ON THE USE OF CORRELATIONS IN CROP YIELDS

David Pawel and Ron Fecso, National Agricultural Statistics Service
David Pawel, USDA-NASS, Room 4801-South, Washington, D.C. 20250

ABSTRACT

The National Agricultural Statistics Service (NASS) uses a separate sampling, imputation, and estimation approach for each objective yield (OY) crop to produce yield statistics. An exploratory analysis indicates that strong correlations exist between the official U.S.-level estimates of yields of several crops. A principal component analysis of these correlations indicates potential useful interpretations of regional and seasonal planting, growth and harvesting patterns. Spatial correlations between state-level objective yield estimates of soybeans and corn are strong and tend to decrease with respect to the distance between states. Consideration of spatial and between-crop correlations may lead to improvements in the survey design, process control, and methodology used for the estimation of production and yield.

1. INTRODUCTION

The art of survey design relies upon the recognition of relationships within known data and insight into relationships that may exist in data to be collected. Because relationships among crop yields almost surely exist, we sought answers to the following questions.

1) What precisely are these relationships?

2) How can these relationships be quantified?

3) How can the National Agricultural Statistics Service (NASS), U.S. Department of Agriculture, use knowledge of these relationships to improve the survey design and ultimately the estimates of quantities such as yield and production?

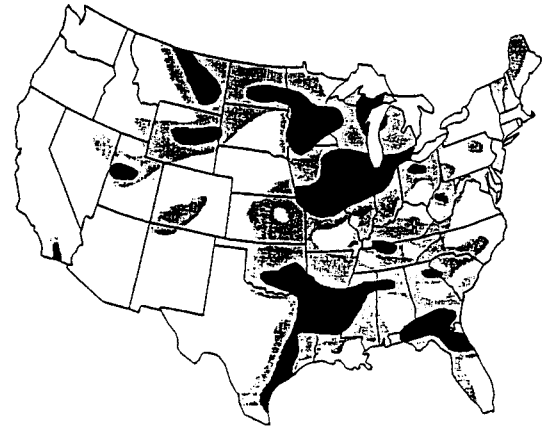
A look back at this year's drought should help to describe two types of relationships that we postulated to exist among crop yields. The "drought" map on this page illustrates how neighboring states such as Iowa and Illinois, or farther north, Minnesota and North Dakota, suffered from practically identical dry weather conditions. In general, crop yields in neighboring states often either benefit or suffer from similar weather conditions, because weather events often affect large contiguous regions.

This suggested the formulation of two hypotheses, 1) crop yields from neighboring states should tend to move in the same direction, and 2) yields of crops grown in similar regions should tend to move in the same direction, since crops grown in overlapping regions, such as corn and soybeans, will be exposed to similar growing conditions.

Our analysis of historic yield data supported these hypotheses. It showed strong correlations in crop yields **between states** (hypothesis 1), and **between crops** (hypothesis 2) that were consistent with relationships in regional and seasonal crop characteristics.

We believe that knowledge of the two types of correlations may be used to improve major components of our survey program by borrowing information collected in one state to improve estimates in a neighboring state, and by improving the estimate of yield of one crop (such as corn) by borrowing information gathered on a related crop (such as soybeans). Program areas to consider would include survey design, small area estimation, and process control. Kriging and procedures based on James-Stein type estimates are two types of methodologies that suggest this type of borrowing of information is feasible. We hope that the presentation of these

Figure 1. Percentage of Normal Precipitation, April 1 to July 31, 1988



Black: Less than 50%
Grey: 50 to 75%

Climate Analysis Center NOAA

results will motivate research on how NASS can use such techniques to improve its estimates of crop yields and production.

2. NASS CROP YIELD ESTIMATES

Information used to establish official crop yield estimates comes from at least two sources.

1) Mail surveys in which farmers are asked questions about crop conditions, expected and realized yields, and so on.

2) Objective yield (OY) surveys in which measurements are made in the field on small parcels of land which we will call plots. A typical plot may be a rectangle, three feet long and two rows wide. Plots are chosen through a multi-staged survey design which, at least in theory, ensures that estimates from the OY program are essentially unbiased. NASS conducts OY surveys of corn, cotton, wheat, soybeans, potatoes, and rice in states that are major producers of these crops. Further details concerning the survey design may be found in Francisco, Fuller, and Fecso (1987).

3. DATA ANALYSIS - BETWEEN CROP CORRELATIONS

In the next two sections, we describe the relationships **between crop** yields Results of our analysis of the data series of the 1971-85 official U.S. level estimates for corn, sorghum, cotton, soybeans, potatoes, sunflowers, sugar beets, winter wheat, durum wheat, spring wheat, and rice are presented. **Between state** relationships in crop yields are discussed in a later section.

The fifteen year data series of official U.S.-level crop yield estimates used for the analysis is shown in Table 1. We felt that fifteen years of data reflected a reasonable compromise between using data from too long a time interval (in which case the analysis would be complicated by significant changes in basic conditions), and too short an interval (in which case the amount of data would be insufficient). The yield estimates in Table 1 were regressed against time using a polynomial model to take into account possible trends. Rice and the three types of wheat required a quadratic term that was difficult to interpret, and suggested that a more complex time series model may be appropriate. For the purposes of this exploratory study, these four crops were eliminated. For all other crops, Durbin-Watson tests indicated that autocorrelation would not be a serious problem.

Correlations were calculated for the residuals associated with the eight remaining crops. The correlations are close to 1 for several pairs of crops, and uniformly positive for almost all pairs. The highest correlations and some moderate correlations are between crops that seem to be the most agronomically related. Table 2 shows high correlations (in bold print) between corn and sorghum (.87), soybeans and corn (.82), soybeans and sorghum (.81), and cotton and sorghum (.77). Moderate correlations exist between sorghum and sunflowers (.65), sugar beets and soybeans (.59), cotton and corn (.57), sunflowers and corn (.57), peanuts and cotton (.55), soybeans and potatoes (.55), and soybeans and cotton (.55). P-values for these correlations were calculated under the assumptions that the random components of the crop yields were normal and i.i.d.. Although only 15 data points were used, the p-values for the eleven highest correlations, mentioned here solely for descriptive purposes, are very small, ranging from .0001 to .0355. All of these crops are "significantly" correlated with at least one other crop. Uncorrelated pairs of crops, such as potatoes and peanuts, sugar beets and sunflowers, and peanuts and sunflowers, are usually grown in separate regions.

To get some indication of the sensitivity of the analysis to outliers and violations of model assumptions, two alternative approaches were employed to calculate the correlations among the official U.S.-level yield estimates. Both used first-differences to eliminate the effect of trend. First-differences were defined to be the year-to-year increases (decreases) in crop yields. In the first approach, correlations among the official U.S.-level yield estimates were calculated directly from these first-differences. In the second approach, the Spearman rank correlation coefficients of the first-differences (correlations of the ranked first-differences) were calculated. Spearman rank correlation coefficients are in general more resistant to outliers (such as the 1980 official U.S.-level peanut yield estimate).

Both approaches relied on the assumption that the 14 (ranked) first-differences associated with an individual crop are uncorrelated. Although, as expected, there appeared to be some negative autocorrelation in both the raw and ranked first-differences for many if not all of the crops, we felt the autocorrelations were small enough for the purposes of our sensitivity analysis.

Correlations resulting from all three methods, shown together in Table 2, are remarkably similar. In particular, the correlations between corn and sorghum (.87 to .75), corn and soybeans (.82 to .73), sorghum and cotton (.77 to .87),

Table 1. Official U.S.-level yield estimates for several crops, 1971-1985.

Year	Crops											
	Corn	Sorghum	Soybean	Winter Wheat	Durum Wheat	Other Spring Wheat	Cotton	Peanut	Sunflowers	Rice	Potatoes	Sugar beets
	Bushels/acre						lbs/acre				Cwt/acre	Tons/acre
71	88.1	53.8	27.5	35.4	32.1	30.7	438	2066	1050	4718	230	20.2
72	97.0	60.7	27.8	34.0	28.6	29.0	507	2203	916	4700	236	21.4
73	91.3	58.8	27.8	33.0	27.2	28.3	520	2323	1080	4274	230	20.1
74	71.9	45.1	23.7	29.4	19.8	22.4	442	2491	957	4440	246	18.2
75	86.4	49.0	28.9	32.0	26.4	26.8	453	2564	1109	4558	256	19.6
76	88.0	49.1	26.1	31.5	29.4	26.8	465	2464	1058	4663	261	19.9
77	90.8	56.6	30.6	31.6	26.4	28.6	520	2456	1252	4412	261	20.6
78	101.0	54.5	29.4	31.8	33.1	30.0	420	2619	1365	4484	267	20.3
79	109.5	62.6	32.1	36.9	27.1	28.2	547	2611	1349	4599	272	19.6
80	91.0	46.3	26.5	36.8	22.4	25.3	404	1645	1016	4413	265	19.8
81	108.9	64.0	30.1	35.9	32.4	30.6	542	2675	1177	4819	276	22.4
82	113.2	59.1	31.5	36.0	34.9	33.8	590	2693	1129	4710	280	20.3
83	81.1	48.7	26.2	41.8	29.3	31.7	508	2399	1044	4598	269	19.9
84	106.7	56.4	28.1	40.0	32.1	35.3	600	2878	1014	4954	279	20.2
85	118.0	66.7	34.1	38.1	36.4	35.4	630	2810	1109	5437	298	20.5

Source: Agricultural Statistics - 1986

Table 2. Comparison of correlations of raw data, first-differences, and ranked first-differences of the official U.S.-level yield estimates for several crops, 1971-1985.

Crops	Sorghum	Cotton	Soybeans	Potatoes	Peanuts	Sugar beets	Sun flowers
Corn	**.87** [1] (.83)[2] .75 [3]	**.57** (.71) .56	**.82** (.78) .73	**.41** (.59) .64	**.31** (.62) .54	**.37** (.47) .59	**.57** (.50) .42
Sorghum	**1.0**	**.77** (.87) .83	**.81** (.84) .83	**.24** (.47) .50	**.39** (.69) .30	**.32** (.64) .59	**.65** (.64) .54
Cotton	-	**1.0**	**.55** (.73) .64	**.04** (.36) .30	**.55** (.66) .41	**-.06** (.35) .32	**.32** (.45) .34
Soybeans	-	-	**1.0**	**.55** (.52) .50	**.36** (.50) .23	**.59** (.40) .50	**.41** (.72) .67
Potatoes	-	-	-	**1.0**	**.42** (.53) .59	**.48** (.18) .34	**.04** (.25) .12
Peanuts	-	-	-	-	**1.0**	**.31** (.33) .16	**.04** (.60) .29
Sugar beets	-	-	-	-	-	**1.0**	**.11** (.27) .27

[1] Correlations of residuals resulting from regressing yield against time are shown in bold type.
[2] Correlations of first-differences are shown in parentheses.
[3] Spearman rank-correlations of first-differences are shown last.

soybeans and sorghum (.81 to .84) are uniformly high, and as will be shown in the next section, are consistent with agronomic realities.

4. PRINCIPAL COMPONENT ANALYSIS

Table 3 exhibits the first four eigenvectors and corresponding eigenvalues of the correlations of the residuals resulting from regression of yield against time. The principal component analysis indicated three factors account for approximately 85% of the variation inherent in the official U.S.-level yield estimates of the eight crops. The components of all three factors seemed to separate the crops into groups, so that within each group, the crops share similar growing conditions and needs.

In some sense, the first factor lumps all crops into one universal group. All of the components of the first factor are positive, which must be due to the relatively high and uniformly positive correlations that exist between yields of all crops. Perhaps the high correlations reflect the likelihood that growing conditions affecting the yields of most crops will be similar, because many of these crops are grown in overlapping regions.

Further, the first factor has components which fall into three groups in a manner consistent with agronomic realities. For corn, sorghum, and soybeans, the components are all close to .9, for cotton the component is .7, and for potatoes, peanuts, sugar, and sunflowers, the components are all between .5 and .6. The growing needs and conditions of the first three crops are very similar; both soybeans and sorghum thrive under environmental conditions favorable to corn production

Table 3. The first four eigenvectors and eigenvalues of correlations of residuals resulting from regression of yield against time.

	Eigenvector			
Crop	1	2	3	4
Corn	.90	-.11	-.18	.11
Sorghum	.93	-.30	-.05	-.07
Cotton	.70	-.47	.49	-.05
Soybeans	.91	.16	-.14	-.03
Potatoes	.51	.69	.03	.51
Peanuts	.56	.24	.70	-.08
Sugar	.51	.66	-.27	-.46
Sunflowers	.58	-.45	-.47	.10
Eigenvalue	**4.17**	**1.51**	**1.08**	**.51**

(Chapman, et. al., 1976). All three are considered to be warm-weather crops that require substantial amounts of moisture, and relatively specific lighting conditions. The three crops are grown during similar parts of the year, and soybeans and corn are typically grown in the same region (in midwestern states such as Iowa, Illinois, and Indiana).

As for the second factor, the loadings for cotton, sorghum, corn, and soybeans are almost consecutive ranging from -.47 to .16, and the ordering induced by these loadings places sorghum in between cotton and corn. Both cotton and sorghum have major crop growing areas in Texas. The loadings also separate sugar beets and potatoes from the remaining crops. Both crops are grown predominantly in states on the northern and western edges of the country.

The third factor may represent a regional effect. It separates the crops into three groups: the primarily central state crops corn, sorghum, and soybeans, the primarily Southern crops, cotton and peanuts, and the northern grown crops, sugar beets and sunflowers.

In general, the principal component analysis supports the assertions that (1) official U.S.-level crop yield estimates are in general positively (and for several pairs of crops strongly) correlated, and (2) correlations between the official U.S.-level crop yield estimates reflect regional growing conditions.

5. DATA ANALYSIS - SPATIAL CORRELATIONS (BETWEEN STATES)

The purpose of the second part of the analysis is to introduce spatial concepts by describing between-state relationships in crop yields. We examined the data series of the 1973-1986 soybean and corn end-of-season objective yield estimates for Nebraska, Minnesota, Iowa, Missouri, Ohio, Illinois, and Indiana -- the only states that have OY programs for both crops. The data series were detrended by calculating first differences.

To help describe the between-state correlations in crop yields, we introduced a crude distance measure. We defined the distance between two states to be the minimum number of state boundaries that must be crossed to travel from one state to the other.

For those familiar with geostatistics, the graphs that follow are similar in concept to semivariograms (Davis, 1986). For both soybeans and corn, correlations in the state-level crop yield estimates are graphed against the distance measure. For both crops, the correlations are quite high for neighboring states, and

tend to decrease with respect to the distance between states.

6. DISCUSSION

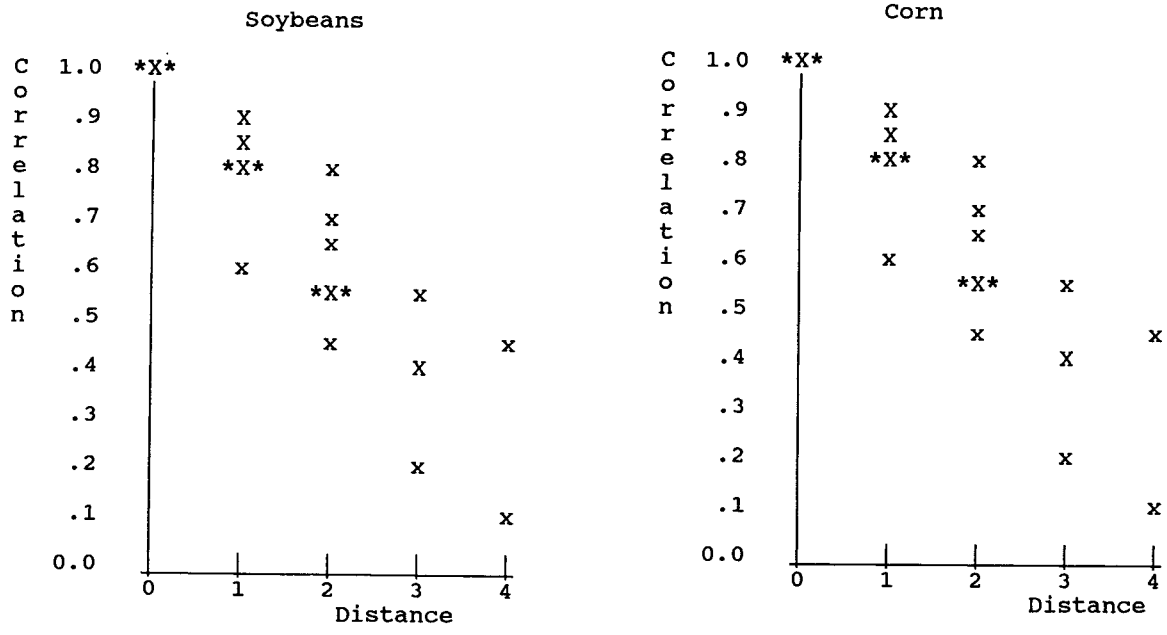
Although the analysis only demonstrates the presence of correlations between yield estimates, it is assumed, because of the effect of yield on the decision-making process concerning crops to be planted, that correlations associated with other quantities, such as production and acreage, may also exist. In this section, an attempt is made to provide a broad outline for the potential use of correlations among quantities such as yield and production to improve our survey design, estimation methodology, and survey process control.

A basic question that can be asked about our survey design is: Which states should be included in our OY programs? At present, ten states, Iowa, Illinois, Nebraska, Indiana, Michigan, Minnesota, Ohio, Wisconsin, Missouri, and South Dakota, conduct corn OY surveys. There is a good reason why these states were chosen. Every year, each of these states are among the top ten or eleven corn producing states in the nation. The ten states account for approximately 85% of U.S. corn production. But there is a drawback to this strategy. Corn is grown in almost every state in the U.S., while the ten corn OY states are all located in the Midwest.

This leads us to consider the survey design itself. The between-state correlations of Figure 2 tell us that the states in the current corn OY program may be too clustered. For example, Texas, ranked eleventh in corn production does not conduct a corn OY survey and has no close neighbor to draw information from while South Dakota ranked tenth has strongly correlated information from the Nebraska, Iowa and Minnesota OY surveys for use in further improving its estimate. We thus become interested in the relative efficiency of spatial estimation techniques using a design which spaces OY states rather than clustering or one with fewer OY samples per state but more states. Further, since some states do not have corn OY surveys but do conduct OY surveys for other crops, the between crop correlations may be useful when combined with the spatial estimation techniques to "expand" survey coverage.

As for small area estimation, the geographical interpretation of the principal components suggests that the analysis continue with the examination of crop yield correlations at the crop district and county levels. The reason given for strong between-crop correlations at the national level would seem to apply at state and substate levels; crops are often grown in well-defined overlapping regions. This point is illustrated by the maps on the next page that show how within Texas, sorghum and cotton are grown in heavily concentrated and remarkably similar regions. Kriging or ratio and regression techniques (where ratios and regression coefficients could be estimated between crops) might be used to increase the precision of yield and production estimates at substate

Figure 2. Correlations of state-level soybean and corn objective yield estimates graphed against distance.



levels where traditional expansion estimates fail because of inadequate sample sizes. Further analysis is needed to make sure that for small regions, spurious fluctuations in crop yields do not dominate the overall spatial correlation structure.

Benefits to the NASS process control program might be found by using correlations in OY statistics and official U.S.-level estimates to identify outlying yield values. Roughly put, the high correlations in the official U.S.-level estimates indicate it is unlikely that a large change in the yield of one crop would not be accompanied by a similar change in the yield of some other crops. Similarly, it is unlikely that yield in a state would dramatically increase when the yields in neighboring states decrease. Formal procedures involving methods such as the construction of multivariate prediction intervals may be used to identify data and estimates that appear to be inconsistent. If correlations are high for production and acreage, then outlying production and acreage estimates may be identified in a similar manner. An agronomic explanation should be offered for any set of final estimates that appear to be inconsistent with correlations implied by past experience.

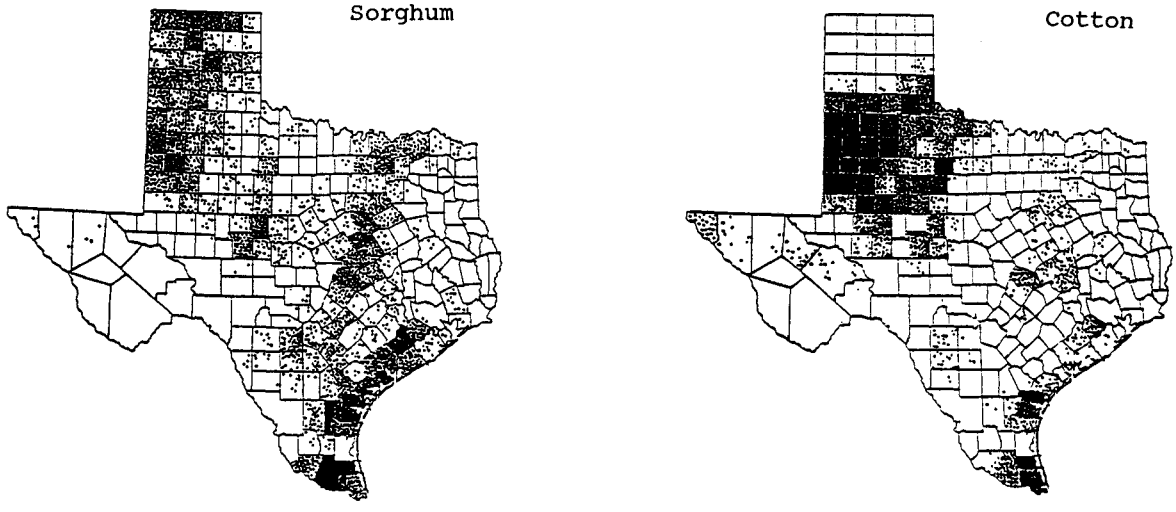
Acknowledgements: The authors would like to thank Ben Klugh for calling to our attention the high between-crop correlations among the first-differences of the crop yields. We would also like to thank Joyce Little[1] for her significant contribution to our analysis of the official U.S.-level yield estimates.

[1] Joyce is a former student assistant with the Nonsampling Errors Research Section of NASS.

6. REFERENCES

1. Chapman, S.R. and Carter, L.P., Crop Production, Principles and Practices, W.H. Freeman & Co., San Francisco, 1976.
2. Davis, J.C., 1986. Statistics and Data Analysis in Geology. John Wiley and Sons: New York, pp. 239-248.
3. Francisco, C., Fuller, W., and Fecso, R. "Statistical Properties of Crop Production Estimators", Survey Methodology, Vol. 13, No. 1, 1987.
4. Pawel, D., Fecso, R., Little, J., "An Examination of Correlations in Board Reported Yields of Several Commodities", U.S. Department of Agriculture, (submitted for review, 1988).
5. U.S. Department of Agriculture, Agricultural Statistics, 1986.
6. United States Department of Agriculture, Texas Crop and Livestock Reporting Service, Texas Field Crop Statistics, 1984.

Figure 3 - Area planted for sorghum and cotton in Texas in 1984



Each dot = 1000 acres

Source: Texas Field Crop Statistics, 1984