EVALUATING THE ADDITION OF WEATHER DATA TO SURVEY DATA TO FORECAST SOYBEAN YIELDS

M. Denice McCormick and Thomas R. Birkett, National Agricultural Statistics Service
M. Denice McCormick, 3251 Old Lee Highway, Fairfax, VA. 22030

KEY WORDS: Precipitation, regression, model evaluation.

INTRODUCTION

In 1990, the National Agricultural Statistics Service (NASS) introduced new models to forecast yield for corn and soybeans on the regional and state levels in a plan to phase out the older, less accurate models (Birkett 1990). An annual survey collects data from randomly selected sample plots in randomly selected fields. The old regression models predicted the components of yield such as number of pods per plant and weight per pod at the plot level based on five years of previous data. Plot level data were then aggregated to the state level. The new models are also regression models, and have initially been developed to predict yield directly rather than the components of yield using survey data aggregated to the regional level. Regions are constructed from a subset of states that participate in the annual survey. A longer period of years in the historic data set must be used since only one data point is used to represent each year.

This research effort evaluates the addition of precipitation data to the models for soybeans, in order to improve the precision of early season (August 1 and September 1) yield forecasts. It considers data for twelve years, 1980 to 1991, for a region of six states: Illinois, Indiana, Iowa, Minnesota, Missouri and Ohio. The performance of the models is examined in this report for August and September, which are the early season forecast periods.

Attempts have been made previously to include weather data in forecast models. Sanderson (1942) used crop condition reports and weather data to forecast the yield per acre of wheat and found gains could be made in forecast accuracy, especially in late season models. House (1977) recommended that weather variables be incorporated into a within-year growth model to forecast corn yields. Sebaugh conducted a number of investigations in this area. Weather data were included in the analysis of the performance of Climatic and Environmental Assessment Services (CEAS) and Thompson models in forecasting spring wheat (1981). Sebaugh also considered its use in Kestle's "Straw Man" model in forecasting corn and soybeans (1981). Later, Sebaugh and Cotter used weather data to forecast soybeans (1983). Others, such as Maas (1982), Sebaugh (1983), and Warren (1990) constructed weather related indices to include in yield forecast models.

This study was limited to evaluating five different model forms incorporating precipitation data and regular survey variables into the multiple regression framework. Since the models show improved performance using aggregated survey data values at the regional level it was anticipated that this would also prove to be an effective method for aggregating weather data.

DATA

Precipitation Data

Precipitation values used in the models represent accumulated precipitation for the growing season at the regional level. The data are provided from a network of National Weather Service weather stations in each state. For the month of August, the growing season is defined as the period from April 1 through July 31. For September, the growing season is the period from April 1 through August 31. The variable is constructed as follows:

$$P_t = \frac{\sum_{s=1}^{S} A_{ts} R_{ts}}{\sum_{s=1}^{S} A_{ts}}, \qquad (1)$$

where

P_t = the estimated accumulated precipitation over the growing season for the region, year t,
S = the number of states covered,
R_{ts}= the estimated accumulated precipitation over the growing season for year t, state s, and
A_{ts}= the acres for harvest for year t, state s.

Further:

$$R_{ts} = \frac{\sum_{d=1}^{D_s} A_{tsd} E_{tsd}}{\sum_{d=1}^{D_s} A_{tsd}},$$

where

A_{tsd}= the acres for harvest for year t, state s, district d,
D_s = the number of districts per state s,

$$E_{tsd} = \frac{1}{W_{tsd}} \sum_{w=1}^{W_{tsd}} U_{tsdw}$$

where

E_{tsd}= the average station accumulated precipitation for year t, state s, district d,
W_{tsd}= number of weather stations for year t, state s, district d, and
U_{tsdw}= accumulated precipitation for year t, state s, district d, weather station w.

Survey Data

The construction of the survey data is discussed by Birkett (1990). For the month of August, the independent variable is the estimated number of lateral branches per eighteen square feet. For September, the independent variable is the estimated number pods with beans per eighteen square feet. The State-level estimates for August are constructed as follows:

$$F_{ts} = \frac{1}{m_{ts}} \sum_{j=1}^{m_{ts}} B_{tsj} L_{tsj},$$

where

m_{ts}= the number of samples in J_{ts} year t, state s,
J_{ts}= the subset of samples classified in maturity categories 2-6 (or 1-6 in southern states), year t, state s,
B_{tsj}= plants per 18 square feet for year t, state s, sample j,
L_{tsj}= lateral branches per plant year t, state s, sample j, and

F_{ts} = number of lateral branches per 18 sq. feet year t, state s.

The State-level estimates are combined to the regional level with current acres harvested used as the weight as follows:

$$Z_t = \frac{\sum_{s=1}^{S} A_{ts} F_{ts}}{\sum_{s=1}^{S} A_{ts}}$$

where

A_{ts}= the acres for harvest for year t, state s

All of the definitions are the same for September except O_{tsj} is substituted for L_{tsj} and Q_{ts} is substituted for F_{ts}, and

O_{tsj}= pods with beans per plant, year t, state s, sample j, and

Q_{ts} = estimated pods with beans per 18 sq. feet year t, state s.

Yield Data

The combined yield values for the six states included in this study were calculated as follows:

$$Y_t = \frac{\sum_{s=1}^{S} A_{ts} V_{ts}}{\sum_{s=1}^{S} A_{ts}} ,$$

where

Y_t= final regional yield for year t, and

V_{ts}= NASS state yield year t, state s.

METHODOLOGY

Statistical analysis methods used to evaluate the performance of precipitation data in combination with survey data are correlation and regression analysis. Multiple linear regression models with associated diagnostics for model fit and forecast accuracy were examined.

The following regression models were examined for each month.

$$1A{:}\quad Y_t = \beta_o + \beta_1 Z_t + \epsilon_t$$

$$1B{:}\quad Y_t = \beta_o + \beta_1 P_t + \epsilon_t$$

$$2A{:}\quad Y_t = \beta_o + \beta_1 Z_t + \beta_2 Z_t^2 + \epsilon_t$$

$$2B{:}\quad Y_t = \beta_o + \beta_1 P_t + \beta_2 P_t^2 + \epsilon_t$$

$$3{:}\quad Y_t = \beta_o + \beta_1 Z_t + \beta_2 P_t + \epsilon_t$$

$$4{:}\quad Y_t = \beta_o + \beta_1 Z_t + \beta_2 Z_t^2 + \beta_3 P_t + \epsilon_t$$

$$5{:}\quad Y_t = \beta_o + \beta_1 Z_t + \beta_2 Z_t^2 + \beta_3 P_t + \beta_4 Z_t P_t + \epsilon_t$$

Model 1A is used by NASS for the August and September forecasts. Model 2A was used in September, 1991, on an experimental basis. Model 5 is the most extensive.

Model Evaluation Criteria

The primary model evaluation criterium is a set of prediction intervals (PI) for the years 1988, 1981 and 1990. These years correspond to the minimum, median and maximum six state regional soybean yields, respectively, over the 12 years in the study. A second criterium is the adjusted coefficient of determination, R_a^2 which provides a measure of correspondence between predicted and actual yields. Both the PI and R_a^2 are based on the sum of squared differences from the least squares analysis used to derive the model parameters. Two other criteria are provided which are based on the absolute relative differences (ARD) between the predicted and actual yields (Sebaugh and Cotter 1983; House 1977). Each of these evaluation criteria is further defined below.

1. The prediction interval (PI) refers to half the confidence interval length for the predicted value of a future Y for a given future year o. That is, at the α significance level,

$$PI = t(1-\frac{\alpha}{2};n-1-p)SD(\hat{Y}_o),$$

where

$$SD(\hat{Y}_o)=s[(x_o'(X_o'X_o)^{-1}x_o) + 1]^{\frac{1}{2}},$$

s = (residual MSE)$^{1/2}$,
x_o= relevant p-dimensional row vector of independent variables for year o (for example, in Model 3: p = 3, x_o = [1, Z_o, P_o]),
X_o= relevant (n-1 x p) matrix of independent variables (excludes x_o),
n= number of years, and
p= number of parameters.

The X_o matrix excludes the row vector x_o, so that the PI reflects the accuracy expected in an operational model where current year data are not included in the model development. A significance level of 0.32 was used for this study, which provides t values near 1.0. Consequently, the future Y will fall within the calculated PI of the predicted Y approximately 68% of the time.

2. R_a^2 is used as a goodness-of-fit test for each model with an adjustment made for the corresponding degrees of freedom (Draper and Smith 1981). R_a^2 is calculated as:

$$R_a^2 = 1-\frac{(RSS_p)/(n-p)}{(CTSS)/(n-1)},$$

where

RSS_p = the residual sum of squares taking the changing number of parameters into account,
CTSS = the corrected total sum of squares,
n = the number of years, and
p = the number of parameters.

3. The average absolute relative difference (ARD) is calculated as:

$$\overline{ARD} = \frac{1}{n}\sum_{t=1}^{n}|RD|_t$$

where

$$|RD|_t = 100\frac{|\hat{Y}_t - Y_t|}{Y_t}$$

Y_t = regional level yield, year t, and

$\hat{Y}_t$ = regional level predicted yield, year t.

The predicted yield $(\hat{Y}_i)$ is based on a model that does not include data from the forecast year. This statistic is a measure of forecast reliability that provides an empirical indication of how closely the model predicted values come within final NASS yields on a percentage basis, without any distributional assumptions.

4. The number of years the ARD is less than 5% provides an empirical basis for comparing how consistently predicted yields are within 5% of the yield.

Outlier Identification

Since the purpose of the models is to make forecasts, the rstudent statistic (also called the studentized residual) was used to help identify outliers to be excluded from the model. This statistic was first recommended in Belsley, Kuh and Welsh (1980). It is similar to the standardized residual:

$$r_{si} = \frac{r_i}{s\sqrt{1-h_i}},$$

where

r_i = i^{th} residual,
s = (residual MSE)$^{1/2}$, and
h_i = $x_i'(X'X)^{-1}x_i$

Here, s is replaced by s(i). S(i) is the estimate of σ with the i^{th} observation deleted. In a forecasting model, rstudent measures how many prediction standard errors the forecast is from the observed Y. Observations with absolute values of rstudent greater than 3.0 were identified as outliers. The rstudent statistic is distributed closely to the t-distribution with n-p-1 degrees of freedom.

The result of the examination of the rstudents found that in September only, 1980 is an outlier for Models 2A, 4, and 5. To test the improvement that occurs within each of these models, 1980 was excluded in a second regression analysis.

RESULTS

Based primarily on comparisons of the PIs, the best model for August is Model 1A, which is the model currently being used by NASS to provide August forecasts. Model 1A consistently has the lowest prediction intervals (PI) of 2.93, 2.58 and 2.58 for years when the minimum, median and maximum yields occur (1988, 1981 and 1990) respectively. Adding the precipitation term (Model 3) increased the length of the prediction intervals by approximately ten percent for all three years, but did produce an equivalent R_a^2 and slightly lower ARD values. In September, Model 2A: Pods and Pods^2, the quadratic model, is the best model when evaluated in terms of prediction intervals. It is the simplest and most cost efficient model. It has relatively low prediction intervals of 2.46, 2.38 and 2.39 for 1988, 1981 and 1990 respectively; a relatively high R_a^2 of .68; a relatively low average ARD value of 4.35 (on average less than 5.0); and predicts within 5% of yield eight out of twelve years.

CONCLUSIONS

In August, there is no evidence that a change from a univariate survey data model is warranted. In September, the quadratic model, using pods and pods^2 (2A), shows definite improvement in all evaluation criteria over the univariate model (1A). Adding the precipitation term investigated for this study to the quadratic model shows limited gains in forecast accuracy at the regional level.

BIBLIOGRAPHY

Belsley, David A, Kuh, Edwin, Welsh, R.E., (1980), Regression Diagnostics, John Wiley & Sons.

Birkett, Thomas R., (1990) "The New Objective Yield Models for Corn and Soybeans", SMB Staff Report Number SMB-90-02, U.S. Department of Agriculture.

Draper, N.R., Smith, H., (1981), Applied Regression Analysis, John Wiley & Sons Second Edition.

House, Carol C., (1977) "A Within-Year Growth Model Approach to Forecasting Corn Yields", Crop Reporting Board, Economics, Statistics, and Cooperatives Service, U.S. Department of Agriculture.

Kaiser, Mark, Sebaugh, Jeanne L., (1984) "Methods for the Evaluation of Real-Time Weather Data for use in Crop Yield Models: An Application to North Dakota., SRD Report Number AGES840424, U.S. Department of Agriculture.

Kestle, Richard A., (1981) "Analysis of Crop Yield Trends and Development of Simple Corn and Soybean "Straw Man" models for Indiana, Illinois, and Iowa. AgRISTARS Yield Model Development Project. Document YMD-2-11-1 (80-11.1), ESS Staff Report AGES810114, U.S. Department of Agriculture.

Maas, Stephan J., (1982) "Forecasting Yields Using Weather-Related Indices", SRD Staff Report Number YRB 8-2-08, U.S. Department of Agriculture.

National Oceanic and Atmospheric Administration, (1987) "TD-3200 Summary of Day Co-operative", U.S. Department of Commerce.

Neter, John, Wasserman, William, Kutner, Michael H., (1983), Applied Linear Regression Models, Richard D. Irwin, Inc.

Sanderson, Fred H., (1942) "Use of Condition Reports and Weather Data in Forecasting the Yield per Acre of Wheat", SMB Staff Report Number YRB 42-01, U.S. Department of Agriculture.

Searle, S.R., (1971) Linear Models, John Wiley & Sons.

Sebaugh, Jeanne L., (1981) "Evaluation of "Straw Man" Model 1, the Simple Linear Model, For Soybean Yields in Iowa, Illinois and Indiana", SRD Staff Report Number AGES811214, U.S. Department of Agriculture.

Sebaugh, Jeanne L., (1981) "One, Two and Three Line Segment "Straw Man Models, Soybean Yields in Iowa, Illinois and Indiana", SRD ESS Staff Report Number AGESS810514, U.S. Department of Agriculture.

Sebaugh, Jeanne L., Cotter, James J., (1983) "Comparison of the CEAS and Thompson-type Models for Soybeans Yields in Iowa, Illinois and Indiana", SRS Staff Report Number AGES830613, U.S. Department of Agriculture.

Sebaugh, Jeanne L., (1983) "Evaluation of the Feyerherm '81 Spring Wheat Models for Estimating Yields in North Dakota and Minnesota", SRS Staff Report Number AGES830609, U.S. Department of Agriculture.

Warren, Fred B., (1990) "An Operational Test Using Weather Data to Forecast Corn Ear Weight, 1988", SRB Staff Report Number SRB-90-05, U.S. Department of Agriculture.