AN INVERTED MATRIX APPROACH FOR DETERMINING CROP-WEATHER REGRESSION EQUATIONS*

Harold F. Huddleston

U.S.D.A. Agricultural Marketing Service
Washington, D. C.

Introduction

We would like to know whether year-to-year changes, or month-to-month changes in crop yields or prospects are consistent with observed weather data. Generally, historical weather records extend back farther than records of crop yields. We wish to make use of weather data for the entire period of record even though yield data may be available for a much shorter period. This paper reports on an exploratory inverted matrix approach used in one phase of a crop-weather study.

The application of multiple regression methods in the study of relationships between crop yields and weather factors is, of course, not new, but the large amount of computational labor involved has discouraged many workers and our people from attempting correlations studies on a very extensive scale. As pointed out by R. A. Fisher, the use of the inverse matrix solution of a set of normal equations greatly reduces the amount of computations when the same set of independent variables is used repeatedly; in addition, it serves to simplify the calculation of sampling errors of the regression coefficients. However, a large amount of computational work is still required when the various dependent variables are available for only relatively few years, and these periods vary from crop to crop because of the fact that the data or series were started at different points in time. We would like some way of utilizing all the weather and crop yield data available. Therefore, we would like to devise what might be called "generalized inverse matrix solution" for a given State or area which could be used whenever the given set of weather factors were appropriate. However, the sampling errors of the regression coefficients cannot be computed using the elements of this generalized solution where the dependent variables are used for only a subperiod.

The inverse matrix solution is obtained for a given set of independent variables (i.e., weather factors) for the entire period of the weather

*Paper given before the Biometric Sessions, Gainesville, Florida March 1954.

records, or at least some fairly long period of years. It is found that the elements of the inverse matrix exhibit stability as the length of the period is increased. The elements or their ratios are used with the covariance terms between yields and each of the weather factors for the various subperiods for which crop yield data are available. Obviously, the underlying assumption which is made is that the interrelationship between given weather factors will remain fairly constant and become more reliable over time. In the example used it is assumed that this stability is over years for fixed months.

Nature of Study

In order to clarify the ideas and procedures suggested in the present study, an example of an application is given. The State of Illinois has arbitrarily been selected for examination and illustrated for corn yields. The study was conducted in the following manner. A linear relationship between yields and monthly rainfall and temperature data was used. Linear regressions have been found to give fairly satisfactory results in many cases for these variables. The functional relationship used was as follows:

$$Y = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

Where

Y = yield per acre
X_1 = average monthly rainfall for State.
X_2 = average monthly temperature for State.
X_3 = product of average monthly rainfall and temperature for the State, or $X_3 = X_1 \cdot X_2$ neglecting decimals (i.e., the product of rainfall and temperature).

Since in many multiple regression studies, joint effects may be important, the product (X_3) of rainfall and temperature was included as a third factor. The utility of the third factor has been pointed out by Hendricks[1] and Scholl where an understanding of the effects of weather is of interest and its inclusion appears desirable for a generalized regression approach.

The period selected for study was 1891–1950. The rainfall and temperature data for the month of July were selected for examination. The inverse matrix solutions were computed for the following subperiods as well as the entire period: (1) 1891–1910; (2) 1911–1930; (3) 1931–1950; and (4) 1911–1950. Table I indicates the C_{ij} values for

[1]Agricultural Experiment Station Technical Bulletin # 74 "Techniques in Measuring Joint Relationships" by North Carolina State College 1943.

TABLE I

Inverse Solutions—Illinois July Weather Data

C_{ij}	20 Year Periods			40 Year Period	60 Year Period
	1891–1910	1911–1930	1931–1950	1911–1950	1891–1950
C_{11}	+3.4509	+1.2288	+7.7944	+14.001	+5.0598
C_{12}	+0.14000	+0.093251	+0.23438	+ 0.45139	+0.16964
C_{13}	−0.047482	−0.015188	−0.10123	− 0.18371	−0.066759
C_{22}	+0.015796	+0.022006	+0.014153	+ 0.019131	+0.0087353
C_{23}	−0.0017839	−0.00091567	−0.0030412	− 0.0058850	−0.0022128
C_{33}	+0.00062649	+0.00019709	+0.0013209	+ 0.0024136	+0.00088277

the various periods, where the C_{ij} are defined by the following set (i.e., $j = 1, 2, 3$) of equations where $k = 20, 40, 60$:

$$\sum^{n_k} x_1^2 C_{1j} + \sum^{n_k} x_1 x_2 C_{2j} + \sum^{n_k} x_1 x_3 C_{3j} = 1, 0, 0$$

$$\sum^{n_k} x_1 x_2 C_{1j} + \sum^{n_k} x_2^2 C_{2j} + \sum^{n_k} x_2 x_3 C_{3j} = 0, 1, 0$$

$$\sum^{n_k} x_1 x_3 C_{1j} + \sum^{n_k} x_2 x_3 C_{2j} + \sum^{n_k} x_3^2 C_{3j} = 0, 0, 1$$

A study of the values in Table I indicates the absolute values of the C_{ij} will vary considerably from one period to the next. However, the ratios of the C_{ij} to each other are of interest in studying the tendency for stability of interrelationships of weather factors. In addition, Table II below shows the ratios of C_{ij} to C_{11} for each period.

TABLE II

Ratios C_{ij} to C_{11}

C_{ij}'	20 Year Periods			40 Year Period	60 Year Period
	1891–1910	1911–1930	1931–1950	1911–1950	1891–1950
C_{11}'	1.00000	1.00000	1.00000	1.00000	1.00000
C_{12}'	.04056	.07589	.03007	.03224	.03353
C_{13}'	− .01376	− .01236	− .01299	− .01312	− .01319
C_{22}'	.004577	.01791	.001816	.001366	.001726
C_{23}'	− .0005169	− .0007452	− .0003902	− .0004203	− .0004373
C_{33}'	.0001815	.0001604	.0001695	.0001724	.0001745

An inspection of the ratios reveals several things; (1) Relationships based upon 20 years of weather data may be expected to have little reliability for subsequent years; (2) in general, it would seem advisable that relationships using weather data for as large an area as a State should be based upon at least 40 years of data in order to obtain stable relationships among the weather factors; and (3) the ratios are fairly stable from period to period in contrast to their absolute values.

The utility of the multipliers for a long period of years which could be used as "population values," i.e., C'_{ij}, as indicated by this analysis appears to be dependent upon: (1) Finding a quick method of estimating a factor of proportionality, K, by which one can convert the ratios to absolute units, or (2) using the ratios of the C_{ij}, as in Table II, to compute regression coefficients proportional to the net regression coefficients; then obtain the relationship between yields and the weather factors by plotting the computed regression values (using the proportional regression coefficients) against the actual yields or deviations from the average yield. Further study of the variances and covariances involved appears necessary before any conclusion can be made concerning the feasibility of determining a suitable value of K a priori.

The multipliers in Table I for any of the periods may be used with any number of crop yields for the same period for the State by computing the respective covariance terms. The computational work is, therefore, considerably reduced. The data in Table II for the 60-year period (last column on right) is thought of as a "general solution".

As an example of the use indicated in (2), the yield of corn is correlated with the July weather data for Illinois. The proportional net regression coefficients are computed as follows:

$$b'_{y2.34} = C'_{11} \sum^{n} x_1y + C'_{12} \sum^{n} x_2y + C'_{13} \sum^{n} x_3y$$

$$b'_{y3.24} = C'_{12} \sum^{n} x_1y + C'_{22} \sum^{n} x_2y + C'_{23} \sum^{n} x_3y$$

$$b'_{y4.23} = C'_{13} \sum^{n} x_1y + C'_{23} \sum^{n} x_2y + C'_{33} \sum^{n} x_3y$$

Where $\sum^{n} x_1y$, $\sum^{n} x_2y$, and $\sum^{n} x_3y$ are sums of products of deviation from means for the yield of corn per harvested acre (Y) with the monthly averages of rainfall (X_1), temperature (X_2) and the product of temperature and rainfall (X_3) for the period 1911–1950 after the yields have been adjusted for trend (i.e., by use of 10-year averages). The C'_{ij} used are based upon the period 1891–1950 in equation 1) and 1911–1950 in equation 2). The regression equation for computing values from the proportional regression coefficients is:

$$Y'_c = b'_{y1.23}x_1 + b'_{y2.13}x_2 + b'_{y3.12}x_3$$

or

1) $$Y'_e = -2.315x_1 - .1295x_2 + .0375x_3$$

The actual regression between the adjusted yields and weather factors determined from the data for the period 1911–1950 is given by the following equation:

$$Y_e = b_{y1.23}x_1 + b_{y2.13}x_2 + b_{y3.12}x_3$$

or

2) $$Y_e = -24.59x_1 - 1.214x_2 + .353x_3$$

The values of Y_e and Y'_e are plotted against $Y - Y_T$ (deviations from trend) in Chart I.

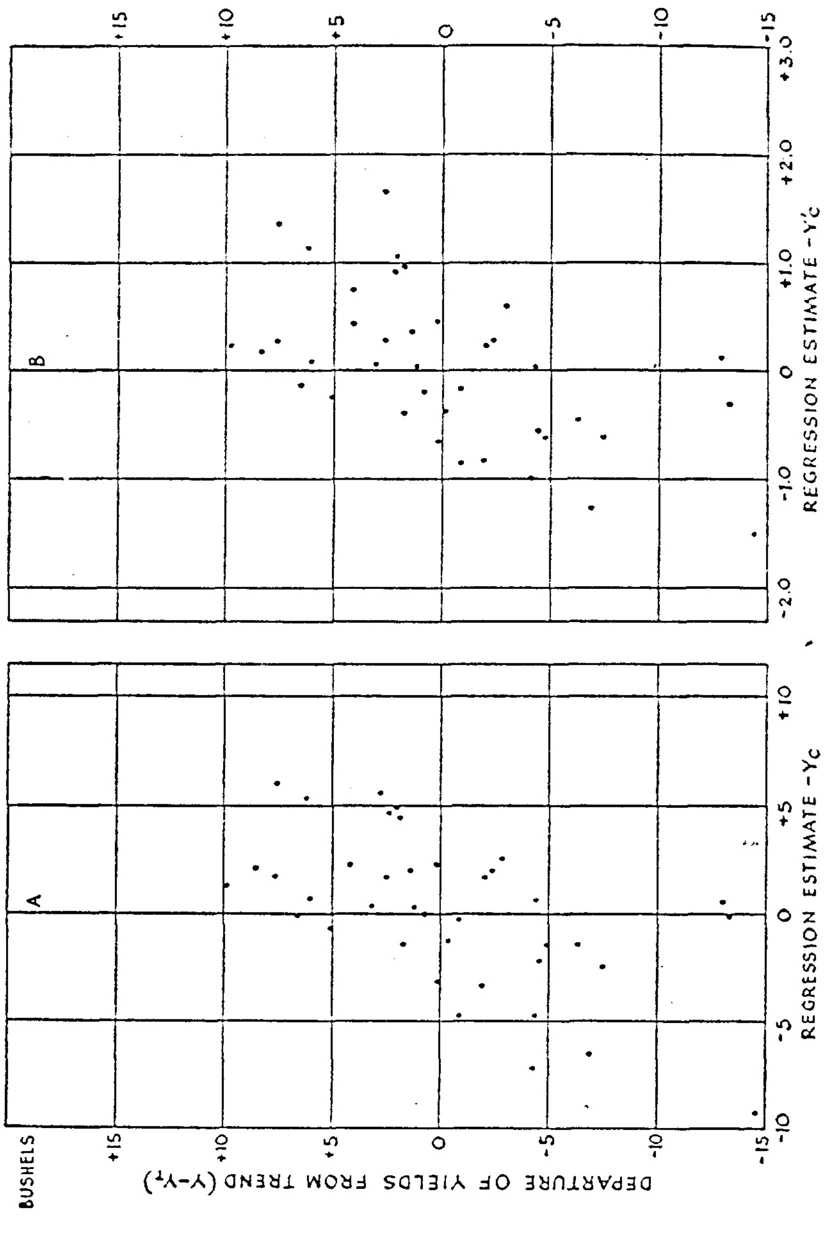
An inspection of Chart I indicates that there is little difference in the relationship found by use of the actual data for the period 1911–1950 and the ratios of the C_{ij} for the period 1891–1950 with the covariance terms for 1911–1950.

A factor (K) which can be used to covert the proportional regression coefficients to the actual values will be equal to the slope of the regression line in Part B of Chart I. That is, K (determined by least squares method) multiplied by the proportional regression coefficients will give the regression in absolute units. A comparison of the coefficients in equations 1) and 2) indicates a factor of about 10 is needed to convert the proportional coefficients to an absolute basis.

The C_{ij}'s (or C'_{ij}'s) in column 5 of Table I (or II) can similarly be used with various subperiods corresponding to the years for which the individual crop yield data are available. However, we would prefer, in general, to express the yield data as a percent of the normal or average yield rather than as deviation in absolute units. If the yield data are expressed in the percentage form, year-to-year changes are indicated by the ratio of the two years. The percentage change can then be converted to bushels per acre rather than determining the regression coefficient in their true or absolute units.

Conclusions: While a fairly large amount of computational work is involved in any multiple regression technique, it is believed that a generalized regression approach may be useful in many situations. The utilization of lengthy weather records to establish stable relationships among weather factors with determination of the covariance terms where yields are available for a much shorter period of time would appear practical based on preliminary results. In addition, the time and costs required to compute the inverse matrix solutions are not nearly so formidable with the aid of modern computing machines as has been the case in the past. It is possible that if work can be expanded

CHART 1 RELATIONSHIP OF ILLINOIS CORN YIELDS WITH WEATHER FACTORS - 1911 - 1950



along these lines a more objective means for estimating the effects of weather factors on crop production from available weather records can be used to supplement the current procedures of the Crop Reporting Board.