97-02

Selection of the Optimum Linear Regression Relationship for Determining Spring Wheat Yields Using NOAA AVHRR Data

Paul W. Cook

U. S. Department of Agriculture
National Agricultural Statistics Service
Fairfax, Virginia 22030

Paul C. Doraiswamy

U. S. Department of Agriculture
Agricultural Research Service
Beltsville, Maryland 20705

1997

ABSTRACT

This research extends earlier work done by the authors in evaluating a linear regression model to relate AVHRR data to spring wheat yields in North Dakota for the 1989-1992 crop season. The earlier evaluation of the model using the EROS Data Center Land Characteristics Data Base to mask out nonspring wheat related data from the county averages had shown encouraging results. The objective of this study was to examine the linear regression relationship of county spring wheat yields to the county averages of Normalized Difference Vegetation Index (NDVI) of individual biweekly composite periods of NOAA AVHRR data for the same four years and to all possible combinations of the county NDVI sums for the available AVHRR periods. An analysis of the available data using S-plus showed that Period 22 (late June to early July) was the best individual period of biweekly data for each year individually. However, combining the four years of data using only Period 22 AVHRR county averages provided a linear relationship with a much lower R-square than did the individual years. The sum of NDVI county averages for Periods 20, 22, 24, and 26 (corresponding to June 22 - August 16 for 1990) for the four years had higher R-squares than did any other possible AVHRR county sums of three or more periods. This sum of periods was the same sum as in the earlier study. This presentation will provide greater detail of the analyses done and provide an evaluation of the relationships formed using NDVI sums of three through ten periods available during the growing season. These analyses show quite conclusively that the previously chosen sum of four biweekly periods of AVHRR data provide the best possible relationship from the four years.

INTRODUCTION

The Agricultural Statistics Board (ASB) of the USDA's National Agricultural Statistics Service (NASS) is responsible for publishing official estimates of crop acreage, yield, and production for the United States. The ASB considers indications from multiple surveys and administrative sources throughout the growing season to calculate these estimates at National and State levels. Survey information includes farmer interviews conducted by enumerators, telephone contacts, infield objective yield measurements, and mail questionnaires. NASS field offices calculate and publish county estimates of crop acreage, yield, and production for major crops. County level estimates must

add to equal the official state estimates and are generally estimated from mail survey and administrative data (Iwig, 1993). State crop yields are obtained by dividing the total State crop production by the total harvested acres for that crop. County crop yields can be decided in this manner as well.

Remote sensing data have had a continuing potential use for monitoring extreme weather conditions that can adversely affect the crop's development (Doraiswamy, 1993). Satellite-based sensors usually have visible, near-infrared, shortwave-infrared, and even thermal spectral bands. The LANDSAT Thematic Mapper sensor collects data in all these spectral bands with a 30-meter (120-meter thermal resolution, but its repeat cycle of 16 days and high processing costs have limited its effectiveness to small geographic areas. Other sensors such as SPOT MultiSpectral Sensor have less thorough spectral coverage(visible and infrared only), but do provide greater spatial resolution and repeat cycles from their two-satellite system. However, the higher resolution sensors, both TM and MSS, are expensive to purchase and process. Even if the cost were less, their extended repeat cycles limit their capabilities to assess crop condition.

Large scale crop yield studies require frequent remote sensing data collection over wide areas. One sensor that provides daily data collection at low cost over wide areas is that of the Advanced Very High Resolution Radiometer (AVHRR) sensor on board the National Oceanic and Atmospheric Administration (NOAA) weather satellite. The AVHRR sensor provides only a 1.1 km pixel (picture element) accuracy. However, by providing daily overpass data from five spectral bands, the AVHRR sensor is more useful for monitoring crop condition over time than the more spatially accurate data from LANDSAT or SPOT. Two AVHRR bands with visible and near-infrared reflectance values can be combined to provide the Normalized Difference Vegetation Index (NDVI = (Near Infrared Band - Red Band)/(Near Infrared Band + Red Band)) that has a definite relationship to crop condition. This study uses NDVI values from ten biweekly AVHRR images to assess the relationship between spring wheat yields and AVHRR data.

METHODOLOGY

The USGS's EROS Data Center (EDC) composites individual date AVHRR images over a 14-day period to create a Biweekly Conterminous digital product for the Unites States that is geo-registered, essentially cloud free, and has the data associated with the maximum NDVI (over the 14-day period) at each pixel location. EROS Data Center provided the AVHRR biweekly composite data for the entire U. S. resampled to one kilometer resolution (Eidenshink, 1992). EDC also provided a categorized U. S. land use categorized product (the EDC Land Characteristics Data Base) product with 167 categories (Brown, and others, 1993) that was also at one kilometer resolution. The North Dakota biweekly AVHRR composite data for the four crop seasons from 1989 through 1992 were chosen for this analysis; both North Dakota and South Dakota were examined in an earlier paper (Doraiswamy and Cook, 1995).

Some categories of the EDC Land Characteristics Data Base provided potential locations for spring wheat production in North and South Dakota. These categories made possible calculating county averages for those pixels likely to have spring wheat acreage. Those counties having less than 100 pixels likely to contain spring wheat were deleted when calculating the value of coefficients for the

prediction equation in this analysis.

Although there is a definite improvement for many counties in the relationship between the 'masked' AVHRR pixels and the spring wheat acreage within each county, many of the 53 counties still have too many remaining non-spring wheat AVHRR pixels (See Figure 1.). An improved mask should increase the accuracy with which the county NDVI averages relate to the spring wheat yields. More recent unpublished work (Stern, Doraiswamy, and Zara 1997) holds some promise in greatly improving on which AVHRR pixels to denote as spring wheat for creating county AVHRR averages. An improved mask would help to ensure that the county NDVI averages represent spring wheat areas rather than other crops.

The NDVI values relate to the "greenness" of the crop and so rise and fall during the crop growing season. Crop stages included in the original analysis were those of flowering to wax ripeness. Weather and planting data for North Dakota (and South Dakota) suggested that the dates of June 13 through August 13 on average would match these growth stages. For example, the equivalent four U. S. AVHRR biweekly Period composites extended from June 22 through August 16 in 1990.

A linear regression model in the earlier paper related official NASS county spring wheat yield estimates to a sum of four selected biweekly NDVI average county values of AVHRR data. However, the objective of this study was to evaluate the AVHRR Spring Wheat model more thoroughly by examining all available data for North Dakota in more detail using the statistical capabilities of S-Plus (Statistical Sciences, 1995). The analysis of the data set using S-Plus evaluated each individual period from the ten available periods from 1989 - 1992 and all possible sums of four or more of the available ten periods for the same four years.

Earlier ANALYSIS of the AVHRR DATA

Use of all the available North Dakota counties that had a minimum of four AVHRR pixels provided a total of 53 counties for the years 1989 through 1992. The Rsquares for the sum of periods 20, 22, 24, and 26 had a low of 0.53 for 1989 and a high of 0.63 for 1992.

Although not always the best Rsquares for all years, a deletion of those counties with 100 or fewer pixels improved Rsquares to 0.57 for 1989 to 0.69 for 1992. Choosing this cutoff of 100 pixels provided a combined regression for all four years of 0.57 with coefficients of -34.95 and 37.76. Standard errors of 4.18 and 2.37, respectively, provided sufficient prediction accuracy. Also, this criterion excluded only four counties from the analysis so that little information was lost. Use of the EDC Land Characteristics Database clearly improved Rsquares while requiring deletion of little county information.

S-PLUS ANALYSIS OF THE DATA SET

Evaluation of the Ten Individual AVHRR Periods

Ten biweekly periods of AVHRR data that corresponded to nearly the entire crop growing year were available for each of the four years. These periods with even numbers were the biweekly Periods 12

through 30. Odd numbered periods are a seven-day composite product between the biweekly imagery periods These ten biweekly periods spanned ten biweekly periods from late May through late September for the four study years for North Dakota. The selection of likely spring wheat pixels for the counties within North Dakota was done with the Land Analysis System on a VAXStation 3100 Model 38. EROS Data Center's Land Characteristics Database provided a subset of categories most related to spring wheat. County averages for the selected pixels were retained for all fifty-three counties in North Dakota for each of the forty available periods.

The original data set consisted of the following information. Each observation contained the specific year, the crop reporting district (a two-digit code for the nine groupings of adjoining counties into which the State was divided), the fips code for each county (a two-digit code by which each county was designated), the county name for each county, and a county NDVI average calculated after eliminating AVHRR pixels not considered to have spring wheat areas for each of the ten biweekly periods for a given year. Conversion to S-Plus took place through an ASCII file input into a data frame for analysis in S-Plus (Statistical Sciences, Inc. 1995). This analysis examined all ten individual AVHRR periods and all possible sums of four or more periods.

Although not examined in greater detail during this study, Figure 2 clearly shows that the county spring wheat county yields were highly variable during the four years at the crop reporting district level. Clearly, Crop Reporting Districts 30 and 60 had higher spring wheat yields for the four years (1989- 1992) than did the remainder of crop reporting districts. Combined with the inaccuracy of the mask that removed nonspring wheat related AVHRR pixels, this variation in spring wheat crop reporting district yields is another difficulty in creating the final relationship between the AVHRR county averages and that of the county spring wheat yields.

The S-Plus Leaps procedure (Statistical Sciences, Inc. 1995) provided the means of examining all regressions using the AVHRR period data either individually or as groups of periods without running each possible least squares regressions individually. The Mallows C_p Statistic (Kotz, S. and Johnson, N. 1982) is

$$C_p = RSS_p / s^2 - n + 2p$$

where RSS_p is the Residual Sum of Squares over the k sets of variables of the residual variables and s^2 is the estimated residual variance using the full model. Also, here, n equals 53 for the number of counties, k equals ten for the number of period variables, and p is the number of parameters in the regression equation and so equals two since the intercept is included. The evaluation of the statistic requires choosing the value of Cp that is close to p, or here two. However, no one variable model is sufficient, so only the smallest Cp helps in choosing the best period. Of course, each analysis included an intercept term within the model definition.

Although the Leaps' procedure provided an evaluation of all possible regressions using one, two, and so on, up to all ten periods of data, our interest in this part of the analysis is with the selection of a best period to use alone. Each year had a varying selection from among the ten periods and so no definite pattern prevailed. First, the Leaps' analysis of the individual PERIODS of AVHRR data

provided insight into the relative importance of each PERIOD considered individually and in groups of two up to ten PERIODS together. However, three years (1989 through 1991) gave PERIOD 22 the best Cp ranking for two factors (the other factor was the intercept). The 1992 data gave the best ranking to Period 26 whereas Period 22 was number four. Table 1 summaries the results of the Leaps' analysis by presenting the Cp values for individual periods and showing the ranking of the variables for each year. Table 2 presents the results of the regressions using Period 22 alone.

Table 1. S-Plus Leaps Analysis of the Individual AVHRR Periods.[1]

Year	*1989*	*1990*	*1991*	*1992*
AVHRR Period	Cp	Cp	Cp	Cp
22	11[1]	37 [1][2]	35 [1]	32 [3.5]
20	17[2]	50 [3]	145 [8]	26 [2]
24	41[3]	45 [2]	55 [2]	32 [3.5]
12	92[4]	141 [5]	92 [3]	56 [6]
18	94[5]	142 [6]	94 [4]	88 [7]
28	108[6]	152 [7.5]	126 [7]	90 [8]
16	109[7]	188 [10]	191 [10]	128 [9]
14	110[8]	178 [9]	158 [9]	117 [10]
30	113[9]	152 [7.5]	122 [6]	34 [5]
26	114[10]	130 [4]	117 [5]	23 [1]

Table 2. Regression Values for Period 22: By Year and the Combined Four Years

Year	*1989*	*1990*	*1991*	*1992*	*89 - 92*
R-Squared	0.64	0.65	0.66	0.60	0.45
Intercept	- 12.7	- 61.6	- 26.1	1.8	- 17.3
Period 22 Coef.	80.0	182.7	110.7	80.1	100.3

Cp values for additional periods in the model were less, however, there was no consistent grouping of periods across years to choose. The intercept and period 22 terms had highly significant t-values

[1] Values in this table are rounded to simplify the presentation.

[2] Values in brackets [ ] are the rank of the specific period for that year.

of 0.0000 for all coefficients except the intercept for 1989 that was significant at 0.0011 and the intercept for 1992 that was not significant at 0.6780. The large ranges of values for the intercept values and regression coefficients for the Period 22 terms plus the large drop in Rsquare values when the years are combined show that the years are not consistent.

Using all ten periods in the regression equation did provide higher Rsquare values for each individual year. Table 3 provides some results of the regressions for the four years. Although the Rsquare values might be quite good for each regression equation, the reality is that none of the regression fits are appropriate. Most of the coefficients for the periods have very large standard errors that can be two to five times as large as the coefficient itself. Thus, very few of the coefficients are significant. Clearly, using all the periods individually will not provide a good predictive equation, so the combined years are not considered.

Table 3. Regression Coefficients for All Periods Regression for 1989 through 1992.

Year	*1989*	*1990*	*1991*	*1992*
Rsquare	0.75	0.83	0.83	0.79
Intercept	- 22.5	- 28.0	6.4	30.3
Period 12	59.7	- 132.6	- 74.6	- 132.9
Period 14	- 16.2	89.2	- 80.5	56.6
Period 16	- 87.0	- 132.0	61.6	- 16.6
Period 18	115.2	30.6	36.6	14.5
Period 20	- 19.0	176.7	- 60.7	31.6
Period 22	28.6	- 44.4	58.7	5.1
Period 24	47.4	42.1	- 1.8	- 25.1
Period 26	- 11.0	- 90.1	51.1	19.3
Period 28	9.6	369.8	- 95.8	25.1
Period 30	- 8.1	- 245.2	90.2	- 8.1

Evaluation of SUMS of Four or More PERIODS

Summing average county NDVI values over periods should be more related to yields than would be individual composite periods (Doraiswamy and Cook, 1995). The next analysis used the S-Plus Leaps' function applied to all possible combinations of sums of the period county averages using four through all ten periods. This procedure evaluated the effectiveness of summing groups of periods besides the four periods in the original study to decide if sums of another group of four or more periods would relate better to the spring wheat yields.

Although few sums produced linearly dependent variables, a subset all possible sums remained in the Leaps' analysis. Period 12 became the first period (denoted as 1) while Period 30 became the tenth (denoted as A) period during the growing season. Sums of the period county NDVI averages were expressed as, for example, SUM.1A, that is, the sum of periods one through ten (A). There were ten sums remaining in the analysis as shown in Table 4, below. SUM.58 has the smallest C_p for each of the four years. However, no year has a C_p that is sufficiently near the optimum value of two to conclude that the regression will fully explain the spring wheat yields.

Table 4. Cp Statistics for SUMS of PERIODS for 1989, 1990, 1991 and 1992.[3]

	1989	*1990*	*1991*	*1992*
Period Sums	Cp	Cp	Cp	Cp
SUM58	77	41	51	15
SUM69	81	68	68	20
SUM59	98	52		
SUM38	119	68		22
SUM49	123	58	76	20
SUM6A	132		77	
SUM49	138	58	76	
SUM5A	139	67	72	
SUM39	141			24
SUM28	149			

The plot in Figure 3 helps to explain more clearly the meaning of the coefficients from Table 4. The three years of 19989, 1991, and 1992 have slopes that are quite comparable. However, the slope for 1990 is clearly steeper than any of the other three years. Indeed, even the years with similar slopes have different intercept values. These intercept differences are due to the four years having spring wheat yields that are not completely comparable.

Table 5 continues the analysis for years and presents the intercept term, standard error of the intercept, and the coefficient the county averages sum of four periods with standard error (std. error). These values vary significantly from year to year and as compared to the combined four-year analysis.

---

[3] Only the original sums of periods for 1989 are in the table, since the remaining sums of periods did not have C_p values that were better than those listed.

Table 5. *Intercept Values and Coefficients for Four Years and 1989-1992 for the SUM.58 analysis.*

Year	*Value*	*Std. Error*	*t-value*	*Pr.(>|t|)*
1989	- 25.7	6.3	- 4.1	0.0002
SUM.58	29.4	3.9	7.5	0
1990	- 49.2	8.7	- 5.7	0
SUM.58	45.1	4.8	9.4	0
1991	- 19.8	5.8	- 3.4	0.0012
SUM.58	28.4	3.3	8.5	0
1992	- 0.7	3.9	- 0.2	0.8661*
SUM.58	22.8	2.2	10.5	0
Comb. 89-92	- 29.4	3.9	- 7.5	0
SUM.58	34.8	2.2	15.5	0

* Not Significant

Combining the four years of spring wheat yields with the corresponding AVHRR NDVI county averages as in Figure 4 shows conclusively that the four years of data are not fully comparable. Those counties having the lowest recorded yields (from 25 - 40 bushels per acreage) during 1992 relate to similar values of AVHRR county NDVI averages as do much lower yields (from 10 to 20 bushels per acreage) for the years of 1989 and 1990. These larger minimum spring wheat yields for 1992 are substantial outliers in the combined regression for the four years.

One way to explain the lack of agreement between the four years of spring wheat yields with the AVHRR NDVI county averages is that the mask that eliminated non-spring wheat AVHRR pixels was not fully accurate (as described earlier). Another difficulty in relating the two data sets is that the spring wheat yields within the crop reporting districts of the State (Figure 2) are not consistent for the four years. This lack of consistency in the four years of data might cause the difficulty.

CONCLUSIONS

These analyses of the North Dakota NDVI data set using all individual periods and all sums of four or more periods (including all ten available periods) have not provided a final answer to developing a model using AVHRR NDVI data to predict spring wheat yields. The possibility of using other periods of NDVI data for spring wheat in North Dakota other than the sum of periods corresponding to the periods 20, 22, 24, and 26 for these four years is excluded. The linear model is properly

relating spring wheat yield and AVHRR data, but not with a sufficiently high degree of accuracy.

Additional analyses of the relationship between spring wheat yields and AVHRR data are needed. Use of Landsat TM data with a resolution of 30 meters may be necessary to establish the characteristics of the yield data in relationship to remote sensing observations. Use of other band combinations is possible in analyzing the more spectrally capable Landsat data. Also, the TM data should have more potential in relating more closely to the varying crop yields with the crop reporting districts for any given year. An improved mask for the location of county spring wheat acreages would also be needed to improve the Landsat TM analyses.

ACKNOWLEDGMENTS

The authors thank Dr. Dan Carr, George Mason University, for encouragement to learn S-plus and examine the data more fully during his Exploratory Data Analysis Course. Our thanks to Rick Mueller, USDA/NASS, for assistance with the graphics manipulation in Framemaker for final printing and to Mike Craig for his editorial recommendations that improved the paper.

REFERENCES

Brown, J. F., T. R. Loveland, J. W. Merchant, B. C. Reed, and D. O. Ohlen, 1993. "Using MultiSpectral Data in Global Landcover Characterization: Concepts, Requirements, and Methods," Photogrammetric Engineering and Remote and Remote Sensing, Vol. 59, No. 6, pp. 977-987.

Doraiswamy, P. ,Hart, G., Craig, M., Cook, P., 1994. The Anomalous '93 Growing Season -- How USDA Used AVHRR Data. Proceedings of the 60th ASPRS, 1:144-151, Reno, Nevada.

Doraiswamy, P. C. and P. Cook 1995. "Spring Wheat Yield Assessment Using NOAA AVHRR Data," Canadian Journal of Remote Sensing, Vol. 21, No. 1, pp. -

Eidenshink, J. C. 1992. "The 1990 Conterminous U.S. AVHRR Data Set," Photogrammetric Engineering and Remote Sensing, Vol. 58, pp. 809-813.

Kotz, S. and Johnson, N. 1982. Encyclopedia of Statistical Sciences, Vol. 2, pp 218-219.

Statistical Sciences, Inc. 1995. S-PLUS User's Manual, Version 3.3 for Windows, Seattle.

Stern, A., Zora, and Doraiswamy, P. "Spring Wheat Yields from Landsat TM Data in North Dakota", in progress.

GRAPHS

Figure 1. North Dakota Spring Wheat County Acreage (1989) vs. Number of AVHRR Spring Wheat Pixels Per County.

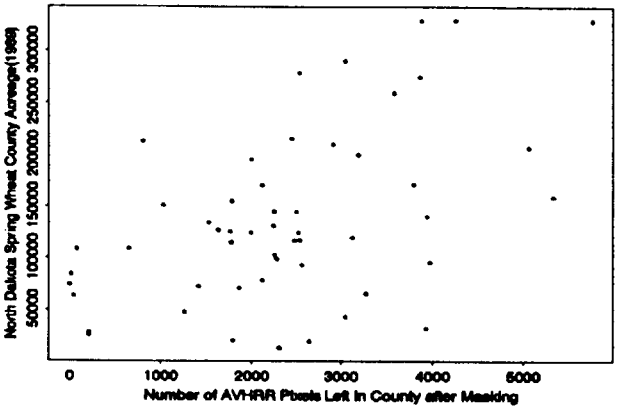

Figure 2. Ranges of North Dakota Spring Wheat County Yields for 1989-1992 by Crop Reporting Districts.

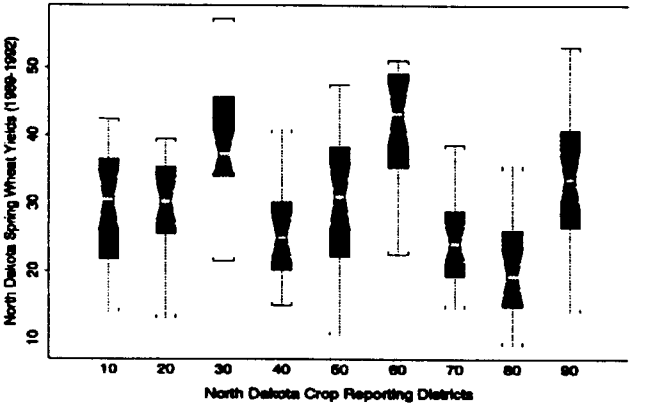

Figure 3. Four Yearly North Dakota Spring Wheat County Yields vs. AVHRR Sums of County Averages for Periods 20 to 26.

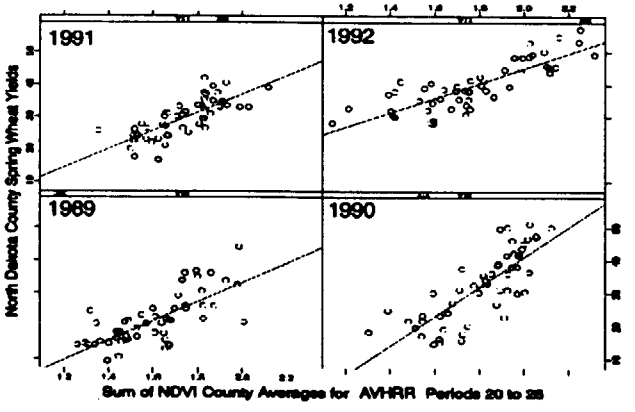

Figure 4. North Dakota Spring Wheat County Yields (1989-1992) vs. AVHRR Sums of County Averages for Periods 20 to 26.