The World's Largest Open Access Agricultural & Applied Economics Digital Library

This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.

Help ensure our sustainability.

Give to AgEcon Search

AgEcon Search
http://ageconsearch.umn.edu
aesearch@umn.edu

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.

Planning, Regional
C

1974

UNIVERSITY OF CALIFORNIA
DAVIS

SEP 9 5 1974

Agricultural Economics Library

Determining Economic Sub-Areas with Principal Component Analysis

by

V. Glenn Chappell, Jr.
Tobacco Tax Council
Richmond, Virginia

Abstract

Principal component analysis is used to incorporate the effects of several socioeconomic variables into an index of regional socioeconomic change. The index is then used as a basis for delineating economic sub-areas within the Tennessee Valley region.

Contributed paper, AAEA meetings, Aug. 1974
College Station, Texas

Determining Economic Sub-Areas With Principal Component Analysis*

V. Glenn Chappell

In recent years the emphasis placed on regional planning and development has stimulated interest in the problem of delineating economic sub-areas within regions. Although several methods have been employed, no universally accepted technique has emerged (2,3,4). In this paper principal component analysis is used to incorporate the effects of a large number of socioeconomic variables into an index of regional socioeconomic change The index is then used as a basis for delineating economic sub-areas within a region.

Principal Component Analysis

Principal component analysis[1] is a statistical technique which may be used to derive a statistically valid, single-valued index of socioeconomic change from several socioeconomic variables. For example, suppose there are P variables X_1, X_2 ... X_P each observed on n individual counties. The jth observation on the ith variable can be denoted Xij and arrayed in a matrix:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{bmatrix} \qquad (1)$$

The single-valued index of socioeconomic change is constructed by deriving a set of principal components from the data matrix through a linear transformation of the variables. The resulting components are defined as linear combinations of variables which have special properties in terms of variances. These special properties come about through the transformation of the original vector variable to the vector of principal components through a rotation of the original response coordinate system to a new coordinate system. This rotation partitions the total variance orthogonally into successively smaller portions. When these portions are distinct, only one set of coefficient vectors evolve. Thus, the method of component analysis rotates the coordinate axes to a new frame of reference in the total variable space. This rotation is an orthogonal transformation in which each set of p original variables is describable in terms of the p new components.

The important characteristic of the new components is that they account, in turn, for all of the variance of the variables. More specifically, the first principal component of the observations (Xij's) in the data matrix in equation (1) is that linear combination of the original variables which contributes a maximum to their total variance and is denoted by

$$C_1 = V_{i1} X_1 + . . . + V_{p1} X_p \quad (2)$$

where C_1 stands for the first principal component, V_{i1} for

for the coefficients of the first component, and X for the corresponding variables in the data matrix.

The second principal component accounts for as much of the residual variance as possible and is denoted by

$$C_2 = V_{i2} X_1 + \ldots + V_{p2} X_p \qquad (3)$$

Additional components are derived until all of the variance has been analyzed. The original variables $X_1, X_2 \ldots X_p$ will be transformed into a new set of variables, $C_1, C_2 \ldots C_p$ which are uncorrelated and have variances in decreasing order.

In the extreme case, the first principal component would explain all the variation in the multivariate system. In the more usual case, the importance and usefulness of the component would be measured by the proportion of the total variance attributed to it. For example, if a significant proportion of the variation in a multivariate system of responses could be accounted for by a single weighted average of the response values, it would appear that the variation could be expressed along a single continuum rather than in multidimensional space. Not only is this more efficient but the relative importance of each original variable in the newly derived component would be indicated by the coefficients of responses. Thus, principal component analysis provides a means for summarizing in fewer variables the variation in a multivariate system.

Application

Principal component analysis was applied to data representing 150 counties in the Tennessee Valley. Primary purpose of the analysis was to derive an index of socioeconomic change which could be subsequently used to delineate economic subareas within the region. Thirteen variables representing various socioeconomic changes during the 1950-1960 period were included in the analysis. These variables were selected primarily because (1) they were readily available from secondary sources, (2) they were available for the same time period and (3) they were available by county for all the counties in the region under study.

Results of the analysis are summarized in Table I. The derived index of socioeconomic change, C_1, is the linear combination given by

$$C_1 = 0.3747X_1 + 0.3843X_2 + 0.2981X_3 + 0.3019X_4 + 0.1567X_5 + 0.2515X_6 + 0.1818X_7 + 0.2066X_8 - 0.2091X_9 + 0.2849X_{10} + 0.2093X_{11} + 0.3578X_{12} + 0.2678X_{13} \quad (4)$$

As the coefficients indicate, the most important variables in the index were change in median income of all families, rate of population change, and change in total payroll per capita.

The index was calculated for socioeconomic changes in each county occuring between 1950 and 1960 in order to compare the relative economic performances of the counties within the region. Counties were then rank-ordered on the basis of

Table I

COMPONENT COEFFICIENTS FOR THE FIRST PRINCIPAL COMPONENT OF THE SELECETED VARIABLES FOR THE 150 COUNTY TENNESSEE VALLEY REGION

Variable Number	Description	Component Coefficient
1	Rate of population change	0.3747
2	Change in median income of all familes	0.3843
3	Change in median income of rural families	0.2981
4	Change in median income of rural farm families	0.3019
5	Change in total property tax	0.1567
6	Change in median value of owner occupied housing units	0.2515
7	Change in the percent of dwellings structurally sound with all plumbing facilities	0.1817
8	Change in the percent employed in agriculture	0.2066
9	Change in retail sales per capita	-0.2091
10	Change in value of farm land and buildings	0.2849
11	Ratio of bank deposits to aggregate income	0.2093
12	Change in total payroll per capita	0.3578
13	Change in the value of manufacturing per capita	0.1568
	Variance	5.1568
	Percentages of total variance	40.0

their index, C_1 and divided into four classifications representing different levels of socioeconomic development. A visual inspection of the county indexes resulted in 13 counties being assigned to the first group representing the most developed counties; 21 counties to the second or moderately developed grouped; 25 counties to the third or slowly developing group and 91 counties to the underdeveloped group. Thus, the use of principal component analysis provided a method for classifying counties into relatively homogenous groups with a maximum amount of difference between the groups.

In order to test the statistical validity of the four groups multiple discriminant analysis[2] was used. Multiple discriminant analysis determines statistically, by an analysis of variance, those linear combinations of county socioeconomic variables which best discriminate among various groups of counties. The sense in which the derived functions discriminate best among these groups is that they represent those linear combinations of variables, each of which (given the preceding variables), maximizes the remaining distance between group means and the variance among groups. The discriminant function obtained in this manner can then be used to evaluate each data point such that the probability of classification into each of the groups is determined. Final classification into a particular group is determined by the selection of the highest probability among each of the groups of every observation.

Twenty-five variables were used to describe the regional economic environment and were utilized in a multiple discriminant analysis program. Each county was assigned four probabilities by the discriminant function developed from the data for the 25 socioeconomic variables (Table II). The largest of the four probabilities determined within which group the individual county was classified. As Table III indicates, fewer misclassifications occurred in the extreme groups, suggesting that the most developed and underdeveloped groups were more easily determined. The most developed group had 12 out of 13 classifications in agreement while the underdeveloped group had 69 out of 91 classifications in agreement. Thus, the results of the discriminant analysis indicated that the four groups were statistically different from each other.

Evaluation

Principal component analysis clearly has potential as a statistical means of assisting researchers in delineating regions into relatively homogeneous sub-areas. The technique's primary advantage is that it gives the researcher the capability to reduce the multiple dimensions of economic development into a single dimension which can be used as an index of economic development. However, while principal component analysis provides a means for determining the total variance and how it is partitioned, it does not provide a measure of a fundamental model for covariance structure.

Although principal component analysis has a strong potential as a technique for systemization and rationalization

Table II

CLASSIFICATION FUNCTION COEFFICIENTS FOR 25 SELECTED SOCIO-ECONOMIC VARIABLES BY GROUP FOR THE 150 COUNTY TENNESSEE VALLEY REGION

Variable Number	Description	Function Coefficient for Group			
		1	2	3	4
1	Percent urban population	-0.5243	-0.5141	-0.5329	-0.4930
2	Percent rural farm population	-2.5774	-2.6155	-2.6353	-2.5809
3	Percent non-white population	-0.1048	-0.0918	-0.0975	-0.0985
4	Percent of population 65 and over	17.8609	19.5337	18.8989	19.6111
5	Percent of population voting in the 1960 presidential election	3.3177	3.4423	3.6246	3.4705
6	Percent of families with incomes under $3,000	4.0330	3.9658	4.1313	4.1042
7	Percent of families with incomes over $10,000	-14.8255	-14.2762	-14.1886	-14.3720
8	Median school years completed	117.7836	115.4467	116.4134	115.0909
9	Percent of migrants from different county	6.9068	6.9110	6.8987	6.9011
10	Percent of population in labor force	41.8459	40.7987	41.1052	40.7381
11	Percent of employment in white collar occupations	4.5392	4.4304	4.4478	4.5351
12	Percent of employment working outside county of residence	-0.6337	-0.6406	-0.6957	-0.6822
13	Average size of household	570.1316	568.4603	571.6893	573.8665
14	Per capita property taxes	-0.5466	-0.5703	-0.5563	-0.5876
15	Number of towns and municipalities	-0.2171	-1.4460	-1.1405	-1.2918
16	Non-worker to worker ratio	417.3268	403.5153	401.5978	395.9311
17	Value of mineral industries shipments and receipts	0.0034	0.0032	0.0034	0.0035
18	Percent of farms with sales under $2,500	3.6070	3.6020	3.5531	3.5950
19	Percent of farms with sales $10,000 and over	-0.6217	-0.5168	-0.8448	-0.8447
20	Average per farm value of farm products sold	0.0121	0.0116	0.0123	0.0121
21	Farm operator level of living index	1.3026	1.3967	1.4237	1.4818
22	Local government general expenditures on natural resources	-0.2534	-0.2413	-0.2418	-0.2352
23	Percent of land in capability Classes I-IV	0.7156	0.6555	0.6218	0.6438
24	Location of TVA steam generator plants (dummy)	14.5173	13.6474	12.7203	15.3088
25	Location of TVA dams (dummy)	27.1768	27.1481	27.3521	27.8332
	Constant	-2778.4655	-2713.7009	-2748.3676	-2733.5870

Table III

CLASSIFICATION MATRIX FOR THE 150 COUNTY TENNESSEE VALLEY REGION

	Function				
Group	1	2	3	4	Total
I	(12)	1	0	0	13
II	1	(10)	4	6	21
III	1	1	(18)	5	25
IV	3	9	10	(69)	91
Grand Total					150

of facts, the method poses some shortcomings. The forms of the components derived are not invarient under changes in the scales of response. Also, no rational criteria exists for deciding when a sufficient proportion of the variances has been accounted for by the principal components. In this study, the first principal component, which became the index of socio-economic change explained only 40 percent of the total variation of the 13 variate system. However, the explanatory power of the technique would probably have been improved by a better specification and selection of the socioeconomic variables used to represent dimensions of socioeconomic development in each county. Future users of principal component analysis will obtain better results by accomplishing more toward refinement in the selection of geographical, economic, political and temporal variables.

Footnotes

* The research reported in this paper was supported by the Department of Agricultural Economics and Rural Sociology of the University of Tennessee. The author acknowledges helpful discussions of this subject with Joe A. Martin.

1. See (5) for a more complete discussion of principal component analysis.
2. See (1) for an explanation of multiple discriminant analysis.

References

(1) Anderson, T. W., Introduction to Multivariate Statistical Analysis, New York, John Wiley, 1958.

(2) Cox, P. Thomas, Bernard Siskin, and Allan Miller, "A More Objective Procedure For Determining Economic Subregions: Cluster Analysis," Southern Journal of Agricultural Economics, December, 1969.

(3) Fox, Karl A. and T. Krishna Kumar, "The Functional Economic Area: Delineation and Implications for Economic Analysis and Policy," Papers and Proceedings of the Regional Science Association, 15: 57-85, 1965.

(4) Hall, W. W. and J. C. Hite, "The Use of Central Place Theory and Gravity-Flow Analysis to Delineate Economic Areas," Southern Journal of Agricultural Economics, December, 1970.

(5) Tintner, Gerhard, Econometrics, New York, John Wiley and Sons, 1952.