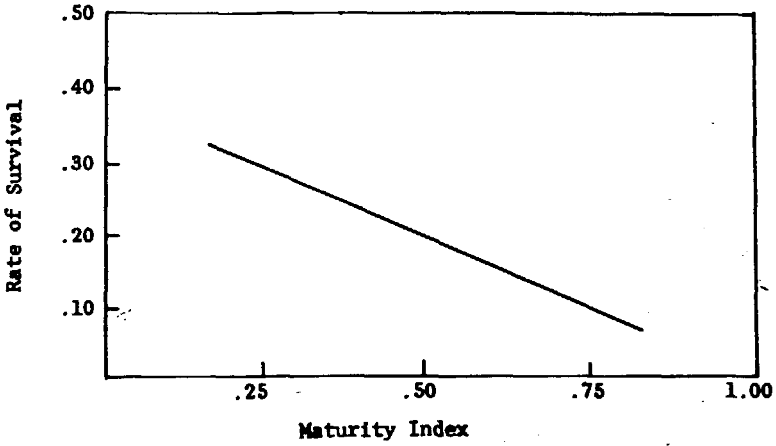

Fig. 1 Survival of Cotton Squares (Buds)

For some crops the counting of fruit by maturity categories may not be necessary. The orange crop is a good example. The forecast of the number of fruit at harvest is simply the product of the present fruit count and the probability of survival which is a function of time. Taking corn as another example, after corn ears have silked, there is practically no disappearance of ears. Hence the forecasted number of ears at harvest is the same as the present count. However, the ears may be classified into maturity categories for purposes of forecasting ear size at harvest.

Average weight or size of fruit. As in forecasting the number of fruit at time of harvest, intensive study of growth patterns of a crop is needed to develop reliable means of predicting the average weight or size of fruit at time of harvest. Study of the growth of citrus, for example, has revealed that the relative increase in size of fruit between September 1 and harvest is nearly constant from year to year. The growth pattern follows a logarithmic curve which provides a good basis for prediction provided stage of maturity, not just calendar date, is taken into account. Projected estimates of fruit size and number of fruit at time of harvest are converted to number of boxes, using information obtained from packinghouses to establish the relationship between fruit size and number of fruit per box. Thus the yield forecasts are expressed as number of boxes.

With regard to corn, two models are presently being used to forecast ear kernel weight at harvest. One is a relationship of harvest weight of kernels per ear (adjusted to 15 percent moisture) to dry weight and total weight at time of observation. In this case, the ratio of dry kernel weight to total kernel weight observed provides a built-in, continuous type crop maturity index for forecasting harvest weight per ear from observed dry matter in the kernels. The other model entails separate predictions by ear maturity categories from the length of the ears at time of observation.

Error of forecast. For the forecasting of crop yields from plant measurements, the Statistical Reporting Service uses essentially the same sample plots that are used for preharvest sampling, table 1. The measurements taken may differ considerably from one date to another during the season. However, tables 2 and 3 give a partial summary of the degree of success with forecasting. Regarding terms of reference used in this paper, the September forecast of cotton is a late season forecast, whereas the August cotton forecast and the May and June forecasts of wheat are early season forecasts.

Table 2.- Forecast,and Sampling Errors for Cotton[1]

Item	Year	August	September	Preharvest	
		FE[2]	FE[2]	Mean	SE[3]
Large bolls (No. per plot)	1963	5.6	3.0	415	6.2
	1964	6.1	3.5	428	6.6
Boll weight (grams per boll)	1963	.025	.024	5.05	.037
	1964	.032	.029	4.88	.042
Gross yield (lbs. lint per acre)	1963	7.8	4.8	570	9.8
	1964	9.2	5.9	588	10.3

[1] Data are from a sample of 1,200 fields representing a region comprised of 5 States.
[2] Forecast error, see text.
[3] Sampling standard error of the mean.

Table 3.- Forecast and Sampling Errors for Wheat[1]

Item	Year	May	June	Preharvest	
		FE[2]	FE[2]	Mean	SE[3]
Number of heads per plot	1962	4.9	5.2	334	4.9
	1963	4.0	2.2	314	2.9
Weight per head	1962	.008	.008	.466	.004
	1963	.008	.007	.522	.004
Gross yield (bushels per acre)	1962	.655	.661	27.9	.42
	1963	.582	.442	28.4	.30

[1] Data from a sample of 900 fields representing 9 States.
[2] Forecast error, see text.
[3] Sampling standard error of the mean.

The errors of forecast shown in these tables are root mean square errors computed as follows:

$$FE = \sqrt{\Sigma\ (y_i - \hat{y}_i)^2/n}$$

where $\hat{y}_i$ = the actual yield, or component of yield, for the two plots in the ith sample field at harvest,

y_i = the corresponding forecast of yield or a component,

n = the number of sample fields, and

FE = error of forecast

The errors of forecast do not include variability associated with selection of fields and of plots within fields. They reflect only between fields within years variability of the forecast component of error. Data for additional years are needed before between years' error of forecast can be adequately measured.

Early Season Forecasts

Research work on early season forecasting from plant measurements has been less extensive than for late season. For tree crops the duration of "late season" is quite long and "early season" forecasts have not been attempted. Cotton, wheat, corn, and soybeans have received the most attention in the development of early season forecast models.

Growth patterns among different plant species are so varied that not much can be said about a general approach for finding a forecasting model. The nature of the problem obviously changes rapidly with the stage of development. An important aid in developing realistic early season forecasting models is the availability or securing of data weekly on fruiting and plant characteristics starting in advance of the first forecast date up to harvest. However, the use of such detailed data from isolated studies poses problems in statistical inference, among which are: (1) The construction of the models, (2) a logical translation of the model into observable characteristics in surveys at less frequent intervals, (3) development of constants or tentative parameters for model which will apply to data observed for a specific forecast date, and (4) relationships for selective or non-probability samples may differ in unexpected ways. For crops which fruit over a relatively long period or have many fruit per plant, a "fruiting model" may be developed based on classifying fruit, or classifying fields where some fields will have fruit and other fields show no fruit present. For a crop such as cotton, many fields (or even plants within fields) will have "squares," bloom, and bolls, while other fields in a nearby area may have only squares or even no fruit observable.

Where part of the fruit has been set, one approach is to add a term for additional fruit expected at harvest from fruit not set to the model discussed on page 8 of the previous section. For example, for the August 1 forecast of cotton, the relationship between "the number of cotton bolls at harvest from fruit not set" and a maturity index can be used--the maturity index being the ratio of large bolls to all bolls at the time of observation. To establish this relationship, fruit set at time of observation must be "tagged" so bolls at harvest from fruit not set can be counted. Incidentally, another type of model for early cotton forecasts called "the rate of fruiting" model has been developed. This model is more complex and will not be discussed here.

For wheat, the May forecast of number of heads is predicted from stalk counts using a relationship established from historical data. Weight of grain per head is related to plant density. Hence head weight is adjusted for plant density rather than using the average for several years.

It appears that an historical average weight per fruit may be a satisfactory basis for a forecast when there is control of cultural practices such as irrigation and the thinning of tree fruits so the density varies little from year to year. An historical average weight may also be satisfactory if the forecast is for a large area, say several States, so the average environment for the whole area is about the same from year to year even though the environment for any given small locality may vary considerably from year to year.

- - - -

SUMMARY

Progress and experience in the development of models and procedures for forecasting and estimating crop yields from plant measurements are discussed in this paper. Three time periods are of importance: (1) A short period prior to harvest when the problem is one of sampling and estimation rather than forecasting, (2) the period after all fruit has been set when the forecasting problem is that of predicting survival of fruit already set and of predicting weight per fruit at harvest, and (3) the growth period prior to the date when all fruit has been set. General experience from work on several crops is presented rather than a detailed report for one or two crops.

RESUME

Dans ce rapport le progrès et l'expérience dans le développement des modes et procédés pour prévoir et estimer les rendements de la cueillette par action de mesurage des plantes sont discutés. Trois phases importantes sont à noter: (1) Une courte période avant la cueillette quand le problème consiste d'échantillonner et d'estimer le fruit plutôt que de le prévoir; (2) La période après que le fruit a été planté c.à.d. lorsque le problème consiste encore de prédire la survivance du fruit déjà planté ainsi que son poids au tempts de la cueillette; (3) La période de croissance avant la date que tout le fruit a été planté. L'expérience générale acquise de plusieurs cueillettes est presentée ici plutôt qu'un rapport détaillé d'une ou de deux cueillettes.