A CLUSTER ANALYSIS APPROACH TO IDENTIFYING SHIFTS IN U.S. COTTON PRODUCTION

G. A. Mumma and D. Hudson
Department of Agricultural Economics
Mississippi State University
Mississippi State, MS

Abstract

Location segments for U.S. cotton production are identified using results from factor analysis and subsequent cluster analysis. Summated scales from the factor analysis are used in hierarchical and non-hierarchical cluster analysis of 16 U.S. cotton producing states. Results for the clusters are reported and are found to compare well with the traditional four cotton producing U.S. regions, the Southeast, Mid-south or Delta, Southwest / or Southern Plains, and West.

Introduction

With the Federal Agricultural Improvement and Reform Act of 1996 (FAIR) in place, the factors which cause shifts in cotton production across states and regions of the US and their effects have become more important to producers, gins, policymakers, and textile manufactures (Martin et al.). Cotton has traditionally been produced in regions generally described as the Southeast, Mid-south or Delta, Southwest and/or the Southern Plains, and West and/or Southern Plains. Supply response studies (Duffy et al., Isengildina) described dramatic regional shifts in cotton production over the past three to four decades. The studies have identified determinants that may be responsible for the acreage shifts, and on a largely arbitrary basis (See Table 1) have clustered the 16-17 cotton producing states into distinct regions. As shown in Table 1, the results from such clustering create an overlap in an accounting of regional shifts in cotton production, which may provide confusing signals to the parties interested in this information.

Cotton was predominantly produced in the Southeast and Mid-south (Delta) regions of the United States between 1960 and 1980. In the early 1980s, cotton production experienced a westward shift to the Southwest and/or Southern Plains and Western and/or Southern Plains regions of the U.S. Since 1986, the shift was back to the Southeast and the Mid-south (Delta) sections of the U.S. Several reasons have been advanced to explain the Westward shift and the subsequent reversal. Studies have looked at the nature of cotton acreage response and identified potential supply inducing factors and variables.

Isengildina and Glade, et al. (1995) posit that the forces influencing the location of cotton production are ultimately reflected in relative returns to basic resources and the cost of inputs. They suggest that soil type, topography, elevation, temperature, sunshine, water availability, irrigation, marketing quotas, program payments, acreage allotments, boll weevil eradication, and the length of production seasons are some among the numerous determinants of where and how well cotton can be produced. Duffy et al. found supply-inducing own-prices, prices of competing enterprises, and government payments for cotton to affect acreage determination. Martin et al. found results that were fundamentally similar to Duffy et al. However, the own-price effects showed conflicting results due to a confluence of other potential acreage determinants.

It is yet to be determined which among the variables are (have been) the primary movers behind the shifts in cotton producing areas of the U.S. A clear-cut and objective method for clustering the cotton producing states based on the identified cotton acreage determinants is yet to be proposed. Comparisons across states, rather than for individual states over time, may be of greater value in identifying production segments and clustering them into distinct regions than the arbitrary methods currently in use. Factor analysis is a generic name for a class of Multivariate statistical methods whose primary purpose is to define the underlying structure in a data matrix (Hair, et al.). Cluster analysis groups objects into subgroups by minimizing the within group differences and maximizing the between group differences (Krause et al.) The objective of this study is twofold. First, the study sought to use factor analysis to identify the underlying structure of the primary factors which may have caused the back and forth shifts in cotton production in the U.S.. The second objective is to use these in a cluster analysis to verify regional (location) segments for cotton production in the U.S.

Methods

Data

A total of 52 variables was (Table 2) initially selected as potentially affecting cotton (Upland cotton) production in each of the 16 cotton producing states in the U.S.. The states were Alabama (AL), Arkansas (AR), Arizona (AR), California (CA), Florida (FL), Georgia (GA), Louisiana (LA) Mississippi (MS), Missouri (MO), New Mexico (NM), North Carolina (NC), Oklahoma (OK), South Carolina (SC), Tennessee (TN), Texas (TX), and Virginia (VA). This study considered ELS cotton as a competing enterprise to Upland cotton in those states in which both were produced. The variables were selected to represent factors that have been previously identified to determine cotton supply response (Duffy et al.; Isengildina; Martin et al.; and Glade et al.,).

For each state, time series data (1979-1996) for farm level cotton price, spot cotton price, mill price, and cotton seed price were selected over their lagged alternatives to represent the supply inducing own-price. Supply inducing

Reprinted from the *Proceedings of the Beltwide Cotton Conference*
Volume 1:240-242 (1999)
National Cotton Council, Memphis TN

prices of competing enterprises were similarly selected. All prices were deflated to 1992 dollars using respective indexes of prices received by farmers (USDA, ERS). In any of the cotton producing states, enterprises in which there was any level of activity during the period under consideration, 1979-1996, competed for resources from Upland cotton. These enterprises were as follows: ELS cotton was produced in AZ, CA, NM, and TX; corn, hay, and wheat was produced in all of the states under consideration; soybeans was produced in all the states except AZ, CA, and NM; sorghum was produced in all the states except AZ and VA; peanuts was produced in AL, FL, GA, NM, NC, OK, SC, TX, AND VA; rice was produced in AR, CA, LA, MS, MO, and TX; barley was produced in AZ, CA, NC, OK, SC, TX, and VA; oats was produced in all states except AZ, FL, LA, MS, NM, and TN; and tobacco was produced in FL, GA, MO, NC, SC, TN, and VA. Both economic and normal net returns above variable costs for cotton and the competing crops were expressed in 1992 dollars for purposes of comparability. The returns above variable costs data were obtained from ERS.

Weather related factors and water availability (based on USDA, ERS and Agricultural census data) were represented by estimated monthly temperature and precipitation, percentage estimated number of irrigated cotton farms, and estimated percentage of irrigated cotton acreage, by state over the 1979-1996 period. The effect of topsoil loss was represented by estimates of sheet and rill erosion of both cultivated and non-cultivated cropland provided by the USDA's National Inventory Directory. Losses in cotton yield that are attributable to the insects were represented by percentage losses due to the boll weevil and other pests and insects. Payments by the government were represented by program payments known as the 1966-85 Diversion Payments, 1980-95 Deficiency Payments, and 1996-98 Production Flexibility Contract Payments that were obtained from the USDA.

Factor Analysis and Cluster Analysis

According to Hair et al., both factor and cluster analyses are interdependence techniques in which all variables are simultaneously considered. Results from a study by Mumma et al. that used R-factor analysis to identify the latent dimensions of the variables that previous studies determined to influence cotton production in the U.S. was adopted for use in this study. By this technique, correlations are computed between variables to provide a resultant factor pattern that demonstrates the underlying relationships of the variables (Hair et al.). The variables considered were of metric measurement, with key indicants identified to be water availability, program payments, competing crop prices, and competing crop acreage. The data for the study was standardized (subtracting the mean and dividing by the standard deviation for each variable) due to the different units and absolute values of the variables. Summated scales were selected over both surrogate variables and factor scores so as to use several variables as indicators, and avoid the use of only a single variable to measure a concept. This process creates a set of variables that posses less random noise to be used in place of the original variable set in the subsequent cluster analysis. Similar to Thomas et al., the Summated scales in this study involved the summation and averaging (around the years, 1979-1984, 1985-1990, and 1991-1996) of the standardized scores of the variables that determine cotton production in the U.S.. Factor loadings greater than 0.60 were considered to be significant (Hair et al.). Summated scales includes variables that load highly on the factor and exclude those that have little impact on the factor; they are easily replicated on subsequent samples, and not necessarily orthogonal (Hair et al.).

Cluster Analysis

According to Hair et al., cluster analysis groups objects based on characteristics they possess. Similar objects are classified together to enable the resulting clusters to exhibit high internal homogeneighty (within-cluster) and high external (between clusters) heterogeneighty. Cluster analysis cannot, by itself, distinguish between the relevance of variables. It is very sensitive to inclusions of irrelevant variables and outliers through non-representative observations, under-sampling, or over-sampling of the group. If the derived clusters are to truly reflect the inherent structure of the data as defined by the variables, selection of variables to be included in the cluster variate must be deliberate and purposeful.

Theoretical, conceptual, and practical considerations embedded in the summated scaling process from factor analysis aided the selection of the variables selected to be included in the cluster analysis. Screening for outliers was conducted and none were found. Wards method, a squared Euclidean distance, method of clustering was used in this study. Similar to factor analysis, data was standardized, thereby avoiding the problem of inconsistencies between the cluster solutions when the scale of the variables is changed.

The data were checked for multicollinearity and representativeness of the sample data. Both hierarchical and non-hierarchical clustering algorithms were used to place similar objects into groups or clusters. Hierarchical clustering algorithms involve agglomerative or divisive structures. This study used agglomerative methods which begin with each object or observation as its own cluster. In subsequent steps, the two closest clusters are combined into a new aggregate cluster, thus reducing the number of clusters by one in each step. In some cases, a third individual joins the first two in a cluster. In others, two groups of individuals formed at an earlier stage may join to form a new cluster.

Non-hierarchical clustering procedures (k-means clustering) are not agglomerative. They assign objects into clusters in steps once the number of clusters to be formed is specified. Three approaches dominate this clustering technique. The sequential threshold sequentially selects cluster seeds and includes all objects within a pre-specified distance. The parallel threshold selects several cluster seeds simultaneously and can be adjusted in the process of clustering to include more or less objects. Optimization is similar to the other two but allows for dynamic reassignment of objects.

Selection of cluster seeds was done in two stages. First, four clusters were pre-specified on the basis of previous studies which identify four U.S. cotton producing regions. Second, both hierarchical and non-hierarchical methods were used to complement each other. The former was used to establish the number of clusters, profile the cluster centers, and identify any obvious outliers. The latter followed to cluster the remaining observations.

Results and Discussion

The results for the non-hierarchical clustering (K-means clustering) method for three periods, 1979-84 (Period I), 1985-90 (Period II), 1991-96 (Period III) are presented in Table 2. The method was used to enable a direct comparison between the resultant clusters of U.S. cotton producing states from this study and the representation of states in the four traditional cotton producing regions in the U.S. Therefore, the algorithm was required to identify exactly four clusters using the nine summated scales that represented the cotton acreage determinants.

Cluster I for the period 1979-84 compared well with the composition of the cotton producing Southeast U.S.. Four of the six traditional Southeast states that produce cotton, AL, FL, GA, and NC, were included in this cluster. However, the group also included OK and NM which have traditionally been classified in the Southwest. The second, third and fourth clusters (Table 4) also compared well with the Delta or Mid-south, Southern Plains / or Southwest, and the West, respectively.

The cluster results formed for Period II (1985-1990) did not compare very well with the traditional cotton producing regions. The first cluster for Period II included most of the states that would be considered to make up the Southeast and the Delta or Mid-south. Similar to the results for Period I, the third cluster included Texas in the Southeast / or Southern Plains. The fourth cluster grouped CA and AR together, despite their largely divergent average annual percentage acreage changes for this period.

Period III (1991-96) compared well with the traditional cotton producing region. Five out of the seven Southeast cotton producing states, AL, FL, GA, SC, and NC were included in the first cluster. Except for VA, the second and third clusters combined, could make up the Delta or Mid-south cotton producing region. Ca, TX, and AZ were grouped together which compares well with the traditional cotton producing Southern Plains / or West.

Table 3 presents the results for the hierarchical clustering technique. The objective for the technique was largely to aid in identifying the seed value to be used in the subsequent and final non-hierarchical clustering. Figure 1, Figure 2, and Figure 3 show the results for this clustering technique. Four clusters were identified for both Period I and Period III (Figure 1, Figure 3). The results for Period I resemble the Southeast and the West. The third and fourth cluster combine to resemble the Delta or Mid-South. Three clusters were identified for Period II (Figure 3). The clusters for this period resemble those for the non-hierarchical technique for the same period.

Table 4 presents the results for the combined use of hierarchical method followed by the non-hierarchical method of clustering. The results are exactly the same as those for the non-hierarchical clustering for Period I and II because the seed value identified by the hierarchical part was four. The three clusters which were identified for period II did not compare the traditional cotton producing states very well.

Conclusions

Summated scales from a previous R-Factor analysis (Mumma et al.) representing some of the primary variables that make up the underlying structure of the factors that may determine cotton production in the different states in the U.S. was used to cluster U.S. cotton producing states into distinct segments. The summated scales represented factors such as water availability, government programs, pest infestation, and a confluence of price and normal and economic returns determining factors.

Results for the subsequent non-hierarchical clustering which required that only four clusters of cotton producing states be generated by the clustering algorithm produced results that compare well with the four arbitrary regions, Southeast, Delta (Mid-south), Southern Plains /or Southwest, and West. Use of hierarchical clustering methods followed by non-hierarchical clustering methods did not bring about any major changes to the results from the non-hierarchical clustering by itself.

The use of factor analysis to examine underlying relationships for the location of cotton production and subsequent use of clustering methods to segment U.S. cotton producing states into their respective regions may provide an objective way to augment the current arbitrary means in place. These methods can provide exploratory tools for further economic analysis of segment related economic problems.

References

Duffy, A. Patricia, J. W. Richardson, and M. K. Wohlgenant. "Regional Cotton Acreage Response." *Southern Journal of Agricultural Economics*, July 1987. (p. 99-109).

Glade H. Edward, Jr., L.A. Meyer, and S. MacDonald. “Cotton: Background for 1995 Farm Legislation.” USDA, ERS. agricultural Economic Report Number 706.

Hair, F. Joseph, Jr., R. E. Anderson, R. L. Tatham, and W. C. Black. *Multivariate Data Analysis*. Prentice Hall, Englewood Cliffs, New Jersey, 1995.

Howell, F. M., J. K. Thomas, Ge Wang, and Don E. Albrecht. “Visualizing Trends in the Structure of U.S. Agriculture, 1982 to 1992.” *Rural Sociology*, 61 (2), 1996, p. 349-374.

Krause, H. Joyce, W. W. Wilson, and F. J. Dooley. "Global Market Segmentation for Value-added Products." *Agribusiness*, Vol. 11, No. 3, 195-206 (1995).

Martin, Steven W., D. Hudson, and G. A. Mumma. "Factors Affecting Regional Cotton Acreage Shifts." *1999 Proceedings, Beltwide Cotton Production Research Conference.* National Cotton Council of America, Orlando, Florida, Jan. 1999.

Mumma, G. A., S. W. Martin, D. Hudson. “Location Determinants for U.S. Cotton Production.” Selected paper to the *Southern Agricultural Economics Association Conference*, Memphis, TN, Feb. 1999.

Statistica. *Statistica for Windows*, Volumes I-V. Statsoft Inc. Tulsa Oklahoma, 1995 Stewart, W. David. "The Application and Misapplication of Factor Analysis in Marketing Research." *Journal of Marketing*, Vol. XVIII (February 1981), 51-62.

Stults, Harolds, E. H. Glade Jr., S. Stanford, L. A. Meyer, J. V. Lawler, R. A. Skinner. “Fibers: Background for 1990 Farm Legislation.” USDA, ERS, AI Bulletin No. 591.

U. S. Department of Agriculture, Economic Research Service (ERS). “Cotton: Background for 1985 Farm Legislation.” AI Bulletin No. 476.

U.S. Department of Agriculture, Economic Research Service. “Cotton and Wool: Situation and Outlook Yearbook.” CWS-1997, Nov., 1997.

U.S. Department of Agriculture, Economic Research Service. Weather in U.S. Agriculture: Monthly Temperature and Precipitation by State and Farm Production Region 1950-1993.” <http://usda.mannlib.cornell.edu/datasets/general/92008/README.DOC> Jan.1996 (24 Nov. 1998)

USDA. “Table 9. “Estimated Average Annual Sheet and Rill Erosion on Non-Federal Rural Land, by State and Year.”, < 1992 National Resource Inventory. May 1994. (9 Dec. 1998).

Williams, R. Michael. “Cotton Insect Losses- 1979-97.” Compiled for the National Cotton Council. Mississippi State University, Mississippi State, Mississippi.

Table 1. Previous Arbitrary Clustering of States into U.S. Cotton Producing Regions

Region Author	South-east	Mid-south	The Delta	South/ Plains	South west	West
Martin et al. (1999)	AL, FL, GA, NC, SC, VA		AR, LA, MS, TN, AR, LA		OK, TX	AZ, CA, NM, NEV
Glade et al. (1995)	AL, FL, GA, NC, SC, VA		AR, LA, MO, MS, TN		KS, OK, TX	AZ, CA, NM
Isengildina (1996)	AL, FL, GA, NC, SC, VA	AR, LA, MO, TN, MS			KS, OK, TX	AZ, CA, NM
Duffy et al. (1987)	AL, FL, GA, NC, SC, VA		AR, LA, MS, MO, TN	NM, OK, TX	AZ, CA	

AL=Alabama, FL= Florida, GA= Georgia, NC =North Carolina, SC = South Carolina, VA= Virginia, AR =Arkansas, LA = Louisiana, MS= Mississippi, MO= Missouri, TN = Tennessee, KS = Kansas, OK = Oklahoma, TX = Texas, AZ = Arizona, CA = California, NM = New Mexico

Table 2. Non-hierarchical Clusters (K = 4) of U.S. Cotton Producing States

Period /	Cluster I		Cluster II		Cluster III		Cluster I	
Cluster	State	AC Δ	State	AC Δ	State	AC Δ	State	AC Δ
Period I	AL	3.20	AR	2.57	TX	-3.90	AZ	-1.09
1979-84	FL	58.65	LA	11.27			CA	-0.04
	GA	5.74	MS	3.48				
	NM	-7.27	MO	14.03				
	NC	20.59	SC	2.46				
	OK	-5.47	TN	9.91				
			VA	0				
Period II	AL	3.48	MO	10.29	TX	5.25	CA	-2.01
1985-90	AZ	4.06					AR	12.12
	FL	13.09						
	GA	10.19						
	LA	5.36						
	MS	4.06						
	NM	4.06						
	NC	22 .29						
	OK	1.44						
	SC	6.17						
	TN	10.09						
	VA	43.33						
Period III	AL	5.65	VA	51.60	AR	0.52	AZ	-2.11
1991-96	FL	17.23	TN	-1.41	LA	2.30	CA	0.69
	GA	28.83			MS	-1.41	TX	-1.41
	SC	8.60			MO	4.34		
	NC	12.86						
	OK	-9.91						
	NM	0.32						

Ac Δ = Average annual acreage percentage change for the state during that period

Table 3. Hierarchical Clusters of U.S. Cotton Producing States

Period /	Cluster I		Cluster II		Cluster III		Cluster I	
Cluster	State	AC Δ	State	AC Δ	State	AC Δ	State	AC Δ
Period I	AL	3.20	VA	0	MO	14.0	TX	-3.90
1979-84	GA	5.74	SC	2.4	TN	9.91	CA	-0.04
	NC	20.59			MS	3.48	AZ	-1.09
	OK	-5.47			LA	11.27		
	NM	-7.27			AR	2.57		
	FL	58.65						
Period II	AL	3.48	TX	5.25	CA	-2.01	-	-
1985-90	FL	13.09	NM	4.06	AR	12.12		
	OK	1.44	AZ	4.06				
	GA	10.19						
	LA	5.36						
	NC	22.29						
	TN	10.09						
	MS	4.06						
	SC	6.17						
	VA	43.33						
	MO	10.29						
Period III	AL	5.65	GA	28.83	VA	51.6	TX	-1.41
1991-96	OK	-9.19	FL	17.23	MS	-1.41	CA	0.69
	NM	0.32			LA	2.30	AZ	-2.11
	SC	8.6			AR	0.52		
	NC	12.86						
	MO	4.34						
	TN	-2.19						

AC Δ = Average annual acreage percentage change for the state during that period

Table 4. Hierarchical and Nonhierarchical Clusters of U.S. Cotton producing States

Period /	Cluster I		Cluster II		Cluster III		Cluster IV	
Cluster	State	AC Δ	State	AC Δ	State	AC Δ	State	AC Δ
Period I	AL	3.20	AR	2.57	TX	-3.90	AZ	-1.09
1979-84	FL	58.65	LA	11.27			CA	-0.04
	GA	5.74	MS	3.48				
	NM	-7.27	MO	14.03				
	NC	20.59	SC	2.46				
	OK	-5.47	TN	9.91				
			VA	0				
Period II	AL	3.48	AZ	4.06	AR	12.12		
1985-90	FL	13.09	GA	10.19	CA	-2.01		
	LA	5.36	NM	4.06				
	MS	4.06	SC	6.17				
	MO	10.29	VA	43.33				
	NC	22.29						
	OK	1.44						
	TN	10.09						
	TX	5.25						
Period III	AL	5.65	VA	51.60	AR	0.52	AZ	-2.11
1991-96	FL	17.23	TN	-2.19	LA	2.30	CA	0.69
	GA	28.83			MS	-1.41	TX	-1.41
	SC	8.60			MO	4.34		
	NC	12.86						
	OK	-9.91						
	NM	0.32						

Ac Δ= Average annual acreage percentage change for the state during that period

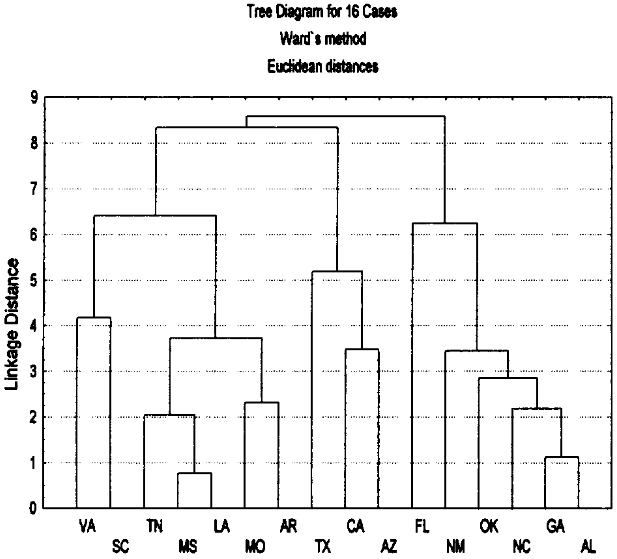

Figure 1. Hierarchical Clusters for Period I (1979-84)

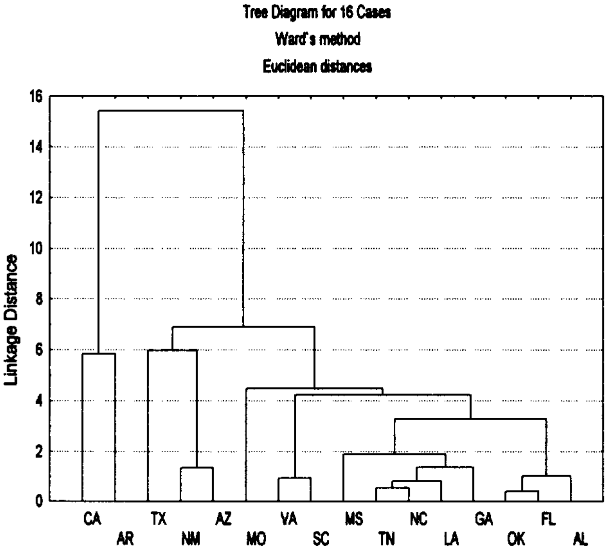

Figure 2. Hierarchical Clusters for Period I (1985-90)

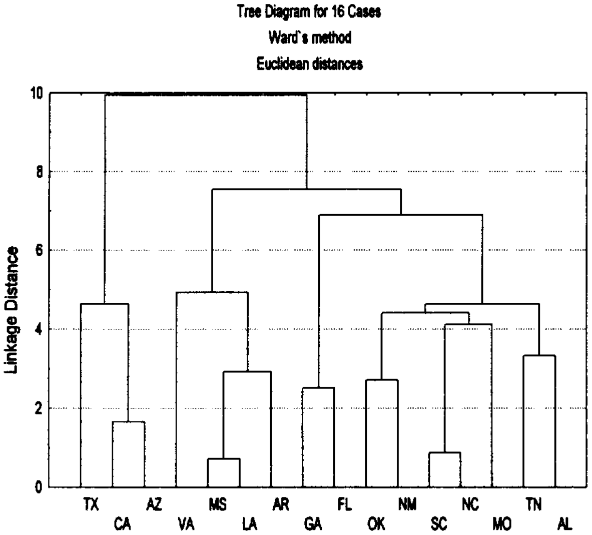

Figure 3. Hierarchical Clusters for Period I (1991-96)