United States
Department of
Agriculture

National
Agricultural
Statistics
Service

Research and
Applications
Division

SRB Staff Report
Number SRB-89-02

March 1989

An Examination of Correlations in Board Reported Yields for Several Commodities

David Pawel
Ron Fecso
Joyce Little

AN EXAMINATION OF CORRELATIONS IN BOARD REPORTED YIELDS OF SEVERAL COMMODITIES. By David Pawel, Ron Fecso, and Joyce Little[1]; Research and Applications Division, National Agricultural Statistics Service; U.S. Department of Agriculture, Washington, D.C. 20250; March, 1989. Staff Report No. SRB-8902.

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 among national board estimates of yields of several commodities. If these correlations in national board estimates accurately portray correlations between the actual yields of commodities, then these correlations could be used to improve the estimation of yields. Investigation into these improvements could involve at least three areas of survey design: sample selection (includes stratification), estimation, and process control. It is recommended that the source of these correlations be determined, and possible relationships between national, state, and segment-level objective yield statistics be studied.

****
* *
* This paper was prepared for limited distribution to the re- *
* search community outside the U.S. Department of Agriculture. *
* The views expressed herein are not necessarily those of NASS *
* or USDA. Use of company names in this publication is for *
* identification only and does not imply endorsement by the *
* Department of Agriculture. *
* *
****

ACKNOWLEDGEMENTS

The authors would like to thank Ben Klugh for his helpful suggestions. In particular, he called to our attention the high correlations that exist between the raw first-differences. We would also like to thank Read Johnston for his patience and unselfishness regarding many aspects of the production of this report.

[1] At the time of this writing, Joyce Little was a student assistant with the Nonsampling Errors Research Section assigned to the Missouri SSO.

CONTENTS

LIST OF TABLES

FIGURES

SUMMARY

A separate sampling, imputation, and estimation approach for each objective yield (OY) crop is being used by the National Agricultural Statistics Service (NASS) to produce yield statistics. The separate sampling approach ignores possible relationships between the yields of commodities which may exist due to similarities in plants. The main purpose of this investigation was to determine whether a multivariate approach to the sampling and estimation of quantities such as yield and production deserves consideration. A secondary purpose of presenting this report is to introduce some concepts of multivariate analysis to a broad agency audience with a simple relevant example. Our exploratory analysis indicates that strong correlations exist between national board estimates of several commodities. Correlations between the national board estimates of corn, sorghum, and soybeans ranged from .77 to .87. An interpretation of correlations among national board estimates of eight crops is given through the use of principal components. If the yields of two crops are correlated, then in the estimation of the yield of one of the crops, the data obtained on the other crop is auxiliary information. "When an auxiliary variable is highly correlated with the characteristic under study, the estimate of the population mean (total) of this characteristic can be improved using the auxiliary variable," (Okafor, 1987, p.110). It would seem to follow that if the correlations observed in the national board estimates accurately portray correlations between the actual yields of commodities, then corresponding correlations in objective yield statistics should be used to improve the estimation of yields. It may also be possible to use correlations in yield statistics to improve our sample selection and process control procedures. At the very least, the source of the correlations in the board estimates should be determined, and possible relationships between national, state, and segment-level objective yield statistics should be studied. Continued analysis is recommended specifically extending the study to determine the nature of correlations associated with other quantities such as production and acreage, and examining statistical methods that may take advantage of such correlations.

AN EXAMINATION OF CORRELATIONS IN BOARD REPORTED YIELDS OF SEVERAL COMMODITIES

By David Pawel, Ron Fecso, and Joyce Little

INTRODUCTION

A separate sampling, imputation and estimation approach for each objective yield (OY) crop is being used by the National Agricultural Statistics Service (NASS) to produce yield statistics. At recent specification meetings, discussions have indicated a desire to standardize the survey operations between crops, a survey design choice, which seems to imply that direct relationships exist. The separate sampling approach ignores possible relationships among the yields of commodities which may exist due to similar characteristics exhibited in many plants such as corn and sorghum, or wheat and rice. Since many crops have similar growing conditions and needs, they are often planted in the same areas and during similar times of the year. This decreases the likelihood that adverse growing conditions would (or would not) affect the yield of only <u>one</u> crop. (Such a situation could occur if, for example, only one of the crops is irrigated).

This exploratory analysis of national board estimates of several commodities was predicated upon the hypothesis that relationships among crop yields can be used to improve our estimates of yield and production. National board estimates were chosen for the analysis for at least three reasons: 1) they are free, 2) they were quickly and easily obtainable, and 3) relationships that could be useful at a national level would likely hold at state levels. The motivation for this exploratory analysis may be best described by a quote from W.E. Deming, "The only useful function of a statistician is to make predictions, and thus provide a basis for action," (found in Wallis, 1980). A main purpose of this report is to provide a partial answer as to whether a multivariate approach to the sampling of quantities such as yield and production deserves consideration. A secondary purpose of presenting this material as a staff report is to introduce the concepts of multivariate analysis to a broad agency audience with a simple, relevant example. It is our hope that the example will show the power of techniques for exploratory analysis and improve the understanding of future research that we will propose.

Our exploratory analysis indicates that strong correlations exist among the national board estimates of the yields of several commodities. An interpretation of these correlations, given through the use of principal components, suggests that these correlations may reflect regional and seasonal characteristics that relate to the planting, growth and harvesting of these commodities. If these correlations, observed in the national board estimates, accurately portray relationships that exist among actual yields, then these relationships can and should be used to improve our

estimates of yields. This would be achieved by making appropriate adjustments in procedures related to one or more of the following areas: sample selection (includes stratification), estimation, and process control.

An example of a (simplified) ratio-type estimate that may be proposed as a way to exploit correlations among yields will be given in a later section. The rationale for the consideration of a ratio-type estimate is: "When an auxiliary variable is highly correlated with the characteristic under study, the estimate of the population mean (total) of this characteristic can be improved using the auxiliary variable," (Okafor, 1987, p.110). If the yields of two crops are highly correlated, data obtained on one crop may be treated as auxiliary information to be used to improve the estimate of the yield of the other crop.

We will also hint at possible directions regarding sample selection and process control.

ANALYSIS

The national board yield estimates for the years 1971-85 for corn, sorghum, cotton, soybeans, potatoes, sunflowers, sugar beets, winter wheat, durum wheat, spring wheat, and rice were chosen for the analysis and are shown in Table 1. It was 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 variables were eliminated, but it should be noted that these crops are in some sense similar to each other and different from the remaining crops. This may be a statistical "warning" that these crops need differing procedures in survey design, forecasting, and/or estimation. Durbin-Watson tests for autocorrelation were computed for the other variables. The critical points for the test were 2.64 for the lower boundary and 2.92 for the upper boundary (Neter and Wasserman, 1974, p.358, 816). All of the Durbin-Watson statistics were between 2.411 and 1.895 which indicates autocorrelation is not a problem in the subset of crops used in subsequent analyses.

To help meet multivariate model requirements, the residuals for the eight variables that required only a linear term were used in this analysis. Table 2 shows high correlations between corn and sorghum (.87), soybeans and corn (.82), soybeans and sorghum (.81), and cotton and sorghum (.77). Moderate correlations also 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). Although only 15 data points were used, the p-values for these correlations were very small, ranging from .0001 to .0355. In fact, most of the crops show significant correlations with at least one other crop, with most uncorrelated pairs being regionally separated crops such as potatoes vs. peanuts, sugar beets vs. sunflowers, and peanuts vs. sunflowers.

SENSITIVITY TO MODEL ASSUMPTIONS AND OUTLIERS

To get some indication of the sensitivity of the analysis to the validity of model assumptions and outliers, two alternative approaches were employed to calculate the correlations among the board yields. The validity of the correlations shown in Table 2 would be substantiated through similar correlations produced through the alternative approaches. Both use first-differences to eliminate the effect of trend. First-differences are defined to be the year-to-year increases (decreases) in crop yields. For example, the 14 first-differences for potatoes are 6, -6, 16, 10, 5, 0, 6, 5, -7, 11, 4, -11, 10, 19. In the first alternative approach, correlations among the board yields are calculated directly from these first-differences. In the second alternative approach, the Spearman rank correlation coefficients of the first-differences (correlations of the ranked first-differences) are calculated. The Spearman rank correlation coefficients would, in general, be more resistant to outliers (such as, for example, the 1980 board estimated peanut yield).

For both alternative approaches, the assumption is made that the 14 (ranked) first-differences associated with an individual commodity are uncorrelated. As may be expected, there does appear to be some negative autocorrelation in both the raw and ranked first-differences for many if not all of the commodities. The Durbin-Watson statistics were calculated for both sets of first-differences, and were found to be not significant for potatoes and sugar, significant for cotton, and inconclusive for the remaining five crops. Autocorrelations for raw first-differences ranged from -.156 for potatoes to -.646 for cotton. Five of the crops had autocorrelations between -.375 and -.5. The autocorrelations for the ranked first-differences were similar. It was felt that these autocorrelations were sufficiently small for the purpose of evaluating the sensitivity of correlations calculated as described in the previous section.

Correlations resulting from all three methods are shown together in Table 3. It is evident that the three sets of correlations 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), soybeans and sorghum (.81 to .84) are consistently high. These numbers certify (almost surely) that the board estimates for these commodities are strongly and positively

related. If these correlations, observed in the board estimates, accurately portray relationships that exist among the actual yields of these commodities, then fundamental sampling would mandate the use of these relationships to improve our estimates.

PRINCIPAL COMPONENT ANALYSIS

Table 4 exhibits the eigenvectors and corresponding eigenvalues of the correlation matrix given in Table 2. The principal component analysis indicates three factors account for approximately 85% of the variation inherent in the board yield estimates of the eight commodities. All of the components of the first factor are positive, which is undoubtedly due to the relatively high and uniformly positive correlations that exist between all crops. This would seem to indicate that some of the information that may be gleaned from yield surveys relates to conditions that affect the yields of <u>all</u> crops. One of the reasons for the relative uniformity of these conditions must simply be that many of these crops are grown in the same regions. In figures 1 to 3, it can be seen that within Texas, soybeans, sorghum, and cotton are all grown in similar regions of the state.

Of course, it is important to note that the components of the first factor are not identical, and in fact seem to fall into three groups. Associated with 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. Certainly, the growing needs and growing conditions of these three crops are very similar; both soybeans and sorghum thrive under environmental conditions favorable to corn production (Chapman, et. al., 1976). All three are considered to be warm-weather crops that require substantial amounts of moisture, and relatively specific lighting conditions. Not surprisingly, 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). Of the remaining crops, cotton in particular seems to share many of the characteristics common to this group of three crops.

As for the second factor, it is interesting to note that 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 also 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.

Factors four (separates potatoes from sugar beets) and five (loads heavily on sunflowers and peanuts) account for only a small portion of the variance of the system, but may reflect conditions governing specialty crops, and thus may prove to be important in smaller area analyses.

Eigenvalues and eigenvectors were also calculated using the correlations of the raw first-differences to get some indication of the sensitivity of the principal component analysis to violations in model assumptions. These are shown in Table 5. The first factor is essentially unchanged; other factors, as may be expected, given their smaller magnitude, do show changes. To summarize, the principal component analysis of both sets of correlations supports the assertions that (1) board crop yield estimates are in general positively (and for several pairs of crops strongly) correlated, and (2) board yield estimates, to some extent, seem to mirror agronomic realities by grouping crops such as corn, sorghum, and soybeans together.

USES

Although the analysis only demonstrates the presence of correlations between board 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 between commodities. These relate to at least three areas associated with our survey design: stratification, estimation, and process control.

The geographic interpretations of the principal components leads to the postulate that correlations obtained at the state level may also be highly significant. As this analysis may have demonstrated, principal component analysis is often a very useful tool for identifying possible sources of variation. If in the future, yield data were to be used to stratify the area frame to increase the efficiency of state and national-level production estimates, then it would seem possible to use principle component analyses of correlations obtained from state-level objective yield data to simplify the stratification process. This would not be the first time principal components were used or proposed for stratification. In the 1940's, the Bureau of Agricultural Economics considered the use of principal components of control (auxiliary) variables to stratify counties to be periodically surveyed for social and economic studies (Hagood and Bernert, 1945). The first principle component of twelve of these auxiliary variables was used to separate the counties into five strata. Then, the second principle component was used to subdivide each of these strata. In effect, this type of analysis could be used to

effectively reduce the number of yield variables to be considered for stratification.

For example, consider a state where soybeans, corn, and wheat are grown. A principal component analysis of the yield data may identify two eigenvectors that would predominate: the first might be one that, in effect, may be associated with the average yield for all crops, the second may be associated with the difference in yield between wheat and the two crops, soybeans and corn. Not only would this simplify the problem by reducing the problem to one that involves only two yield variables, but this may also result in fewer strata.

The existence of high correlations between the board yield values indicates that, perhaps, regression and ratio estimation methods (Cochran, 1977, p.150), or new methods yet to be developed may be used by NASS to increase the precision of its estimates. An example of such a ratio-type estimate is outlined below. For the purpose of exposition, the simplifying assumptions are made that within each state, the land upon which (say) soybeans and corn are grown is divided into N segments of equal size, and that yields are determined with complete precision for separate simple random samples of n segments, selected for both crops. Of course, the design used by our objective yield survey is multi-staged, and in particular, yield statistics are obtained for zero to four samples within each segment included in the June Enumerative Survey sample. However, the underlying concepts of ratio estimates for one-stage and multi-staged designs are similar. For a general discussion of more complicated ratio-estimation techniques, please see Sukhatme and Sukhatme (p.289-312).

Let

m = the number of segments for which yields are determined for both crops.

y_m = average corn yield per segment for the m segments for which yield is determined for both crops.

x_m = average soybean yield per segment for the m segments for which yield is determined for both crops.

y_n = average corn yield per segment for the n segments for which corn yield is determined.

x_n = average soybean yield per segment for the n segments for which soybean yield is determined.

a, b be constants so that $a+b = 1$.

Ratio-type estimates for corn yield would be of the form:

estimate of total corn yield = $N\ (a * y_m/x_m * x_n + b * y_n)$.

Both a and b would be chosen with the objective of minimizing the mean-squared error of the estimate. An estimate of this kind should be considered if the coefficient of variation associated with the overall estimate of corn yield is as large as the coefficient of variation of the soybean estimate, and the segment-level correlation between soybean and corn yields is sufficiently high.

Of at least equal importance are the possible benefits to the NASS process control program. This should begin with an identification of the source of the correlations observed in the board yield estimates. Similar correlations, if observed in objective yield statistics, may then be used to identify outlying estimates of yield values. Roughly put, such high correlations would indicate that 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. Thus, whenever it appears that one or more estimates are behaving inconsistently, those estimates and the data used to support those estimates should be investigated. Formal procedures, such as the construction of multivariate prediction intervals should be used to identify such outliers. If correlations are also 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.

CONCLUSIONS AND RECOMMENDATIONS

This study shows significant correlations between the national board yield estimates of several commodities. It should be determined whether these correlations accurately portray relationships that exist among actual yields, and whether these correlations are present in objective yield statistics in sufficient strength at state and segment levels. The study should be extended to investigate the nature of correlations associated with other quantities such as production and acreage. The potential use of these correlations would relate to at least three areas of survey design: stratification, estimation, and process control. For example, principle components of the correlation matrix may be used for stratification of the area frame. The correlations may be used to optimize the statistical properties of regression or ratio type estimators, or to determine multivariate confidence intervals for the identification of outliers. A multivariate approach to survey design at NASS deserves consideration. Continued analysis is recommended to study the relationship of survey and board data in several geographically distinct states.

REFERENCES

1. Chapman, S.R. and Carter, L.P., Crop Production, Principles and Practices, W.H. Freeman & Co., San Francisco, 1976.

2. Cochran, W.G., Sampling Techniques, John Wiley & Sons, New York, 1977.

3. Hagood, M.J. and Bernert, E.H. "Component Indexes as a Basis for Stratification in Sampling", Journal of the American Statistical Association, 40, 231, 1945, p.330-341.

4. Neter, J. and Wasserman, W., Applied Linear Statistical Models, Richard D. Irwin, Inc., Homewood, Ill., 1974.

5. Okafor, F.C., "Comparison of Estimators of Population Total in Two-Stage Successive Sampling Using Auxiliary Information", Survey Methodology, 13, 1, 1987, p.109-121.

6. Snedecor, G.W., Cochran, W.G., Statistical Methods, Iowa State University Press, Ames, Iowa, 1980.

7. Sukhatme, P.V., and Sukhatme, B.V., Sampling Theory of Surveys with Applications, Iowa State University Press, Ames, Iowa, 1970.

8. United States Department of Agriculture, Agricultural Statistics, 1986. .

9. United States Department of Agriculture, Texas Crop and Livestock Reporting Service, Texas Field Crop Statistics, 1984.

10. Wallis, W.A., "The Statistical Research Group, 1942-1945", Journal of the American Statistical Association", 75, 370, 1980, p.321.

Table 1 - Board yield estimates for several commodities,

1971-1985.

Year	Commodities											
	Corn	Sorg-hum	Soy-beans	Cot-ton	Pea-nuts	Pota-toes	Sugar beets	Sun-flowers	Winter Wheat	Durum Wheat	Spring Wheat	Rice
	- Bushels/acre -			- lbs/acre -		Cwt/acre	Tons/acre	lbs/acre	- Bushels/acre -			lbs/acre
71	88.1	53.8	27.5	438	2066	230	20.2	1050	35.4	32.1	30.7	4718
72	97.0	60.7	27.8	507	2203	236	21.4	916	34.0	28.6	29.0	4700
73	91.3	58.8	27.8	520	2323	230	20.1	1080	33.0	27.2	28.3	4274
74	71.9	45.1	23.7	442	2491	246	18.2	957	29.4	19.8	22.4	4440
75	86.4	49.0	28.9	453	2564	256	19.6	1109	32.0	26.4	26.8	4558
76	88.0	49.1	26.1	465	2464	261	19.9	1058	31.5	29.4	26.8	4663
77	90.8	56.6	30.6	520	2456	261	20.6	1252	31.6	26.4	28.6	4412
78	101.0	54.5	29.4	420	2619	267	20.3	1365	31.8	33.1	30.0	4484
79	109.5	62.6	32.1	547	2611	272	19.6	1349	36.9	27.1	28.2	4599
80	91.0	46.3	26.5	404	1645	265	19.8	1016	36.8	22.4	25.3	4413
81	108.9	64.0	30.1	542	2675	276	22.4	1177	35.9	32.4	30.6	4819
82	113.2	59.1	31.5	590	2693	280	20.3	1129	36.0	34.9	33.8	4710
83	81.1	48.7	26.2	508	2399	269	19.9	1044	41.8	29.3	31.7	4598
84	106.7	56.4	28.1	600	2878	279	20.2	1014	40.0	32.1	35.3	4954
85	118.0	66.7	34.1	630	2810	298	20.8	1109	38.1	36.4	35.4	5437

Source: Agricultural Statistics

Table 2 - Correlations and approximate p-values for board yield values of several commodities for the years 1971-1985.

Commodities	Sorghum	Cotton	Soybeans	Potatoes	Peanuts	Sugar	Sunflowers
Corn	.87 (.00)[1]	.57 (.03)	.82 (.00)	.41 (.13)	.31 (.25)	.37 (.17)	.57 (.03)
Sorghum	1.0	.77 (.00)	.81 (.00)	.24 (.39)	.39 (.14)	.32 (.24)	.65 (.01)
Cotton	-	1.0	.55 (.03)	.04 (.90)	.55 (.03)	-.06 (.81)	.32 (.24)
Soybeans	-	-	1.0	.55 (.03)	.36 (.19)	.59 (.02)	.41 (.12)
Potatoes	-	-	-	1.0	.42 (.12)	.48 (.07)	.04 (.89)
Peanuts	-	-	-	-	1.0	.31 (.27)	.04 (.89)
Sugar beets	-	-	-	-	-	1.0	.11 (.69)

[1] Approximate p-values are given in parentheses.

Table 3 - Comparison of correlations of raw data, first-differences, and ranked first-differences of board yield estimates of several commodities, 1971-1985.

Commodities	Sorghum	Cotton	Soybeans	Potatoes	Peanuts	Sugar beets	Sun flowers
Corn	.87 (.83)[1] .75[2]	.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 first-differences are given in parentheses.

[2] Spearman rank-correlations of first-differences are given in bold type.

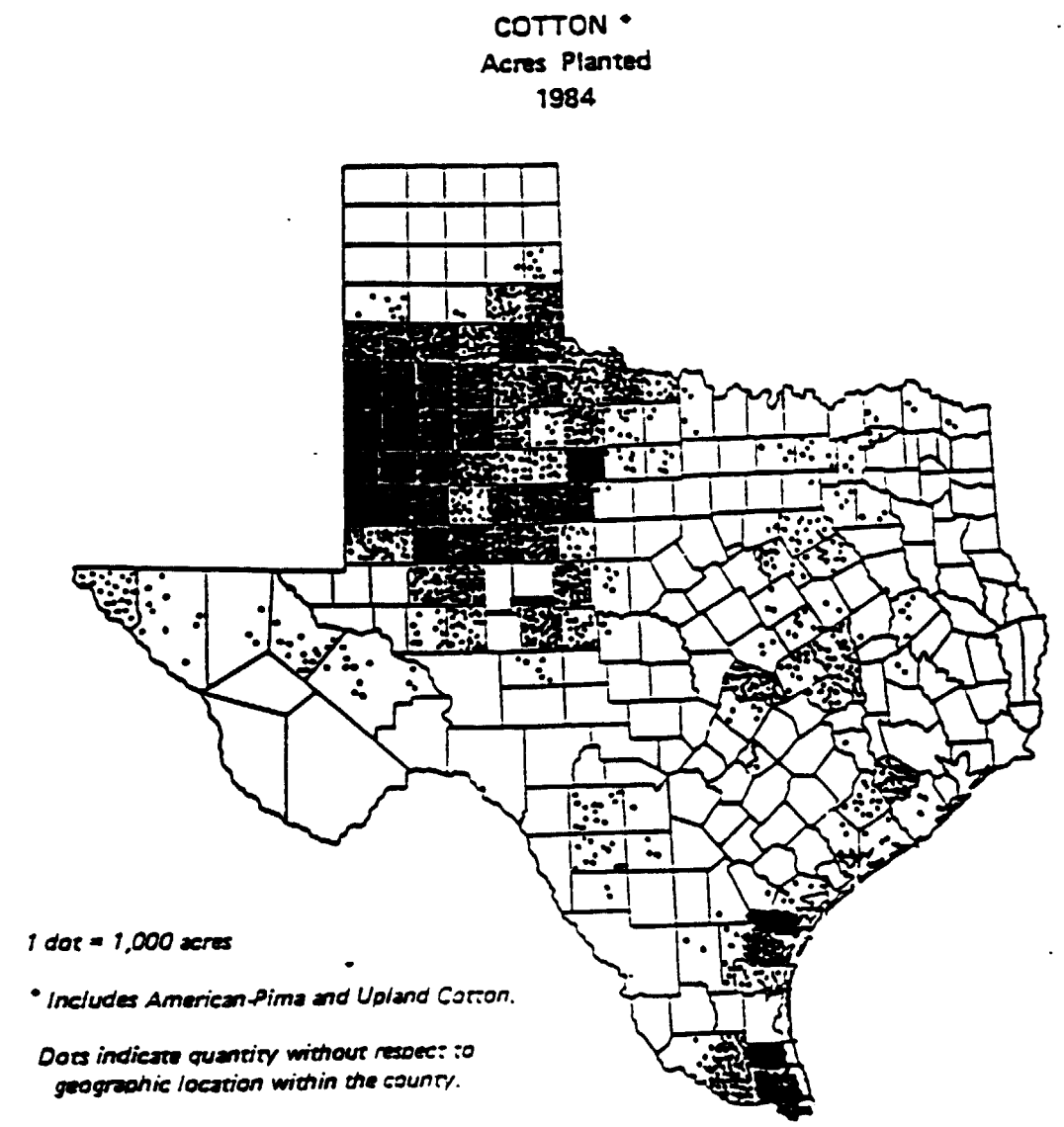
Table 4 - Eigenvectors and eigenvalues of correlation matrix given in table 2.

Commodity	Eigenvectors							
	1	2	3	4	5	6	7	8
Corn	.90	-.11	-.18	.11	-.15	.32	-.04	.05
Sorghum	.93	-.30	-.05	-.07	-.05	.01	.13	-.12
Cotton	.70	-.47	.49	-.05	-.14	.13	.05	.11
Soybeans	.91	.16	-.14	-.03	-.25	.17	-.15	-.04
Potatoes	.51	.69	.03	.51	.03	.07	.07	.02
Peanuts	.56	.24	.70	-.08	.35	-.08	-.06	-.04
Sugar	.51	.66	-.27	-.46	.08	.01	.06	.06
Sunflowers	.58	-.45	-.47	.10	.47	.08	-.03	.03
Eigenvalue	4.17	1.51	1.08	.51	.47	.17	.06	.04

Table 5 - Eigenvectors and eigenvalues of correlation matrix of the raw first-differences.

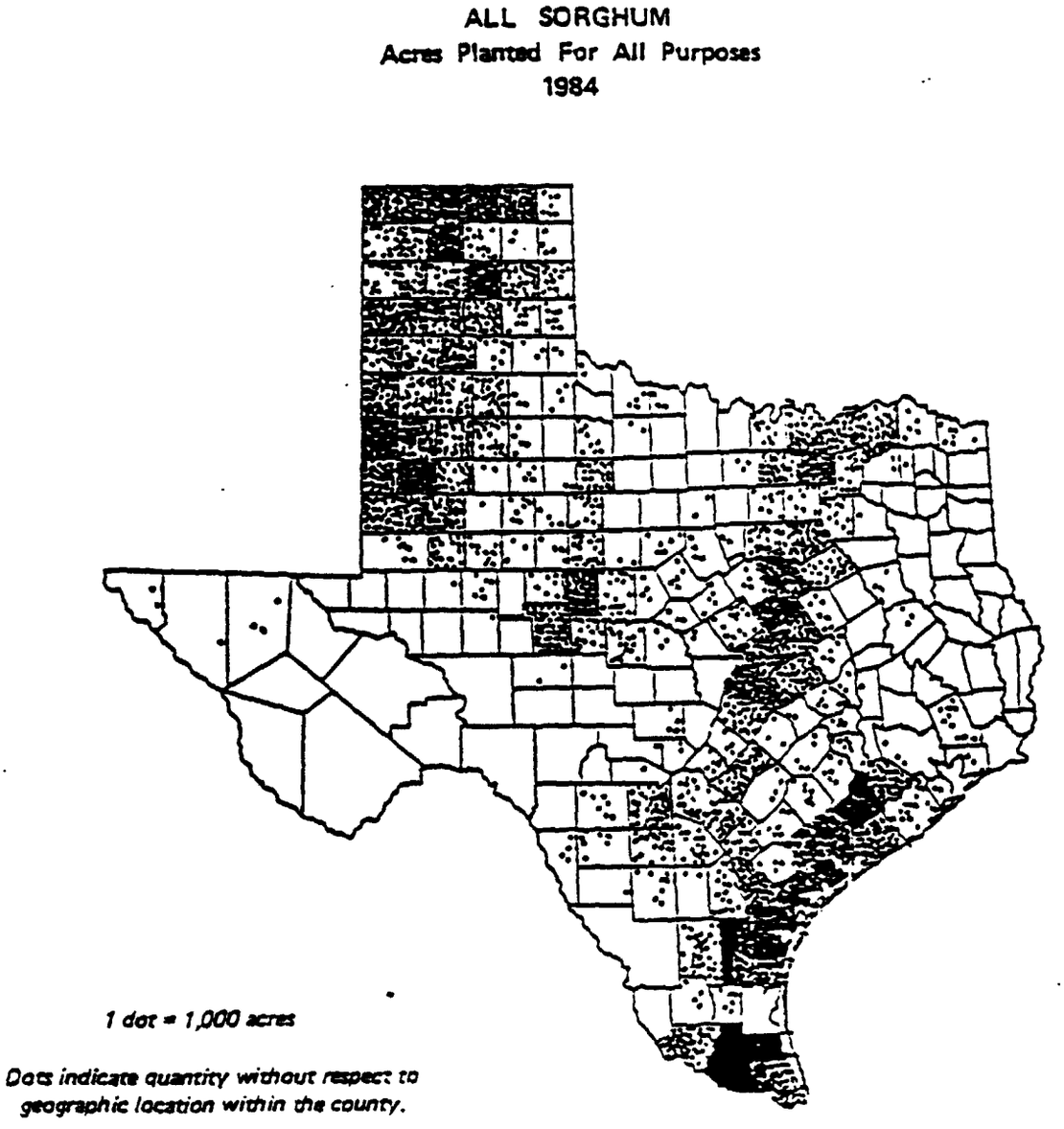
Commodity	Eigenvectors							
	1	2	3	4	5	6	7	8
Corn	.88	-.06	.18	-.07	-.14	-.38	-.01	.02
Sorghum	.96	.19	.02	-.10	-.01	.06	-.15	-.07
Cotton	.84	.03	-.04	-.51	.08	.13	.00	.07
Soybeans	.89	-.02	-.15	.00	-.40	.10	.14	-.04
Potatoes	.61	-.59	.44	.24	-.08	.14	-.04	.03
Peanuts	.79	-.21	-.07	.06	.57	-.04	.08	-.04
Sugar	.56	.67	.38	.27	.08	.06	.04	.03
Sunflowers	.71	.01	-.60	.35	-.03	.00	-.05	.05
Eigenvalue	5.01	.89	.77	.54	.52	.21	.53	.02

Figure 1 - Acres planted for cotton in Texas in 1984



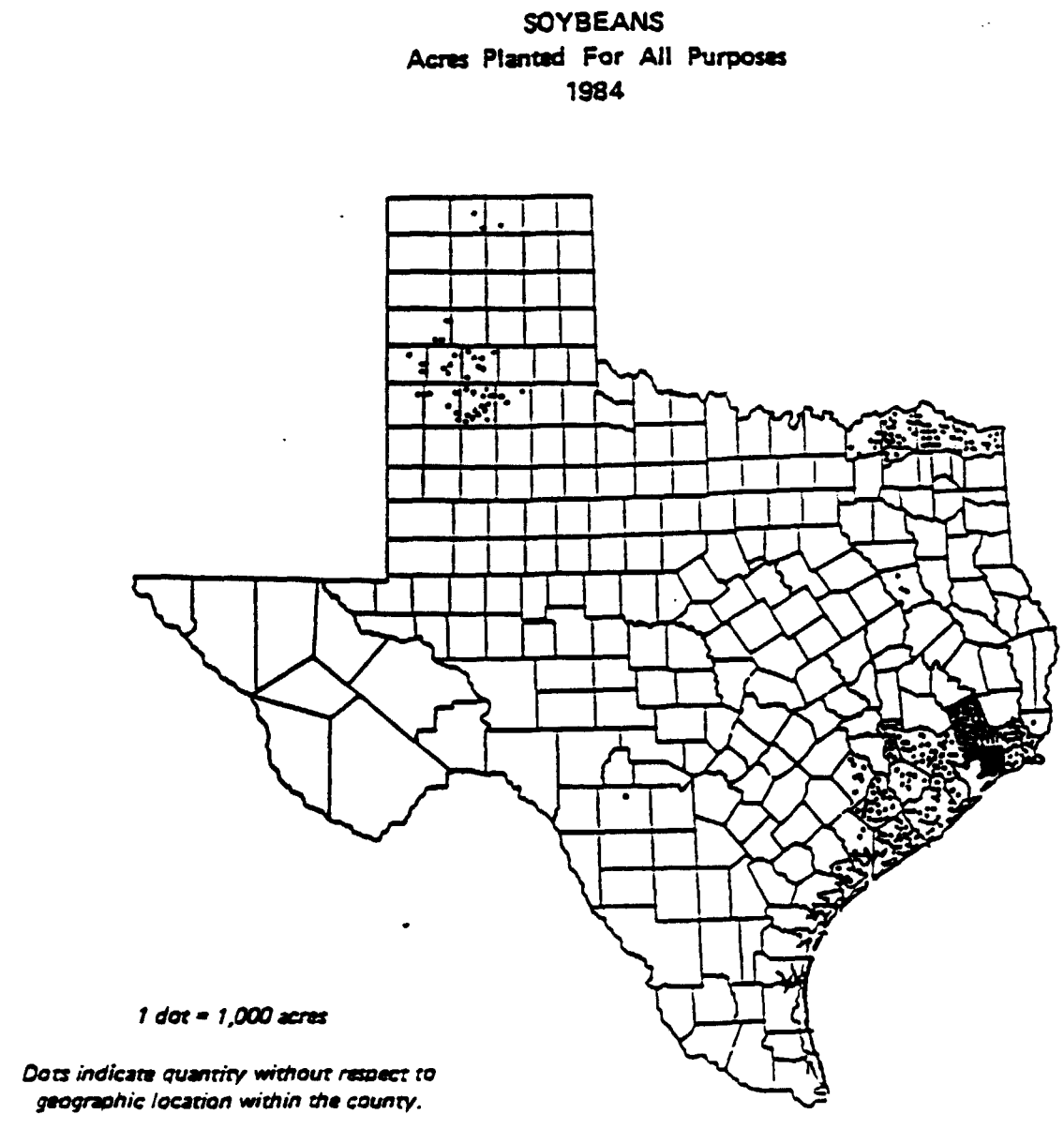
Source: Texas Field Crop Statistics, 1984.

Figure 2 - Area planted for sorghum in Texas in 1984



Source: Texas Field Crop Statistics, 1984.

Figure 3 - Area planted for soybeans in Texas in 1984



Source: Texas Field Crop Statistics, 1984.

APPENDIX A: INTRODUCTION TO PRINCIPAL COMPONENTS

The purpose of this brief introduction to principal components is to provide for the reader some of the background that is necessary to fully understand the analysis of the national board estimates of commodity yields that was given in this report. In this analysis, the principal components of the correlation matrix given in table 2 were examined to lend insight to the relationships that may exist between the yields of these commodities. The use of principal components for this particular purpose of analyzing correlation structures was first proposed by Hotelling in 1933.

In order to introduce some of the more fundamental concepts and to illustrate how principal components may be used, the residuals from the corn and cotton yield data were recorded in table 4 and then plotted in figure 1. For two-dimensional data such as this, there is a very nice analogy between principal components and least squares regression. Let X = the cotton yield residual, and Y = the corn yield residual. Recall that in simple regression, the least-squares line, $Y = a + bX$, is the line that minimizes the sum of the squares of $(Y - Y)^2$, the vertical distances from the data points to the line. The first principal axis may be defined to be the line that minimizes instead the sum of squares of the perpendicular distances from the data points to the line, (that is, the distance is now taken from the point perpendicular to the principal axis). The principal axis and the perpendicular distance from one of the data points to this axis are shown in figure 4. In general, the principal axis always passes through the point (X,Y). In this example, since both X and Y represent residual random variables, the principal axis passes through the origin (0,0).

For two-dimensional data, the second principal axis may then be defined to be the line perpendicular to the first principal axis that passes through the point (X,Y). The second principal axis is the line that maximizes the sum of squares of the perpendicular distances from the data points to the line. These distances, which for our purposes may be defined to be the first principal components of the respective data points, are taken along lines that are parallel to the first principal axis. Thus, in a rough sense, the first principal axis defines the direction in which the variance of the sample points is maximum.

In general, the first principal component is a linear combination of both X and Y. The first principal component, Z_1, for the sample data given in the example is defined by the next equation.

$$Z_1 = .9935X + .1136Y$$

This in itself is not a very meaningful result. The relative size of the coefficients, .9935 vs. .1136 is as much a reflection of the relative size of the units used to measure corn and cotton yields as of anything else. (Note that the range for cotton yields

is 180 vs. 40 for corn). One common remedy for this dilemma dictates that the principal components should instead be calculated using standardized scores where the cotton and corn components of the data points would be divided by their respective sample deviations. (This is akin to calculating correlations vs. covariances). The first principal component for the standardized data would be given by

$$Z_1 = .7071X + .7071Y \ .$$

This last equation has a pleasing interpretation. Since the first principal component determines the direction in which the (standardized) sample points have maximum variance, the fact that the coefficients for X and Y are equal indicates that the yields of corn and cotton tend to vary in the same direction. In other words, the coefficients of the first principal component indicate that large cotton yields are usually accompanied by large corn yields, (and small cotton yields are usually accompanied by small cotton yields).

The second principal component of the standardized data is given by

$$Z_2 = .7071X - .7071Y \ .$$

the variances of Z_1 and Z_2 may be obtained from the correlation matrix of the random vector $(X,Y)^T$. The variance of the first principal component, Z_1, is equal to the largest eigenvalue (see Appendix B) of this correlation matrix; the variance of the second principal component is equal to the other eigenvalue. The relative importance of these two principal components may thus be measured by comparing the sizes of the two eigenvalues. In our example, the eigenvalues are 1.574 and .426. Thus factors that by themselves would cause the correlation between cotton and corn to be negative account for a component of variation (for standardized yields) equal to .426, and factors that by themselves would cause the correlation between cotton and corn to be positive account for a component of variation equal to 1.574. (The correlation between the corn and cotton yields is .5740).

In general, for multivariate data involving k variables, the first principal component is the linear combination of the k variables whose coefficients determine the direction in which the variance of the sample points is maximum. The second principal component is then the linear combination of the k variables whose coefficients determine the direction <u>perpendicular to the first principal axis</u> in which the variance is maximum. Similarly, the third principal component is associated with the direction perpendicular to both the first and second principal axes in which the variance is maximum, etc. The variance attributed to the jth principal component is equal to the jth eigenvalue of the covariance matrix (correlation matrix for standardized data).

More specifically, let $\underline{U}^T = (U_1, U_2, \ldots, U_k)$ be a k-dimensional random vector with mean vector and covariance matrix C. Let c_{ij} symbolize the ij^{th} element of C. The first principal component of random vector $\underline{U}$ may then be defined to be the linear combination

$$Z_1 = p_{11}U_1 + p_{12}U_2 + \ldots + p_{1k}U_k$$

$$= \underline{p}_1^T\underline{U}$$

of the component variables U_1 through U_k whose variance

$$\mathrm{Var}(Z_1) = \sum_{i=1}^{k}\sum_{j=1}^{k} p_{1i}p_{1j}\,c_{ij}$$

$$= \underline{p}_1^T\underline{p}_1$$

is greatest for all coefficient vectors normalized so that

$$\underline{p}_1^T\underline{p}_1 = 1.$$

Similarly, the second principal component of random vector $\underline{U}$ may be defined to be the linear combination

$$Z_2 = p_{21}U_1 + p_{22}U_2 + \ldots + p_{2k}U_k$$

$$= \underline{p}_2^T\underline{U}$$

of the component variables U_1 through U_k whose variance

$$\mathrm{Var}(Z_2) = \sum_{i=1}^{k}\sum_{j=1}^{k} p_{2i}p_{2j}\,c_{ij}$$

$$= \underline{p}_2^T\underline{p}_2$$

is greatest for all coefficient vectors subject to the constraints

$$\underline{p}_2^T\underline{p}_2 = 1 \quad \text{and}$$

$$\underline{p}_1^T\underline{p}_2 = 0.$$

With respect to the problem described in the report, $\underline{U}$ may be defined to be the random vector associated with the commodity estimate residuals. In particular, let $B_i(t)$ represent the random variable associated with the board yield estimate for the i^{th} commodity at time t and $E\{\ \}$ be the expectation operator. The model used to analyze the board yield estimates is symbolized by the following two equalities.

$$B_i(t) = E\{B_i(t)\} + e_i(t), \quad t = 71, 72, \ldots, 85,$$

$$= a + bt + e_i(t) \ ,$$

where $e_i(t)$ represents the residual random variable at time t and the $e_i(t)$ are assumed to be independent. Assume now that t is a realization of a uniformly distributed random variable T. The i^{th} component of vector $\underline{U}$ would then be defined by

$$U_i = e_i(T) \ , \text{ and}$$

$\underline{U}$ is given by

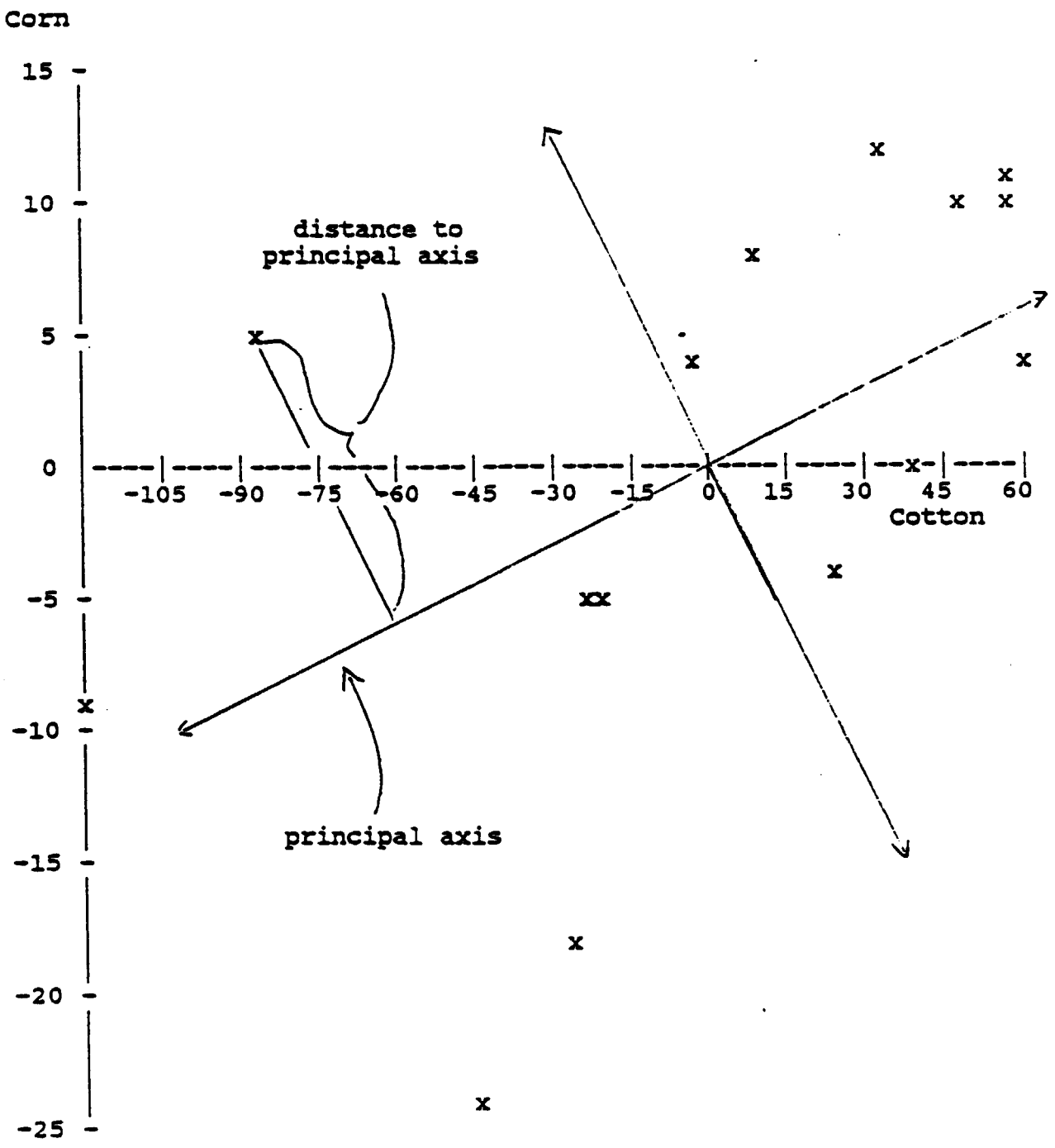
$$\underline{U}^T = (e_1(T), e_2(T), \ldots, e_k(T)) \ .$$

Through the SAS procedure REG, the parameters a and b, and the residuals $e_i(t)$ were estimated. This provided estimates of 15 independent realizations of vector $\underline{U}$. These estimates were then used to calculate a sample covariance matrix and correlation matrix. The latter, of course, is the matrix displayed in table 2.

Table 6 - Predicted values and residuals based upon the board estimates of objective yield for corn and cotton, 1971-1985.

Year	Corn			Cotton		
	Board Estimate	Predicted Value	Residual	Board Estimate	Predicted Value	Residual
	- Bushels per acre -			- Bushels per acre -		
71	88.1	84.35	3.76	438.0	440.6	-2.6
72	97.0	86.04	10.96	507.0	449.9	57.1
73	91.3	87.73	3.57	520.0	459.2	60.8
74	71.9	89.42	-17.52	442.0	468.5	-26.5
75	86.4	91.11	-4.72	453.0	477.8	-24.8
76	88.0	92.81	-4.81	465.0	487.1	-22.1
77	90.8	94.50	-3.70	520.0	496.4	23.6
78	101.0	96.20	4.81	420.0	505.7	-85.7
79	109.5	97.89	11.61	547.0	515.0	32.0
80	91.0	99.58	-8.58	404.0	524.3	-120.3
81	108.9	101.27	7.63	542.0	533.7	8.3
82	113.2	102.96	10.24	590.0	543.0	47.0
83	81.1	104.66	-23.56	508.0	552.3	-44.3
84	106.7	106.35	.35	600.0	561.6	38.4
85	118.0	108.04	9.96	630.0	570.9	59.1

Figure 4 - Plot of corn yield residuals v. cotton yield residuals



APPENDIX B: EIGENVECTORS AND EIGENVALUES

Let C be a square matrix with n rows and columns. An eigenvector of C is any vector $\underline{v}$ such that

$$C\underline{v} = k\underline{v}$$

for some constant k. If C is a covariance matrix then the constant k must be greater than or equal to 0. The constant k is referred to as the eigenvalue associated with the eigenvector $\underline{v}$. There may be as many as n distinct eigenvectors of matrix C. It is common practice to list the eigenvalues (and thus the associated eigenvectors) of a matrix in decreasing order. Thus, if there are n distinct eigenvalues of matrix C, the eigenvalues might be denoted as $k_1, k_2, \ldots, k_n$. The corresponding eigenvectors could then be denoted by $\underline{v}_1, \underline{v}_2, \ldots, \underline{v}_n$ where $\underline{v}_1$ would be the eigenvector associated with the largest eigenvalue k_1, $\underline{v}_2$ would be the eigenvector associated with the second largest eigenvalue, k_2, and so on.