SAS APPLICATIONS TO A SPRING WHEAT AND SUNFLOWER YIELD MODELING PROJECT

P. Held, M. Ulmer, J. Knuteson and D. Patterson
North Dakota State University

ABSTRACT

SAS was applied to a yield modeling project by soil scientists at North Dakota State University. Soil, climatic, and management variables were used to develop yield prediction models for spring wheat and sunflower. The large number of climatic variable combinations warranted the use of SAS arrays. SAS proved to be an effective language for variable recombination and statistical analysis.

Additional Index Words: Array elements, IBM-VSPC, PROC REG, DO OVER.

Introduction

The manipulation of over 480 variables is a difficult manual task, but it represents an effective computer application. In this project, models were developed for predicting spring wheat and sunflower yields using soil, climatic, and management data from agricultural experiment station research sites. The data base consisted of information collected over a period of years from about 100 wheat sites (492 trial years) and about 50 sunflower sites (233 trial years). The North Dakota State University IBM 370/158 computer with MVS (Multiple Virtual Storage) operating system was used to store and manipulate the data. Most programming and editing was done on VSPC (Virtual Storage Personal Computing) operating system run on a sister IBM 4341 computer.

Format and Design

All data were keypunched on cards in columns and transferred to a direct access VSPC library where the data were stored, edited, and used as input in various SAS programs. Once the data were on disk, changes could be made efficiently using an IBM 3270 terminal in the view mode.

Files containing the parameters rainfall, growing degree days, number of days with temperatures greater than 90°F (32.2°C), and total degrees greater than 90°F (32.2°C) for the growing season were handled in the same way. In this paper, the rainfall file is used as an example.

Growing season rainfall was totaled for biweekly periods with each biweekly period representing one rainfall value. For example, rainfall for the last two weeks in April (variable name R4_16_30) was one value; rainfall for the first two weeks in May (variable name R5_1_15) was another value. Twelve biweekly values were required for the growing season. Planting and harvest dates varied for each observation or trial year. A method was needed for determining which combination of the 12 successive biweekly rainfall values had the greatest effect on yields of spring wheat and sunflower.

Variable Recombination by Computer

SAS arrays were used to recombine the biweekly values into new variable groups. A temporary array was used to store the 12 original values. Each array consisted of elements Q1 through Q12 with each element representing a biweekly value. An example of the array is shown in Table 1. The biweekly rainfall values are illustrated by the variables R6_16_30, R5_1_15, etc. These variables represent calendar periods April 16 through 30 and May 1 through 15, respectively.

Table 1. Temporary array used for storing biweekly rainfall values

| 10 | ARRAY Q (M) Q1-Q12 |
| 20 | Q1 = R4_16_30; Q7 = R7_16_31 |
| 30 | Q2 = R5_1_15; Q8 = R8_1_15 |
| 40 | Q3 = R5_16_31; Q9 = R8_16_31 |
| 50 | Q4 = R6_1_15; Q10 = R9_1_15 |
| 60 | Q5 = R6_16_30; Q11 = R9_16_30 |
| 70 | Q6 = R7_1_15; Q12 = R10_END |

The analysis required an evaluation of rainfall for time periods having a common starting point (planting). Another array, designated by array elements R1 through R12 and subscript J, was established to reorder the elements of array Q. R1 contained the first biweekly period from planting, R2 the second, etc. For example, if planting did not occur until Q3 (the last two week period in May), this period became the first valid biweekly period and its value was assigned to R1. R2 contained Q4, the value for the second biweekly period after planting. This procedure is illustrated in Table 2.

Table 2. Procedure used for reordering arrays*

| 10 | ARRAY R (J) R1-R12 |
| 20 | ARRAY Q (M) Q1-Q12 |
| 30 | W = XB-1 |
| 40 | ZQ = Z+1 |
| 50 | DO J = 1 to Z |
| 60 | M = J + W |
| 70 | R = Q |
| 80 | END |
| 90 | IF ZQ LE 12 THEN |
| 100 | DO J = ZQ to 12 |
| 110 | R = ; |
| 120 | END |

*Variables XB and Z are explained in text.
In Table 2:

Line 30: \( W \) is equal to the array element number (\( M \)) of the first valid biweekly period (\( X_B \)) in array \( Q \) minus 1.

Line 50: \( Z \) is equal to the total number of biweekly periods from planting to harvest in array \( Q \).

Line 60-70: This statement determines the proper placement of array \( Q \) (\( Q_1 \) to \( Q_{12} \)) in array \( R \) (\( R_1 \) to \( R_{12} \)). Note that \( J \) is the index variable of array \( R \), \( M \) is the index variable of array \( Q \), and \( W \) is the array element number of the biweekly period containing the planting date minus one. For example, suppose that planting occurs at \( Q_3 \), the last two weeks in May. \( Q_3 \) must be placed in \( R_1 \) (the first element in array \( R \)) because \( R_1 \) must contain the first period from planting. The first time through the loop, \( J \) has a value of 1. \( W \) has a value of 2 (\( Q_3 \) is the first valid rainfall period so \( W=3-1 \) or \( W=2 \)). Therefore, \( Q(M) \) becomes \( Q(1+2) = Q(3) \), the desired result. The next time through the loop \( Q(M) \) becomes \( Q(2+2) = Q(4) \), etc. Line 70 places the value of the \( Q \) array element into the \( R \) array where \( Q(M) \) is defined in line 60.

Lines 90-110: \( Z_0 \) is period following harvest. Inserted from this point.

This algorithm permitted the use of a second technique - grouping the time periods from harvest to planting. To accomplish this, lines 30 and 60 that changed \( J \) to \( W+X_B+1 \) plus 1. The analysis (manipulation of data) from harvest was done by moving the valid elements at the end of array \( Q \) to the beginning of array \( R \). If the last valid biweekly period before harvest was \( Q_{10} \), this period was put into \( R_1 \), the first biweekly period from harvest, and \( Q_{10} \) was placed in \( R_2 \), etc. In these arrays, \( R \) represents rainfall data where all observations have the harvest period as a common starting point.

To determine which combination of biweekly rainfall periods was most important, the biweekly values in array \( R \) were used to create additional variables. The consecutive biweekly periods were grouped by twos (\( R_1 + R_2 = R_{2W1} \), \( R_2 + R_3 = R_{2W2} \), \( R_3 + R_4 = R_{2W3} \), \( R_4 + R_5 = R_{2W4} \), etc.) and named array \( R_{2W} \). Groups of three (\( R_1 + R_2 + R_3 = R_{3W1} \), \( R_3 + R_4 + R_5 = R_{3W2} \), \( R_5 + R_6 + R_7 = R_{3W3} \), etc.) and named array \( R_{3W} \). Groups of twelve (\( R_1 + R_2 + R_3 + R_4 + R_5 + R_6 + R_7 + R_8 + R_9 + R_{10} + R_{11} + R_{12} = R_{12W1} \)) and named array \( R_{12W} \).

The procedure outlined for rainfall was repeated for growing degree days, number of days with temperatures greater than 90°F (32.2°C), and degrees greater than 90°F (32.2°C). After selecting the climatic variables (combinations of climatic periods), soil and management variables were added to the model for further regression analysis with yield. The procedure described above reduced the total number of climatic variables required for entry in a STEPWISE procedure.

Table 4. Regression of yield on rainfall.

| PROC REG; MODEL YIELD = R1 RV1 RV2; |
| PROC REG; MODEL YIELD = R2 RV2 RV3; |
| PROC REG; MODEL YIELD = R3 RV3 RV4; |
| PROC REG; MODEL YIELD = R12 RV12 RV12 R12W; |
| PROC REG; MODEL YIELD = R2W1 RV2W1 R2W2; |
| PROC REG; MODEL YIELD = R2W2 RV2W2 R2W3; |
| PROC REG; MODEL YIELD = R2W3 RV2W3 R2W4; |

The procedure outlined for rainfall was repeated for growing degree days, number of days with temperatures greater than 90°F (32.2°C), and degrees greater than 90°F (32.2°C). After selecting the climatic variables (combinations of climatic periods), soil and management variables were added to the model for further regression analysis with yield. The procedure described above reduced the total number of climatic variables required for entry in a STEPWISE procedure.
Summary and Conclusions

Arrays often are overlooked in preparing large amounts of data for statistical analysis. Once appropriate arrays are defined, data can be restructured rapidly.

The use of the DO OVER statement resulted in fewer programming steps when arrays were used to recombine climatic parameters as functions of time periods. Recombination was required to select the appropriate climatic variables before using them in the regression models along with soil and management variables. The procedures described in this paper permitted the elimination of many variables from the final STEPWISE models.

Arrays and the DO OVER statement in SAS provided an efficient means of restructuring and manipulating large amounts of data in preparation for statistical analysis.
THIS ROUTINE ARRAYS FROM PLANTING,
***** A COMMON STARTING POINT;
***** Q IS THE TEMPORARY ARRAY FOR RAINFALL;
ARRAY Q (M) Q1-Q12;
Q1=R1_16_30;
Q2=R2_1_15;
Q3=R3_16_31;
Q4=R4_1_15;
Q5=R5_16_30;
Q6=R6_1_15;
Q7=R7_16_31;
Q8=R8_1_15;
Q9=R9_16_31;
Q10=R10_1_15;
Q11=R11_16_30;
Q12=R12 END;
***** ZZ IS THE TEMPORARY ARRAY FOR TEMPERATURE;
ARRAY ZZ (M) ZZ1-ZZ12;
ZZ1=T1_16_30;
ZZ2=T2_1_15;
ZZ3=T3_16_31;
ZZ4=T4_1_15;
ZZ5=T5_16_30;
ZZ6=T6_1_15;
ZZ7=T7_16_31;
ZZ8=T8_1_15;
ZZ9=T9_16_31;
ZZ10=T10_1_15;
ZZ11=T11_16_30;
ZZ12=T12 END;
***** THE FOLLOWING ROUTINE DETERMINES ENDING
***** BIWEEKLY PERIOD;
IF ZZ12 NE 0 THEN XE=12;
IF ZZ10=0 THEN XE=10;
IF ZZ9=0 THEN XE=9;
IF ZZ8=0 THEN XE=8;
IF ZZ7=0 THEN XE=7;
***** THE FOLLOWING ROUTINE FINDS THE FIRST
***** BIWEEKLY PERIOD;
***** XB REPRESENTS THE FIRST VALID BIWEEKLY
***** PERIOD;
IF ZZ1 NE 0 THEN XB=1;
IF ZZ=0 THEN XB=2;
if ZZ=0 THEN XB=3;
if ZZ=0 THEN XB=4;
if ZZ=0 THEN XB=5;
***** Y IS THE PREREQUISITE VARIABLE TO
***** DETERMINE Z;
Y = XE-XB;

TABLE 5

Z IS THE TOTAL AMOUNT OF BIWEEKLY PERIODS
FROM PLANTING TO HARVEST;
Z=T+1;
ZQ IS THE BIWEEKLY AT WHICH (.) START,
FOLLOWING HARVEST;
IN OTHER WORDS, ZQ ASSIGN A MISSING
VALUE TO THE BIWEEKLY VALUES;
WHICH FOLLOW HARVEST;
W IS THE BEGINNING POINT OF THE TEMPORARY
ARRAY PREVIOUS ONE - WHICH
WHEN ADDED TO SUBSCRIPT (J), REPRESENTS
THE STARTING POINT OF THE
FINAL ARRAY - SEE LINE BELOW: (M=J+W);
ZP=E+1;
W=XB-1;
Z1=Z-1;
Z2=Z-2;
Z3=Z-3;
Z4=Z-4;
Z5=Z-5;
Z6=Z-6;
Z7=Z-7;
Z8=Z-8;
Z9=Z-9;
Z10=Z-10;
FINAL ARRAYS USED IN THE PROC STATEMENTS;
RENAME;
ARRAY X (J) X1-X12;
ARRAY T (J) T1-T12;
ARRAY HQ (J) HQ1-HQ12;
ARRAY DQ (J) DQ1-DQ12;
SUBSCRIPT OF NEW ARRAY - WHICH IS BEING
ASSIGNED VALUES (T & R);
SUBSCRIPT OF TEMP ARRAY - OF WHICH
VALUES ARE BEING RELOCATED INTO
THE NEW ARRAY SO THAT ALL OBSERVATIONS
START FROM PLANTING;
DO J=1 TO Z;
T=ZZ; R=Q; HQ=ZH; DQ=DZ; END;
IF ZQ LE 12 THEN DO;
THIS DO LOOP PUTS IN MISSING
VALUES AFTER
HARVEST;
DO J=ZQ TO 12;
R=.; T=.; END;
END;
DROP Q1-Q12;
DROP ZZ1-ZZ12;
DROP Z1-Z10;
DROP ZH1-ZH12;
DROP ZD1-ZD12;
END OF RENAMING SESSION;