RESEARCH NOTE NC-33

NORTH CENTRAL FOREST EXPERIMENT STATION, FOREST SERVICE—U.S. DEPARTMENT OF AGRICULTURE
Folwell Avenue, St. Paul, Minnesota 55101

A COMPUTER PROGRAM FOR PRELIMINARY DATA ANALYSIS

ABSTRACT. — A computer program written in FORTRAN has been designed to summarize data. Class frequencies, means, and standard deviations are printed for as many as 100 independent variables. Cross-classifications of an observed dependent variable and of a dependent variable predicted by a multiple regression equation can also be generated.

A preliminary look at a set of measurements is often desirable before settling on a particular analytical technique. Questions of possible interest at this stage include how well the measurements represent the range of field conditions and what relationships apparently exist between variables. Values predicted by a series of estimating equations might be compared so that the "best" such equation can be chosen.

The NCBREAKS computer program discussed here is designed to summarize distributions so that checks for apparent relationships and rough comparisons can be quickly and easily made. Measurements on each variable of interest are subdivided into several classes. The numbers of observations and their means and standard deviations are then printed in a series of tables for each of these classes.

An illustrative application of the NCBREAKS program is provided by the problem of predicting gross cubic-foot volume growth in natural and planted jack pine stands, given measurements made in 480 sample stands. A standard 80-column Hollerith card is punched with the basic data for each stand: the past gross growth in cubic feet, the number of trees per acre, and a code indicating whether the stand is natural or planted. To determine the spread of the sample stands, stocking classes were established by providing their upper limits, the class "breaks," as program input.

Break specified as program input	*Interval code and no. of trees per acre*
.	1. 0
0	
	2. 1- 200
200	
	3. 201- 400
400	
	4. 401- 600
600	
	5. 601- 850
850	
	6. 851-1100
1100	
	7. 1101-1400
1400	
	8. over 1400

If the data were summarized at this point, all stands would be lumped together. To separate the planted from the natural stands, a special *transgeneration* option of the program was utilized. Variables could be generated through common arithmetic operations on the observed variables; or logarithms, powers, or inverses of the observations could be taken. For the jack pine problem, the additional option of creating "zero-one" dummy variables (as occasionally used in multiple regression equations) was used to separate the stands. The result of this manipulation is that three variables occur in the printed tables: (1) the stocking level of the 480 stands, (2) the stocking level of just the natural stands, and (3) the stocking level of the planted stands.

A frequency table shows the number of each kind of stand by stocking classes (table 1). For example, while 30 natural stands with less than 200 trees per

Table 1. —NUMBER OF OBSERVATIONS ASSOCIATED WITH INTERVALS

INDEPENDENT VARIABLE	INTERVAL 1	INTERVAL 2	INTERVAL 3	INTERVAL 4	INTERVAL 5	INTERVAL 6	INTERVAL 7	INTERVAL 8
1	0	32,	115,	104,	91,	46,	29,	63,
2	125,	30,	101,	87,	53,	27,	13,	44,
3	355,	2,	14,	17,	38,	19,	16,	19,

Table 2. —INTERVAL MEANS OF INDEPENDENT VARIABLES

INDEPENDENT VARIABLE	INTERVAL 1	INTERVAL 2	INTERVAL 3	INTERVAL 4	INTERVAL 5	INTERVAL 6	INTERVAL 7	INTERVAL 8
1	0	,1574E+03	,3025E+03	,4818E+03	,7231E+03	,9742E+03	,1212E+04	,2899E+04
2	0	,1556E+03	,3016E+03	,4845E+03	,7297E+03	,9581E+03	,1206E+04	,2928E+04
3	0	,1850E+03	,3084E+03	,4678E+03	,7139E+03	,9971E+03	,1217E+04	,2832E+04

Table 3. —INTERVAL MEANS OF OBSERVED DEPENDENT VARIABLE

INDEPENDENT VARIABLE	INTERVAL 1	INTERVAL 2	INTERVAL 3	INTERVAL 4	INTERVAL 5	INTERVAL 6	INTERVAL 7	INTERVAL 8
1	0	,5543E+02	,7749E+02	,9464E+02	,9778E+02	,1221E+03	,1371E+03	,1026E+03
2	,1173E+03	,5677E+02	,7963E+02	,9695E+02	,9080E+02	,1009E+03	,1211E+03	,8011E+02
3	,8681E+02	,3525E+02	,6211E+02	,8282E+02	,1075E+03	,1523E+03	,1500E+03	,1546E+03

acre (interval 2) were sampled, that stocking level is represented by only 2 planted stands. Note that the first interval must be carefully interpreted. It is rather awkwardly defined as the number of stands that do not fit under a given variable. There are 125 stands with zero naturally occurring trees, i.e. there are 125 planted stands.[1]

To determine the relationships between stocking levels and gross growths, two sets of tables were generated. The first set includes the means and standard deviations (only the means are shown in table 2) of the independent variables — the stocking levels — for the specified classes. The second set lists the means and standard deviations of the gross growths of the stands in each of these classes (table 3). Combining the three tables tells us, for example, that the 19 planted stands in the highest stocking class (interval 8 of the third variable) average 2,832 trees and an annual gross increment of 155 cubic feet per acre.[2] (Interval 1 states that the 125 not-natural stands averaged about 30 cubic feet per acre more growth than did the 355 not-planted stands.)

The most direct way to consolidate this information is through a graph (fig. 1). While all of the minor irregularities in the 480 stands will not appear, this loss of detail is more than offset by the speed of plotting just the class means. Such smoothed plots can often suggest whether a linear or logarithmic or some other curve form would be appropriate — or, in terms of the present example, whether there is a significant or important difference between the growth-stocking level relationships of planted and natural stands.

Suppose that the single quadratic equation

$$\text{gross growth} = 0.26\ (\text{No. trees/A}) - 0.00013\ (\text{No. trees/A})^2$$

seems to be worth investigating. By entering the appropriate coefficients into the program and transgenerating the required squared term, mean predicted values of gross growth corresponding to the stocking level classes of the independent variables will be generated in the usual tabular form. A rough test of the prediction equation can then be made by plotting the results (fig. 2).

The preceding discussion has emphasized the most common uses of the program. An unlimited number of observations on as many as 100 variables, each divided into appropriate intervals, can be summarized. Any single variable will be cross-classified with

---

1 *For convenience, the stocking classes for all three variables were defined by the same set of breaks. Additional breaks ordinarily would be specified for other stand characteristics, such as site index, that might be examined.*

2 *The symbol E+xx following a number indicates that the number is to be interpreted after moving the decimal point xx places to the right.*

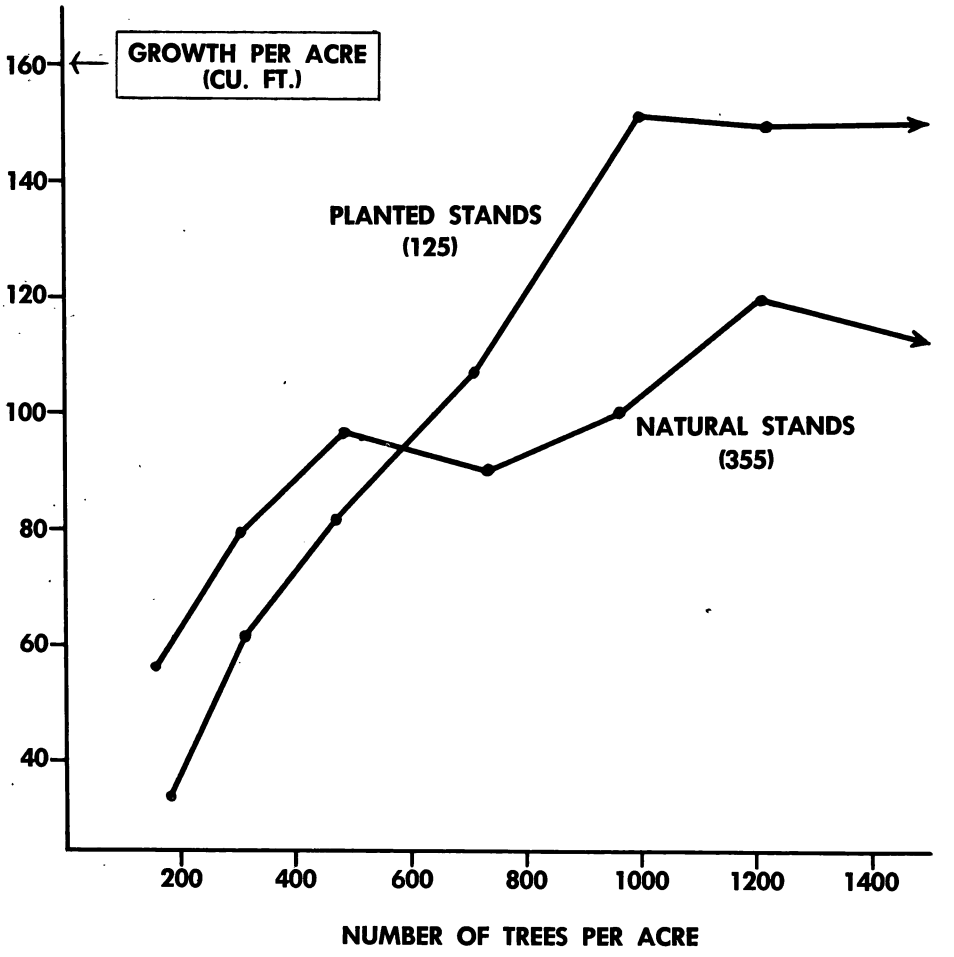

Figure 1.—Observed annual gross growth per acre on 480 jack pine stands by mean stocking per acre.

all others if it is designated as the dependent variable. Entering a constant and coefficients for any subset of the independent variables, perhaps in combination with one of the transgeneration options, results in a similar cross-classification of predicted values. The tables of standard deviations are also useful, for example, to test the reasonableness of an assumption of constant variance. In all these instances, graphs intended to give a preliminary indication of the nature of a set of data can be quickly prepared from the interval means.

The NCBREAKS program was written in FORTRAN 60, specifically for the Control Data Corporation 1604 computer of the Numerical Analysis Center of the University of Minnesota, but is adaptable to most computers with the following characteristics:

(1) A FORTRAN 60 or FORTRAN II or compatible compiler capable of utilizing
   (a) two-dimensional arrays,
   (b) a variable data format (not in source program);
(2) Core storage of 16,000 data words in addition to the compiler storage requirement.

Complete program documentation and a punched copy of the source deck are available from the Biometrical Section of the Station.

DENNIS L. SCHWEITZER
Associate Economist
August 1967

See page 4

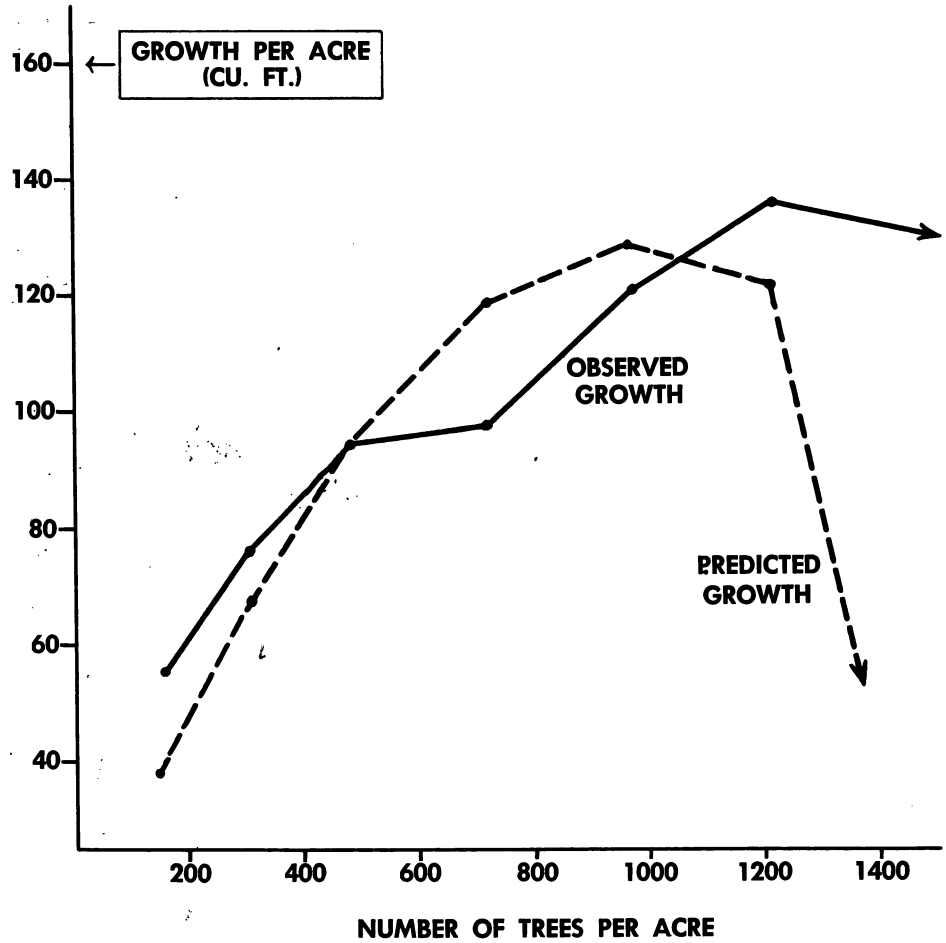

Figure 2.—Observed and predicted annual gross growth per acre on 480 jack pine stands by stocking per acre.