Michigan
Crop
Reporting
Service
Michigan Dept. of Agriculture
U.S. Department of Agriculture

MICHIGAN DRY BEANS 1976 RESEARCH REPORT

M I C H I G A N D R Y B E A N S

1 9 7 6 R E S E A R C H R E P O R T

MICHIGAN CROP REPORTING SERVICE
201 FEDERAL BUILDING -- P.O. BOX 20008 -- LANSING, MICHIGAN 48901

Telephone (517) 373-9020

CHARLES A. HINES . . . . . Statistician in Charge
STEVE J. PSCODNA . . . . Assistant Statistician in Charge

Designed
Supervised and Analyzed
by
GERALD J. LANGLEY
Mathematical Statistician

Funding was provided by the Michigan Department of Agriculture and Research Division, SRS, Washington, D.C.

Technical support was provided by the Methods Staff, E.D. and Research and Development Branch, R.D., SRS, Washington, D.C.

CRSB-77-01

CONTENTS

INTRODUCTION

In the late 60's Michigan Agriculturalists expressed an interest in developing a dry bean objective yield program for the state. This interest resulted in four years of data collection by the Michigan State Statistical Office (SSO) from 1969 through 1972 and the present 1976 research. This year's research was based partly on the analysis of the 1972 data which gave indications that plant characteristic counts would produce the best method to forecast yield. Using this and other background information from the potato and soybean objective yield programs, the 1976 research was designed to satisfy four main objectives:

1. Develop a probabilistic technique for estimating statewide bean acreage.
2. Determine an optimal sampling design for a preharvest biological yield estimate.
3. Explore models for a "September 1" forecast.
4. Determine if operational procedures could be developed for a full scale survey in 1977.

In July 1976 a multiple frame survey to estimate acreage was conducted using a list frame, developed for this purpose, and the June Enumerative Survey area frame. Using telephone and field enumerators to follow-up on mailed questionnaires, a total of 700 bean producers were contacted over a 12 day period.

Thirty objective yield samples were randomly selected from the multiple frame sample to study the possibility of forecasting and estimating yield. Nine of the original 30 objective yield samples were either refusals or had no navy beans. Data from the remaining 21 randomly selected samples was collected by 6 enumerators, working in pairs. They made a "September 1" visit in late August to lay out the units and make counts. If at this time the bean field was not ready for harvest, enumerators made another visit in September when pods were harvested and mailed to the laboratory. Gleaning of harvest losses and post harvest interviews were conducted on even numbered samples.

I. ACREAGE ESTIMATION

A check against non-overlap operators on the June Enumerative Survey (JES) area frame showed the list frame to be about 51 percent complete. The non-overlap operators expanded acreage, adjusted by the "July Update" bean acreage ratios, was used to estimate the acreage not represented in the list frame. The list frame was stratified by reported 1975 acreage into 8 acreage categories (Table I). The first seven strata were randomly sampled using sample sizes obtained from the following formula:

$$n_h = \left[S_h \cdot N_h \Big/ \sum_h (S_h \cdot N_h) \right] n$$

where N_h = the number of operations in stratum h.

S_h = standard deviation of 1975 bean acreage for stratum h.

n_h = sample size for stratum h.

n = total number of samples from strata 1 through 7.

Sample allocation according to the above formula was not strictly adhered to, however. It was decided that the higher strata should be sampled heavier than the equation indicated. The sample sizes in Table I reflect this decision.

Table I

Acreage Statistics from 1975 Dry Bean List

Stratum Index	Acreage range	Standard deviation(S_h)	Mean acreage ($\bar{X}_h$)	Number of operations(N_h)	($N_h \cdot S_h$)	Sample size(n_h)
1	1 - 45	11.2299	24.867	1,368	15,362.5	70
2	46 - 80	10.6937	61.266	640	6,844.0	56
3	81 - 115	7.3969	98.325	369	2,729.5	30
4	116 - 175	16.1588	138.613	266	4,298.2	35
5	176 - 290	28.0929	216.511	178	5,000.5	76
6	291 - 400	41.7536	340.931	58	2,421.7	35
7	401 -	112.9867	529.500	24	2,711.7	24
8	Blank & 0	--	0	1,884	--	374
Total	--	--	--	4,787	--	700

Of the 700 operations sampled, 75 were refusals or inaccessible. Results of the 1976 survey are summarized in Table II.

Table II

Summarization of the 1976 Acreage Survey

Stratum Index	Mean acreage ($\overline{X}_h$)	Standard deviation(S_h)	Expanded acreage (Y_h)	Relative error of Y_h (%)
1	38.219	74.31	52,283.25	24.30
2	60.889	46.65	38,968.89	10.20
3	82.720	42.41	30,523.68	10.25
4	101.094	61.98	26,890.94	10.84
5	187.685	96.48	33,407.92	6.02
6	248.444	153.81	14,409.78	11.90
7	423.810	269.67	10,171.43	13.89
8	40.833	61.24	76,929.05	8.27
Total	--	--	283,584.93	5.52

The non-overlap acreage was estimated at 284,727 with a relative error of 12.1 percent, whereas the acreage represented by the list frame was 283,585 with a relative error of 5.52 percent. These combined to give an estimate of 568,312 planted acres for the state with a relative error of 6.76 percent. Since the relative error of the "list frame acreage" estimate is reasonable, the sample size of 700 operations is recommended for future use.

Improvements in the multiple frame estimate should be sought by first building the list frame to an acceptable level of completion, 80 percent or more. This would reduce the relative error of the state acreage estimate by making it less dependent on the small number of bean growers presently enumerated in the JES. Adding area segments from the bean producing region to the JES is a costly procedure and should; therefore, be viewed as a last alternative. However, if in the future the building of the dry bean list frame becomes too expensive, the possibility of adding area segments to the JES should be investigated.

II. YIELD RESEARCH

A. Sample Selection

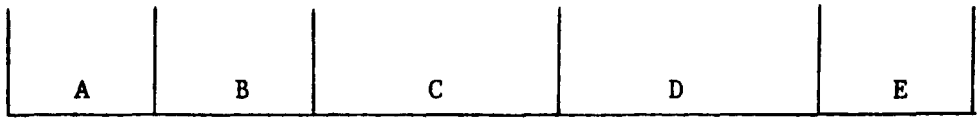
Due to monetary and workload considerations, the 1976 objective yield research was limited to a sample size of 30. Since the list frame sample was used to represent 51 percent of the state bean acreage and the non-overlap 49 percent, it was decided that 15 samples would be randomly selected from each source. The selection was proportional to expanded navy bean acreage. One large field from a non-overlap operation was selected twice; therefore, two samples were laid out in that field.

B. Data Collection

Data was collected in August and September 1976 from 21 randomly selected navy bean fields in Michigan. Nine of the original 30 were either refusals or had no navy beans. Each sample contained 4 randomly located units, with each unit consisting of two adjacent 42 inch rows. The rows were subsequently divided into five sections -- the first two sections (A & B) being 6" in length, followed by two 12" sections (C & D) and a final 6" section (E) - (Figure I).

Figure I

Row Section Designations



Of the 21 samples, 7 were not suitable for forecast model development, either because they were harvested too early or because of improper enumeration.

Data collected for the 14 other samples included:

for all sections -- plant counts

for 6" sections -- number of lateral branches
number of racemes
number of small pods ($\leq$ 1" long)
number of large green pods ($>$ 1" long)
number of mature yellow pods.

Approximately 3 days prior to when farmers harvested each of the sample fields, enumerators made check counts and harvested, by section, all 4 units of each sample. The pods were then mailed to the Michigan SSO laboratory. In the laboratory, pods with beans were counted and weighed. After separating the beans and pods, beans were weighed and then tested for moisture content.

Within three days after harvest, post harvest interviews and gleanings of harvest losses were performed on the even numbered samples. Six of the nine "lost" samples were even numbered, leaving only nine samples with which to estimate harvest loss.

C. Preharvest Analysis

The optimal solution for three-stage sampling is found by minimizing the "cost x variance" equation. The following cost and variance equations were used:

1. $\text{Cost}\ (k) = n_1\ c_1 + (n_1\ n_2)\ c_2 + (n_1\ n_2\ n_3)\ c_{3,k}$

2. $$\sigma^2_{\bar{\bar{y}}} = \frac{\sigma^2_{1,k}}{n_1} + \frac{\sigma^2_{2,k}}{n_1\ n_2} + \frac{\sigma^2_{3,k}}{n_1\ n_2\ n_3}$$

For optimum allocation, the following notation is used:

n_1, n_2, n_3 = number of samples, units, and rows, respectively.

C_1 = between sample cost.

C_2 = between unit cost.

$C_{3,k}$ = per row cost for row length k.

$S^2_{1,k}$ = estimate of the between sample variance component using rows of length k.

$S^2_{2,k}$ = estimate of the between unit within sample variance component using rows of length k.

$S^2_{3,k}$ = estimate of the between row within unit variance component using rows of length k.

$S^2_{\bar{\bar{y}}}$ = estimate of the variance of the sample mean ($\bar{\bar{y}}$).

The optimum number of units per sample and the optimum number of rows per unit for a row of length k are estimated by:

3. $$\text{Opt}\ (n_2) = \sqrt{\frac{c_1}{c_2}\ \frac{S^2_{2,k}}{S^2_{1,k}}}$$

4. $$\text{Opt}\ (n_3) = \sqrt{\frac{c_2}{c_{3,k}}\ \frac{S^2_{3,k}}{S^2_{2,k}}}$$

To estimate the variance components, nested analysis of variance (NANOVA) was performed on the estimate of hundredweight per acre for the seven row lengths (6" through 42"). The different sections and combinations of sections on which the NANOVA was run are outlined in Table III with reference to the letter designations as shown in Figure I.

Table III

NANOVA Row Length Designations Tested

Row Length in Inches						
6	12	18	24	30	36	42
A	A & B	B & C	A & B & C	B & C & D	A & B & C & D	(All Five Sections)
B	C	D & E	C & D	C & D & E	B & C & D & E	
E	D					

Time studies done in conjunction with the 1976 data collection produced the following cost figures for preharvest work (all in terms of minutes):

$$c_1 = 90, \quad c_2 = 10, \quad c_{3,k} = g(k), \quad \text{where } g(k) = k$$

Between sample cost includes the average amount of time enumerators spent traveling to the next sample, conducting the initial interview, and getting to the sample field's starting corner. Also included is the average mileage between samples, converted to minutes of enumerator time. Between unit cost is the average time between the ending time of one unit within a field to the time counts start on the next unit in the same field. Per row cost for the preharvest work was found to be one minute per inch.

Solving for opt (n_2) and opt (n_3), using the 15 sets of variance components along with the cost figures on page 9 and then averaging by row length, produced Table IV.

Table IV

Optimum Allocation by Equations 3 and 4

	6	12	18	24	30	36	42
Opt (n_2)	4.451	4.596	4.288	4.199	4.175	4.241	4.321
Opt (n_3)	2.579	.893	.567	.443	.351	.282	.233

The figures in the above table are somewhat misleading for row lengths 12 through 42 inches, i.e. where opt (n_3)<1. The numbers of rows per unit must, of course, be at least one. A better estimate of the optimum number of units per sample, opt (n_2), for when opt (n_3)<1, is found by setting $n_3 = 1$ in the cost x variance equation before solving for n_2. Table V summarizes the resulting opt (n_2) for row lengths 12 to 42.

Table V

Opt (n_2) with n_3 Fixed = 1

	12	18	24	30	36	42
Opt (n_2)	4.311	3.218	2.759	2.444	2.243	2.099

The optimal integral solutions for allocation are as given in Table VI.

Table VI

Integral Solution for n_2 and n_3

	6	12	18	24	30	36	42
Opt (n_2)	4	4	3	3	2	2	2
Opt (n_1)	3	1	1	1	1	1	1

As with most survey design research done in SRS a desired variance for the sample mean ($\sigma^2_{\bar{y}}$) is specified and the most efficient way to obtain this is then determined. The necessary n_1, nec (n_1), for each row length (Table VII) is derived by using the opt (n_2)'s and opt (n_3)'s in Table VI and a fixed variance.

$$\text{nec } (n_1) = \left(S_{1,k} + S_{2,k} \Big/ n_2 + S_{3,k} \Big/ n_2 n_3 \right) \Big/ \sigma^2_{\bar{y}}$$

Table VII

Sample Sizes Needed (by Row Lengths) so that $S^2_{\bar{y}} \leq 0.22$

	6	12	18	24	30	36	42
nec (n_1)	56.43	55.71	60.71	54.29	69.29	66.43	63.57

To determine the "best" row length we use the appropriate nec (n_1), opt (n_2), and opt (n_3) in the cost equation 1, page 8, and then compare the total costs of the different row lengths. These cost figures, in terms of minutes, are displayed in Table VIII.

Table VIII

Best Survey Costs by Row Length

Relative Cost in Minutes	6	12	18	24	30	36	42
	15,755	13,816	14,722	14,400	16,409	16,704	17,150

The 12 inch row length with 4 units per sample and 1 row per unit is the optimum design for estimating hundredweight per acre as indicated by the 1976 data. Assuming the attrition rate due to lost samples, etc. will be similar to the rate in 1976, a sample size of 80 samples is recommended for future surveys. This should result in 50 to 60 useable samples which would produce a C.V. of the estimate of around 5 percent.

A paired t - test was run on the yield estimate of the combined intensive count areas (A + B + E) versus the 12 inch (plant counts only) areas (C + D) to check for the effect of the intensive counts on yield. A value of t = -1.4737 is significant at the .1 level of significance for a one-tailed test. Therefore, it is suggested that the units be divided into three areas, one for preharvest work, one for forecasting counts, and a buffer zone separating these two. At harvest time both the preharvest area and the count area should be picked so that forecast models can be developed using paired observations.

D. Forecasting Analysis

As with the soybean yield forecasting procedures, units within fields were classified into maturity categories. The criteria for the classification of a unit into a certain maturity code is summarized below:

Code 1 -- Pods Still Forming or Earlier

The unit will be classified as "1" until the plant has progressed through the bloom stage. Any pods formed will still be green and units are expected to be in this stage in late July.

Code 2 -- Pods Set, Leaves Still Green

In general, there should be no blooms on the plants except possibly for a late plant in the unit which may have an occasional bloom or two on the top node of the main stem or near the end of a lateral branch. Most of the pods will still be filling and all leaves will still be green.

Code 3 -- Pods Filled, Leaves Turning Yellow

Leaves will be yellowing on nearly all plants, but green leaves may still be more numerous on the plants than yellow or partially yellow leaves. Almost all the pods will be filled and some will be ripening.

Code 4 -- Pods Turning Color, Leaves Shedding

Most leaves will have turned yellow and some leaves will have fallen. The pods will have their full size. Pods will be changing color from green to brown, with less than 10 percent still green.

Code 5 -- Pods Brown, Almost Mature

Almost all pods will be brown and easily opened so the beans can be removed. The beans are white and have shrunk inside the pod. Most of the leaves have been shed by the plants.

Code 6 -- Mature

The pods will be brown and ready to combine. All leaves will have fallen from the plants, except for an occasional late plant in the unit.

There were no units classified as code 1 or 6 on the "September 1" visit. Twenty-seven were classified as code 2, eight as code 3, eighteen as code 4, and three as code 5. Because of the small sample sizes of code 3 and code 5, codes 2 and 3 were combined into code A and codes 4 and 5 were combined into code B. Correlation coefficients of the different plant characteristic counts with the yield estimate from the preharvest data is summarized by maturity codes in Table IX. Distinguishing between racemes and lateral branches is extremely difficult after leaves have begun to shed. Therefore, no counts of racemes and branches were made for codes 4 and 5 (code B).

Table IX

Correlations of Plant Characteristic Counts with the Preharvest Estimate of Yield/Probability of a Greater $|r|$ under $H_0: \rho = 0$

Codes	Branches	Racemes	Racemes + Branches	Small Pods ≤ 1"	Green Pods > 1"	Mature Pods	Large Pods *	All Pods **
A n=35	.734 / .0001	.736 / .0001	.754 / .0001	.09 / .61	.858 / .0001	.053 / .7617	.871 / .0001	.781 / .0001
B n=21	-- / --	-- / --	-- / --	-.166 / .5223	-.061 / .7875	.805 / .0001	.976 / .0001	.973 / .0001
2 n=27	.648 / .0005	.635 / .0006	.655 / .0004	-.079 / .6963	.816 / .0001	.253 / .2013	.816 / .0001	.679 / .0002
3 n=8	.734 / .0597	.927 / .0035	.894 / .0073	.580 / .1713	.865 / .0124	.384 / .6016	.940 / .0023	.939 / .0025
4 n=18	-- / --	-- / --	-- / --	-.155 / .5446	-.038 / .8766	.780 / .0003	.974 / .0001	.970 / .0001
5 n=3	-- / --	-- / --	-- / --	0 / 1	.481 / .6748	.995 / .0652	.992 / .0845	.992 / .0845

* Large pods = mature pods + green pods > 1".
** All pods = large pods + small pods.

The number of large pods is the most highly correlated variable with yield in every maturity category except code 5. It is logical that the number of mature pods would be a better predictor of yield after the unit has shed most of its leaves and hence development has stopped. However, it is significant that the number of large pods stays so highly correlated over such a wide range of bean development. Its usefulness as a forecasting variable should be thoroughly utilized.

To determine the optimum length of the intensive count area, cost analysis was performed using variance component estimates from nested analysis of variance of the number of large pods. Solutions for cost equations, for each area length (6, 12, and 18 inches), can be found after determining the necessary sample sizes to attain a fixed C.V. of the estimate. Note that n_2 and n_3 are now fixed at 4 and 1, respectively.

Time studies on the September 1 visits produced the following cost figures:

$$c_1 = 90, \qquad c_2 = 12.5, \qquad c_{3,k} = g(k) = k$$

Summarization of the forecasting cost estimates in terms of minutes and the necessary sample size for a C.V. of 5 percent are given in Table X.

Table X

September 1 Forecast

Data Collection Costs in Minutes by Row Length and Maturity Code
Necessary Sample Size for Fixed Variance for a C.V. of 5 Percent

Code	6	12	18
A	15416 / 94	9776 / 52	8692 / 41
B	13940 / 85	13724 / 73	14628 / 69

The 12 inch section is the most efficient for code B. For code A, the 18 inch section is slightly better in terms of time, but for the sake of the enumerators, and therefore accuracy and consistency, it may be better to limit the length to 12 inches. Enumerators complained this year of being in the field too long and often the effects of this could be detected in the data. Also, since the minimum number of samples for the preharvest work is 55, the 12 inch section, as far as code A is concerned, would be the most convenient, since the number of samples needed for it is 52. Therefore, it is recommended that count sections be 12 inches long in future years.

The number of large pods is seen as the "best" forecasting variable for both codes A and B. However, in the regression analysis on the earlier stages several other variables, racemes, branches, and small pods, are found in a few significant models. For example, backwards elimination on sections A and B (combined) on the second row of each unit showed small pods to be a valuable addition to large pods at a greater than .0002 level of significance. On the other hand, analysis on the first row shows small pods adding very little. The same type of indications are found for branches and racemes.

The dry bean crop development was unusually early in 1976. In a more "typical" year or in a late year, small pods, branches, and racemes would probably be found to be more important for the September 1 forecast.

For the more mature stages, a forecasting model containing the number of large pods as the only independent variable will produce the best prediction equation. In the regression analysis, (backward, forward, and stepwise model selection procedures) the count of large pods is in every significant equation. Sometimes

all pods or mature pods would appear in combination with large pods in a significant model, but always their regression coefficients were not significantly different from zero. Also, all three variables are so highly correlated with each other that using just the best one in our model will give us practically all the information available from the other two.

Combining large green pods and mature pods for future forecasting work will eliminate a subjective decision process from the enumerators' task. That is, enumerators would not need to decide if this pod is yellow enough to be counted as a mature pod. The classification procedure would be simply to decide (measure) if the pod was greater than or less than one inch -- a much more objective process.

In further research in forecasting, counts should be made on small pods, branches, racemes, and large pods as long as the unit is in maturity code 3 or less (i.e. leaves have not yet begun to fall). For codes 4 and above the count of large pods alone would probably be sufficient. Since mature pods had a higher correlation in code 5 with yield than did large pods (recall the small sample size, n=3), it may be advisable to make a count of large pods and then go over the section again counting mature (yellow-dry) pods. There was insufficient data this year to determine whether or not this procedure would be worthwhile. Future work should try to answer this question.

A summarization of the significant models ($\alpha \leq .2$) produced by regression analysis of 12 inch rows is shown in Table XI. These are displayed only to give an idea as to the general form of the "promising" models. They are <u>not</u> proposed forecasting models. Future forecasting research will be aimed at building a model for each maturity code, 1 through 6. Whereas, the models in Table XI are for codes A and B and would, therefore, be of no use in future model building.

Table XI

Various Significant Models for Codes A and B

Code A	Y = 86.05 + 0.5036 LP - 0.2883 RA
	Y = 80.10 + 0.3898 LP
	Y = 20.94 + 0.3331 LP - 0.6564 SP + 0.4238 RA
	Y = 31.38 + 0.5031 LP - 0.5397 SP
	Y = 6.8 + 0.4117 LP + 0.3781 BR
	Y = 33.02 + 0.4756 LP
	Y = 7.33 + 1.1988 LP - 0.6665 AP
	Y = 25.548 + 0.4360 LP
	Y = 23.41 + 0.5123 LP + 0.8214 BR - 0.4168 RABR
	Y = 37.34 + 0.5132 LP - 0.5625 SP + 0.3945 MP
	Y = 37.198 + 0.5229 LP - 0.582 SP
Code B	Y = -30.72 + 0.5432 LP - 0.0601 GP
	Y = -39.66 + 0.5384 LP
	Y = -26.94 + 0.5377 LP
	Y = -7.25 + 0.5627 LP - 0.8361 PLANTS
	Y = -16.706 + 0.4438 LP + 0.1048 MP
	Y = -59.59 + 0.5801 LP - 0.0782 GP
	Y = -74.74 + 0.5777 LP
	Y = -22.81 + 0.5340 LP
	Y = -17.249 + 0.5502 LP - 0.067 GP

Y = biological yield in grams per 18 square feet.
LP = number of large pods per 18 square feet.
RA = number of racemes per 18 square feet.
SP = number of small pods per 18 square feet.
BR = number of branches per 18 square feet.
AP = number of all pods per 18 square feet.
RABR = number of racemes and branches per 18 square feet.
MP = number of mature pods per 18 square feet.
GP = number of green pods per 18 square feet.
PLANTS = number of plants per 18 square feet.

E. Harvest Loss Estimation

The gleaning samples each consisted of two randomly located rectangular units, 10.0 by 1.5 feet in size. The units were laid out perpendicular to the direction of the bean rows. Beans and pieces of beans were gleaned, packaged, and sent to the Michigan SSO laboratory to determine moisture content and weight. Bean weights for the 15 square feet gleaning areas were expanded and averaged to estimate harvest loss for the state. This years estimate was 1.8 hundredweight per acre with a variance of .094. The coefficient of variation for the estimate was 17 percent. It is recommended that post harvest interviews and gleanings be done on all samples in the future. Using the projected attrition rate of samples and the recommended sample size of the preharvest work, a C.V. for the harvest loss mean of about 7 percent is projected.

SUMMARY

Cost x Variance analysis on preharvest counts and weights of Michigan dry beans indicate a field structure of 4 randomly located 12 inch rows as the optimal sampling design. To provide a mean yield with a five percent coefficient of variation, a sample of about 80 fields would be necessary. The count of pods greater than one inch long is seen as the "best" forecasting variable. In the earlier stages of growth, racemes, branches and/or small pods will probably produce the best models for prediction of yield. Gleaning of harvest losses should be conducted on every sample.

The multiple frame estimate of planted dry bean acreage for the state was 568,312 acres, with 283,585 from the list frame and 284,727 from the non-overlap. This compares to the preliminary published state estimate of 540,000 acres. Improvements in the multiple frame estimate should be sought by building the list frame to an acceptable level of completeness, 80 percent or more.

The estimate of biological yield for the state was 10.7 hundredweight per acre with a variance of .075. Subtracting the harvest loss estimate from 10.7 produces an estimate of harvested yield per acre of 8.9 hundredweight. The preliminary state yield estimate (harvested yield), derived from the standard composite techniques, was 9.3 hundredweight per acre.

.75
.22