THE ARL LIBRARY INDEX AND QUANTITATIVE RELATIONSHIPS IN THE ARL

A Report by
Kendon Stubbs
for the Committee on ARL Statistics
(Kendon Stubbs, Richard Talbot, Anne Woodsworth)

Association of Research Libraries
Washington, D. C.
November, 1980

CONTENTS

I. Introduction

At a meeting in May, 1980, the ARL members adopted new criteria for membership in the Association. For academic libraries the criteria, in large part, are based upon what the membership committee report called the ARL Library Index. The present paper is intended to explain in more detail than was possible in that report how the Index is derived and what its implications are.

It is worthwhile to note in the beginning, however, that the Index is but one answer to a broad question which can be asked about the ARL statistics. The annual compilation of statistics reports on categories of data, or variables, concerning collections, staffing, and expenditures. An obvious question about these data is whether they point to common characteristics of the ARL libraries and, if so, how these characteristics can be measured. This question in turn gives rise to two narrower questions:

1. Among the ARL libraries what is the nature and degree of association among specific variables -- for example, between number of professional staff and volumes held, or between the data on library expenditures, on the one hand, and the data on enrollments, Ph.D. fields, and Ph.D.'s awarded, on the other hand?

2. What is common to the variables taken together? That is, what characterizes the ARL members as a whole?

The Committee on ARL Statistics and its predecessor, the Task Force on ARL Statistics, have been seeking answers to the first of these questions through the statistical methods of correlation and regression analysis; and to the second question, through the method of factor analysis. Since an understanding of the former method can serve as a background to the latter, we turn in the next section to a discussion of regression and correlation.

II. Measuring Relationships Among ARL Variables: Regression and Correlation

A basic question to ask about the ARL statistics is what the nature and degree of association is between two or more categories of data, or variables. Note that this question has two parts: (1) What is the nature or form of association between the variables; and (2) what is the strength or degree of association?

Consider a specific instance -- the relation between number of professional staff and volumes held. It seems almost axiomatic that the smaller ARL libraries, in terms of volumes held, have fewer professionals, and the larger, more professionals. How can we describe this relationship more specifically, however, and how strong is it? In answer to this query, a useful starting point is to prepare a graph, or plot, of the data on professional staff and volumes held. Figure 1 presents a plot of the 1978-1979 data for selected ARL libraries.

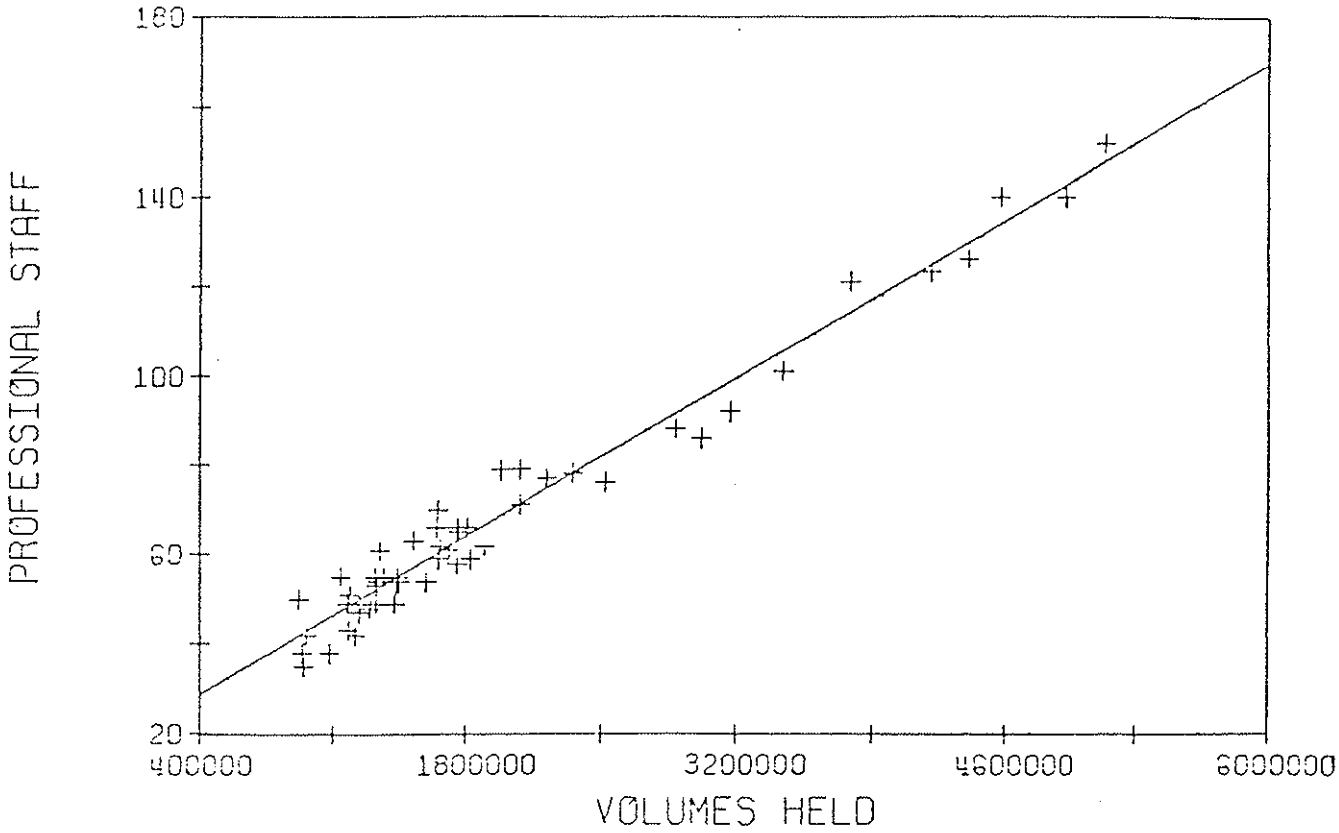


Figure 1: Relationship between professional staff and volumes held, for selected ARL libraries, 1978-1979

Volumes held are measured along the horizontal, or x, axis; professional staff is measured along the vertical, or y, axis. For a given library we measure off the values of volumes held and professional staff on the axes, and the intersection of (imaginary) lines drawn through the two measured points perpendicular to each axis is marked by a cross on the plot. Each cross therefore represents one library's data on volumes held and professional staff.

It should be fairly clear that the crosses in Figure 1 lie more or less in a straight line. A line is in fact shown in the figure, and the crosses are clustered closely around it. The line could have been fitted to the crosses by eye. It can be shown mathematically, however, that a straight line will fit the crosses most closely when the squares of the vertical distances from all the crosses to the line are a minimum. From elementary geometry we also know that the formula for a straight line is $Y = a + bX$. That is, for any value of X, measured along the horizontal axis, a value of Y can be calculated when the values of a and b are known. These mathematical techniques therefore enable us to calculate values of a and b for the line that best fits the crosses in the plot. This line is the regression line.

For the 1978-1979 data on professional staff and volumes held for all ARL libraries (not just the selected libraries shown in Figure 1) the values of a and b in the formula for the line of best fit are approximately

$$a = 20.8$$
$$b = .000025$$

The straight line formula is thus $Y = 20.8 + .000025X$. If the number of volumes in a particular library is substituted for X in the formula, a value of Y can be calculated. The value of Y is a point on the straight line which represents the predicted number of professional staff in that library. The predicted number is usually different from the actual number of professionals because the cross indicating that library's actual number is at some distance from the line. Nevertheless, what is important is that these predicted values for the ARL libraries as a whole are closer to the actual values than any other values that one might derive systematically. In other words, for a given number of volumes in an ARL library, the best prediction that we can make of the number of professionals in that library is through the use of the regression line and its formula $Y = 20.8 + .000025X$.

This formula provides an answer in a specific instance to the first question with which we opened this section: what is the nature or form of association between two variables in the ARL statistics? In spite of considerable differences among ARL libraries, throughout the ARL there was a relation between professionals and volumes held in 1978-1979. From the formula we can see that any increase in the value of X (volumes held) leads to a proportional increase in Y (professional staff). Specifically, substitute 40,000 for X in the formula. Then .000025 times 40,000 equals 1. Thus, the formula would state that Y (predicted professional staff) equals 20.8 plus 1, when X equals 40,000 volumes. Each additional increment of 40,000 adds 1 to the value of Y. Consequently, over and above 20.8 professionals the formula predicts 1 additional professional for each 40,000 volumes held. If a library has 2,000,000 volumes (or 40,000 times 50), the formula predicts that that library has 50 plus 20.8 professional staff, or 71 professionals.

The regression line and its formula are therefore a guide for describing the association between the sizes of ARL libraries in volumes and the numbers of professional staff. This technique permits us to say that, on the whole, in 1978-1979 academic research libraries had 1 professional for each 40,000 volumes held (above a base of 20.8 professionals). Note that this statement is neither prescriptive nor causal. It does not say that research libraries should have 1 professional for each 40,000 volumes -- merely that this was, in general, the ratio in the ARL in 1978-1979. The particular circumstances of given libraries might dictate fewer or more professionals than the formula predicts, since it does not take account of differing effects of automation, resource-sharing, dispersal of library collections on a campus, and other unmeasured influences on staffing.

We have seen that a regression line can be fitted to the data on volumes

and professionals, and a formula for the line can be derived. Returning to the second question with which this section began, we need to assess the strength of the association between volumes and professionals. How reliable is the formula as an indicator of a relationship throughout the ARL? Or in geometrical terms, how close, relatively, are the crosses in Figure 1 to the line? In applied statistics two measures of the strength or degree of association are in common use: the correlation coefficient, designated by r, and the coefficient of determination, designated by r^2.[1] The correlation coefficient measures the degree and kind of association. It can have values from -1 through 0 to +1. Where the association is positive (that is, where the more volumes a library has, the greater its number of professionals), the coefficient is positive (between 0 and +1). For negative association, where high values of one variable are associated with low values of the other, r ranges from 0 to -1. If there is perfect positive association or correlation, r equals 1. For example, if each ARL library had exactly one professional for each 40,000 volumes held, above the base of 20.8 professionals, there would be perfect correlation between volumes and professionals. If we knew how many volumes a library has, we could calculate precisely how many professionals it has. In this situation all of the crosses in Figure 1 would fall exactly on the regression line. In fact, of course, the crosses do not fall on the line, and r is less than 1. If two variables did not have any relationship that could be described by a straight line, r would equal 0. Thus, the closer r is to 1, the more certain we can be that the regression formula accurately describes the association between two variables.

The strength of association is also measured by r^2, the coefficent of determination. This coefficient specifies the proportion of variation in one variable that is associated with the other variable. The values of r^2 range from 0, when none of the variation is explained, to +1, when all variation is associated with the second variable. As an example, the number of professionals in ARL libraries in 1978-1979 varied from 26 to 263. r^2 indicates how much of this variation is associated with the variation in the number of volumes held.

For the relation between professionals and volumes in 1978-1979 r equals .91, and r^2 equals .82. Thus, there is a strong association between professional staff and volumes held in the ARL. 82% of the variation in the number of professionals, moreover, can be explained by the number of volumes. That is, knowing how many volumes the ARL libraries hold, we can account for 82% of the variation in numbers of professionals, merely by using the regression formula $Y = 20.8 + .000025X$. The formula is a powerful tool for making a highly accurate statement about a relationship in the ARL.

It is interesting to turn from the high correlation between professionals and volumes held to a contrasting case in the ARL. There is sometimes thought to be a certain degree of association between the sizes of ARL libraries in volumes held and the numbers of fields in which Ph.D.'s are offered at the parent institutions. Figure 2 shows a plot of volumes held and Ph.D. fields in 1978-1979. A relation between volumes and Ph.D. fields is hard to discern in this plot. Some of the largest libraries are in institutions with only a moderate number of fields, whereas some of the highest numbers of fields are associated with smaller libraries. For volumes and Ph.D. fields r equals .40,

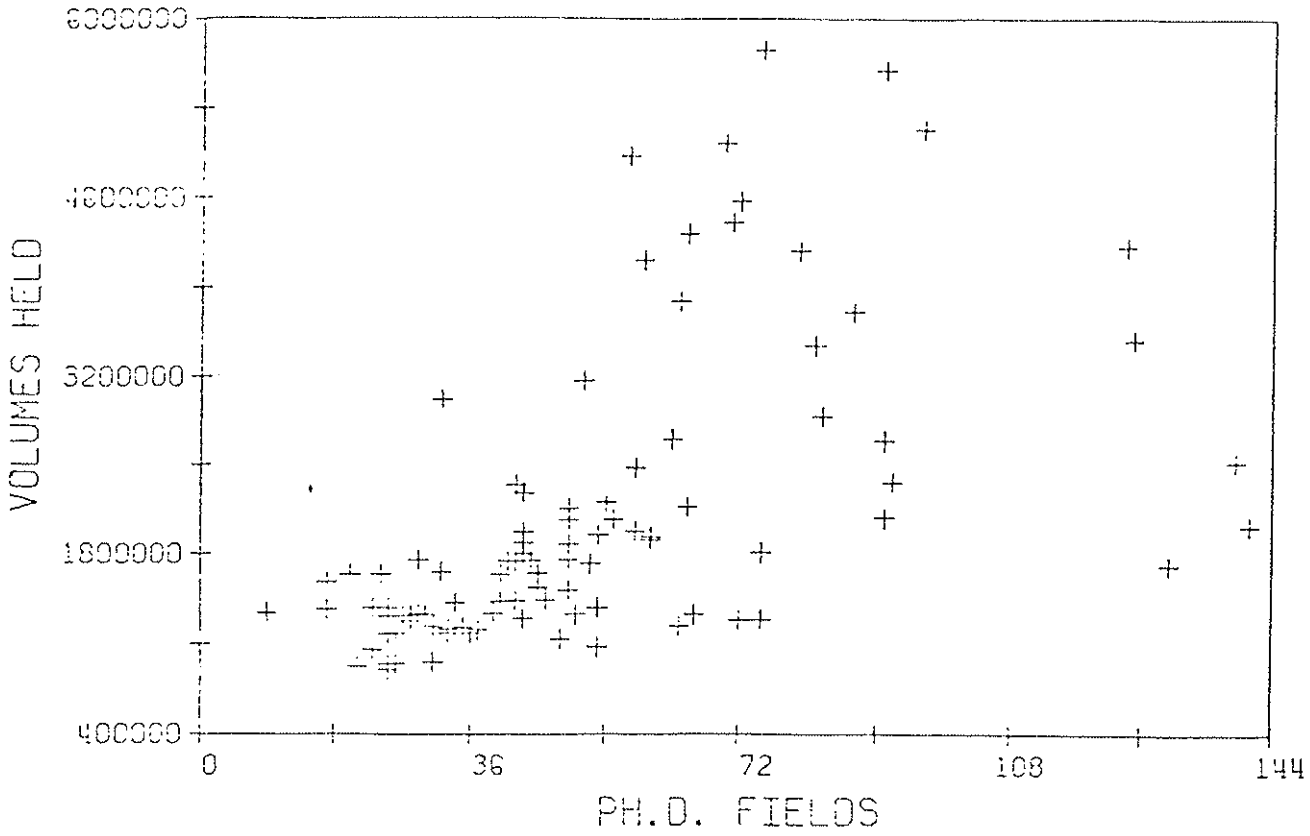


Figure 2: Relationship between volumes held and Ph.D. fields in ARL institutions, 1978-1979

and r^2 equals .16. Only 16% of the variation in the numbers of volumes can be accounted for by Ph.D. fields. Or conversely, Ph.D. fields fail to account for 84% of the variation in volumes. It is therefore fair to say that Ph.D. fields are not a useful guide to the sizes of ARL libraries in volumes.

Just as the correlation coefficient r has been calculated for professionals and volumes, as well as for volumes and Ph.D. fields, so we can compute r for any pair of variables in the ARL Statistics. Table 1 displays the values of r for selected library variables in 1978-1979, and Table 2 displays values of the correlation between library variables and institutional variables. Some of the relationships indicated in these tables are obvious. We would expect, for example, a high correlation (r = .92) between total salaries and professional staff. Other associations are of more import: for example, high correlations between current serials and professional staff; relatively high correlations between library variables such as current serials and professional staff, on the one hand, and graduate students, on the other; considerably lower correlations between these library variables and total students. It is also interesting to compare these correlations for 1978-1979 data with the coefficients from

Table 1: Correlation Coefficients of Selected Library Variables, 1978-1979

	Vols.	Vols. Added Gross	Microf.	Current Serials	Expend. Lib. Mats.	Expend. Binding	Total Salaries	Other Op. Expend.	Prof. Staff	Nonprof. Staff
Vols.	1.00	.84	.35	.89	.76	.75	.85	.82	.91	.82
Vols. Added Gross	.84	1.00	.42	.77	.80	.70	.79	.68	.81	.78
Microf.	.35	.42	1.00	.38	.51	.43	.39	.31	.41	.40
Current Serials	.89	.77	.38	1.00	.73	.74	.83	.79	.81	.74
Expend. Lib. Mats.	.76	.80	.51	.73	1.00	.84	.80	.67	.81	.75
Expend. Binding	.75	.70	.43	.74	.84	1.00	.76	.69	.75	.63
Total Salaries	.85	.79	.39	.83	.80	.76	1.00	.79	.92	.90
Other Op. Expend.	.82	.68	.31	.79	.67	.69	.79	1.00	.78	.69
Prof. Staff	.91	.81	.41	.81	.81	.75	.92	.78	1.00	.86
Nonprof. Staff	.82	.78	.40	.74	.75	.63	.90	.69	.86	1.00

Table 2: Correlation Coefficients of Selected Library and University Variables, 1978-1979

	Vols.	Vols. Added Gross	Microf.	Current Serials	Expend. Lib. Mats.	Expend. Binding	Total Salaries	Other Op. Expend.	Prof. Staff	Nonprof. Staff
Total Students	.31	.46	.38	.37	.44	.29	.47	.22	.45	.40
Graduate Students	.71	.62	.37	.75	.58	.63	.77	.63	.74	.61
Ph.D.'s Awarded	.68	.59	.37	.76	.59	.52	.65	.54	.66	.57
Ph.D. Fields	.40	.40	.34	.41	.38	.32	.46	.33	.44	.38

a decade earlier in Baumol and Marcus.[2] Most of the correlation coefficients are approximately the same, indicating a persistence of the same relationships in academic library data. The chief difference is that the correlation between expenditures and volumes added is significantly lower in 1978-1979 than ten years earlier. In other words, nowadays from a knowledge of volumes added it is harder than in the past to predict how much libraries spent for materials, and vice versa.

We have been considering the association between two variables, such as professional staff and volumes held. In multiple regression it is also possible to assess the relationships between one variable and two or more other variables. For example, for the relation of total library expenditures to volumes r^2 equals .77. That is, volumes held explain 77% of the variation in total library expenditures. If volumes added are included in the regression, however, then volumes held and volumes added explain 80% of total expenditures. If we know how large an ARL library is in volumes held and how many volumes were added, we can predict with a high degree of accuracy how much the library spent in total. Through a further procedure called canonical correlation we can even test the strength of association between two sets of variables. Consider library data on expenditures, on the one hand, and university data on enrollments, Ph.D.'s awarded, and Ph.D. fields, on the other hand. The university data in 1978-1979 explained about 62% of the variation in library data. Thus, if we are given university figures on enrollments, degrees, and fields of study, up to a point (62%) we can predict what kind of expenditures each library may have. But about one-third of the variation in ARL library expenditures cannot be explained by reference to the common measures of university size.[3]

III. Measuring What Is Common to ARL Variables: Factor Analysis and the ARL Library Index

Regression and correlation allow us to analyze the relationships among the categories of data in the ARL statistics - to predict one category from the knowledge of another. It is useful to ask also what is common to the categories taken together. That is, what characterizes the ARL libraries as a whole, or sums up, as it were, what is in the ARL statistics? To these questions the statistical procedures of factor ananlysis provide answers.[4]

Factor analysis is a group of statistical techniques widely used in social sciences such as psychology, education, political science, and sociology. Because of its rather extensive calculations factor analysis is nowadays performed through standard computer programs. The objective of the analysis is data reduction. Data reduction means finding and characterizing underlying patterns in a large set of data, such as the ARL statistics. Factor analysis has both analytical and descriptive uses: analytical, in that it uncovers the underlying dimensions or factors of the data; descriptive, in that it shows through scores for individual libraries how each library is related to the underlying pattern and to other libraries.

In its analytic aspect factor analysis begins with the correlation coefficients for all the variables reported in the ARL statistics -- 25 variables, or categories of data. The analysis asks, so to speak, what groupings of variables there are in the ARL statistics. In a way we can see what is going on in the analytic phase by looking again at Tables 1 and 2. It is obvious that library variables for collections, staff, and expenditures have considerably higher correlations among themselves than they do with variables for enrollments and Ph.D.'s. Similarly, enrollments and Ph.D. variables are more strongly related among themselves than they are with library variables. There are, in other words, two discernible patterns of relationships in Tables 1 and 2 -- a pattern of library relations and a pattern of university relations. Factor analysis measures these patterns through a rigorous, mathematical procedure. The analysis determines that there are four strong patterns or factors in the whole ARL data (as well as weaker factors); and these factors underlie four groupings of the ARL variables:

1	2	3	4
Vols.	ILL orig. borrowed	ILL orig. lent	Total students
Vols. added net	ILL copies borrowed	ILL copies lent	Grad. students
Vols. added gross	ILL total borrowed	ILL total lent	Ph.D.'s awarded
Microforms			Ph.D. fields
Serials			[Student assts.]
Expend. lib. mats.			
Expend. serials			
Expend. binding			
Total salaries			
Operating expend.			
Total expend.			
Prof. staff			
Nonprof. staff			
[Student assts.]			
Total staff			

These results are not surprising. As one would expect, they confirm chiefly that the data on ARL libraries and their universities exhibit two underlying dimensions. It is of more interest to find that interlibrary loans lent represents a grouping separate from interlibrary loans borrowed, and that both of these groups are distinguished from that of factor 1. This factor clearly represents library size and resources deployed, since the variables associated with it are those for collections, expenditures, and staffing. It should be pointed out that the four factors are not totally distinct from one another. There is, for example, a correlation of .63 between factors 1 and 4, that is, between library size and university size. On the other hand, there is a correlation of only .27 between factors 1 and 2, library size and interlibrary borrowing; and this is some evidence that there is not much association between how large a library is and how much it borrows.

Among the fifteen variables grouped under factor 1 some are simply subsets or combinations of others and, so to speak, repeat the information in the others. For example, volumes added net are a subset of volumes added gross, and expenditures for serials of expenditures for library materials; total staff equal professionals plus nonprofessionals. The student assistants variable is peculiar in that it is the only one of the 25 variables that has relations equally to two factors, 1 and 4, and both relations are weak. If we exclude student assistants and the redundant variables like total staff and volumes added net, we are left with ten variables from which a measure of library size can be derived:

Volumes held
Volumes added, gross
Microforms
Current serials
Expenditures for library materials
Expenditures for binding
Total salaries
Other operating expenditures
Professional staff
Nonprofessional staff

It should be re-emphasized that these ten variables, which characterize ARL library size, have been arrived at in an exacting, objective way. We began with all of the categories of data in the ARL statistics and submitted the data to analysis. The analysis determined that fifteen variables are more closely related to each other than to the others. These variables delineate an underlying dimension or factor of library size. From the fifteen we subtract five which are merely subsets or combinations of the others; and we are left with the ten variables above.

From this analysis of all of the ARL variables we turn next to the descriptive uses of factor analysis in characterizing our ten library size variables. For descriptive purposes a simple and direct variant of factor analysis called principal component analysis is used.[5] As before, we begin with the correlation coefficients, but this time for only the ten variables. The analysis determines that there is one significant underlying dimension or component, of which the ten variables are expressions. It then calculates the correlations of the ten variables with this hypothetical component. The component is derived in such a way that it has the highest possible correlations with the ten variables. In regression we asked how strong the relationship is between one or more known variables, such as volumes and volumes added, and some other known variable, such as total expenditures. In factor and principal component analysis we ask what hypothetical variable (or factor or component) is most closely related to the ten library size variables. Again, a look at Table 1 helps us to gain some idea of what the analysis is doing in a rigorous way. It is clear, for example, that variables like total salaries and volumes held have the highest correlations, on the whole, with the other nine variables; whereas the variable microforms has the weakest relations. That

is, total salaries and volumes have strong positive associations with all the variables which, in some way, measure ARL library size. Component analysis determines that total salaries have a correlation of .94 with the hypothetical dimension which we are calling library size; volumes have a correlation of .91; and microforms, .52.

From the correlations between the ten variables and the library size component, what are called component score coefficients are derived. These coefficients are directly proportional to the correlations. The coefficients are in fact no more than weights for each of the ten variables. Here are the component score coefficients for 1978-1979:

Volumes held	.12431
Volumes added, gross	.11905
Microforms	.07064
Current serials	.12297
Expenditures for library materials	.12284
Expenditures for binding	.11364
Total salaries	.12743
Other operating expenditures	.10920
Professional staff	.12674
Nonprofessional staff	.11923

These weights are multiplied by the data from an individual library; and the sum of the ten products is an ARL library size score for that library.

It is worth pausing for a moment to look again at what these weights mean. One of the problems in computing an index or scale from library data has been in deciding what weight to give various categories of data. Are volumes added more important, for example, than current serials in indicating something about library size? Such a question can be debated at length. In the previous ARL membership criteria variables such as volumes held, volumes added, and current serials were treated as if they were of equal weight. In component analysis, on the other hand, weights are derived which are precisely proportional to the correlation of each variable with library size. A high correlation means that there is a strong relationship between a given variable and library size in the ARL. And where the correlation is high it follows that the given variable explains much of the variation in ARL library size. Or, more loosely, what the ARL libraries have most in common is associated with high correlations. Thus, the highest component score coefficient above is for total salaries -- .12743. This implies that ARL libraries are most alike in the strong association between the total salaries each library pays and its other data. The libraries are least alike in the relationships between microforms and the other variables. The coefficients or weights accurately mirror the ways in which ARL libraries are alike or different; they reflect and measure what is common among the ARL libraries.

The weights are multiplied by the data for a given library (after the data are transformed to a standard normal form) to produce a component score. (For the calculation of component scores, see the Appendix.) A library's score

therefore indicates its rank or position in respect to the full range of ARL library size. Taken together, the scores for all of the libraries have the property that they are approximated by a normal curve, the most familiar expression of which is the bell-shaped curve. In this kind of curve or distribution the midpoint (that is, the mean and median) is 0. Most of the values of the distribution fall between +2 and -2, with only a few values higher than +2 or lower than -2. A valuable feature of this kind of distribution is that it permits useful probability statements. For example, in any standard normal distribution approximately 84% of the values are higher than -1, and 95% of the values lie between +2 and -2. We would expect about 84% of the ARL members to have component scores above -1; and about 95% to score above -2 and below +2; and both of these proportions approximately hold true. We can infer, moreover, that if scores are calculated for non-members, and if there is a group of libraries like ARL members, then 84% of this group should score above -1. Or in different terms the chances are only about 16% (100% minus 84%) that a non-member essentially like the ARL members would fail to score higher than -1. The new ARL membership criteria adopted in May require applicants to the ARL to score higher than -1. If a non-member is in fact like the ARL members in library size and resources deployed, this requirement posits an 84% probability that the non-member will pass the test. Similarly, membership can be withdrawn from an ARL library if it continues to score below -1.75. In a standard normal distribution about 96% of the values will be above -1.75. Consequently, if a library shares the essential library size characteristics of the ARL members, the chances are 96% that it will score above -1.75 and only 4% that it may score below -1.75.

The ARL component scores for 1978-1979 are displayed in Table 3. These scores form a scale or index of library size and resources deployed; hence they have been called the ARL Library Index.

How should the scores be interpreted? They are merely a sum of the data from each library on its collections, expenditures, and staffing, weighted in accord with the ways in which ARL libraries are similar or different. They are, in other words, simply mathematical transformations of the data for each library. Except for the weights derived from the whole ARL, there is no more in a library's score than that library's data. The scores indicate library size, but this is an amalgam, so to speak, of volumes held, volumes added, microforms, serials, and so on. These ten variables, from which the scores are produced, may be called inputs. They measure the size of what goes into the library system. What they do not measure, and what the component scores, as a result, do not measure, is the outputs of the system -- the services provided and the use of the system by patrons. We must be careful not to assume that a high score indicates a high quality of the collections and services offered by a library, or a low score the opposite. Whether the relations are strong or weak between library size and quality of use is a question that still needs to be assessed.[6]

Table 3: ARL Library Index, 1978-1979

(Principal component scores based on Volumes Held; Volumes Added, Gross; Microforms; Current Serials; Expenditures for Library Materials; Expenditures for Binding; Total Salaries; Other Operating Expenditures; Professional Staff; Nonprofessional Staff)

1.	Harvard	3.05	50.	South Carolina	-.32
2.	Calif., Berkeley	2.18	51.	Connecticut	-.33
3.	Yale	2.12	52.	Syracuse	-.34
4.	Indiana	1.97	53.	Missouri	-.35
5.	Calif., Los Angeles	1.92	54.	Johns Hopkins	-.35
6.	Toronto	1.91	55.	Tennessee	-.36
7.	Illinois	1.88	56.	MIT	-.39
8.	Stanford	1.80	57.	Western Ontario	-.39
9.	Washington	1.70	58.	Washington U-St. Louis	-.40
10.	Texas	1.62	59.	Utah	-.40
11.	Michigan	1.62	60.	Wayne State	-.41
12.	Columbia	1.54	61.	Nebraska	-.51
13.	Cornell	1.47	62.	Arizona State	-.51
14.	Wisconsin	1.40	63.	Temple	-.52
15.	Minnesota	1.03	64.	Louisiana State	-.52
16.	British Columbia	.96	65.	Texas A&M	-.53
17.	Chicago	.90	66.	York	-.56
18.	North Carolina	.87	67.	Purdue	-.56
19.	Rutgers	.83	68.	Cincinnati	-.56
20.	Florida	.76	69.	Iowa State	-.56
21.	Virginia	.72	70.	Boston	-.58
22.	Princeton	.72	71.	Joint University	-.60
23.	Pennsylvania State	.66	72.	Brigham Young	-.65
24.	Northwestern	.63	73.	SUNY-Stony Brook	-.67
25.	Ohio State	.59	74.	Emory	-.67
26.	Pennsylvania	.54	75.	Colorado	-.71
27.	Calif., Davis	.51	76.	Massachusetts	-.72
28.	New York.	.46	77.	Rochester	-.72
29.	Alberta	.40	78.	Georgetown	-.72
30.	Southern California	.30	79.	Miami	-.73
31.	Pittsburgh	.29	80.	Howard	-.82
32.	Georgia	.29	81.	Brown	-.89
33.	Michigan State	.27	82.	Oklahoma	-.90
34.	Duke	.26	83.	Queen's	-.91
35.	SUNY-Buffalo	.21	84.	Oregon	-.91
36.	Iowa	.19	85.	New Mexico	-.97
37.	Arizona	.17	86.	Calif., Riverside	-.99
38.	Houston	.14	87.	SUNY-Albany	-1.05
39.	Kansas	.11	88.	McMaster	-1.05
40.	Maryland	.08	89.	Dartmouth	-1.13
41.	McGill	.03	90.	Colorado State	-1.14
42.	Calif., San Diego	.02	91.	Tulane	-1.21
43.	Southern Illinois	-.03	92.	Case Western Reserve	-1.22
44.	Kentucky	-.03	93.	Guelph	-1.24
45.	Hawaii	-.11	94.	Notre Dame	-1.28
46.	VPI & SU	-.12	95.	Alabama	-1.39
47.	Calif., Santa Barbara	-.17	96.	Kent State	-1.60
48.	Florida State	-.20	97.	Oklahoma State	-1.87
49.	Washington State	-.31	98.	Rice	-1.91

IV. ARL Component Scores, 1969-1970 through 1978-1979

The ARL Library Index, or component scores, for 1978-1979 is listed in Table 3. It is possible similarly to compute scores for each year of ARL data. These scores can then be plotted on a graph to show year-by-year increases or decreases in scores.

The following pages display plots of the component scores for the last ten years for the 75 libraries which have been ARL members throughout the decade. Libraries which joined the ARL during the seventies were not included in the calculations because the scores would have been somewhat inflated or deflated with the introduction of new members (just as, to a greater extent, the ARL medians have decreased relatively from year to year as smaller libraries have joined). In the plots years are shown on the horizontal axis, where "1969" means 1968-1969, and similarly for the other years. The component scores are indicated on the vertical axis. Note that each library requires a different scale of scores; but the range is the same for all but seven of the plots, since the high and low points differ by 1.0.[7] (That is, the scale for Alabama is from -1.4 to -2.4; for Alberta from 0 to 1.0; and so on.) This sameness of range indicates that the scores for a given library have not increased or decreased by more than 1.0 during the decade.

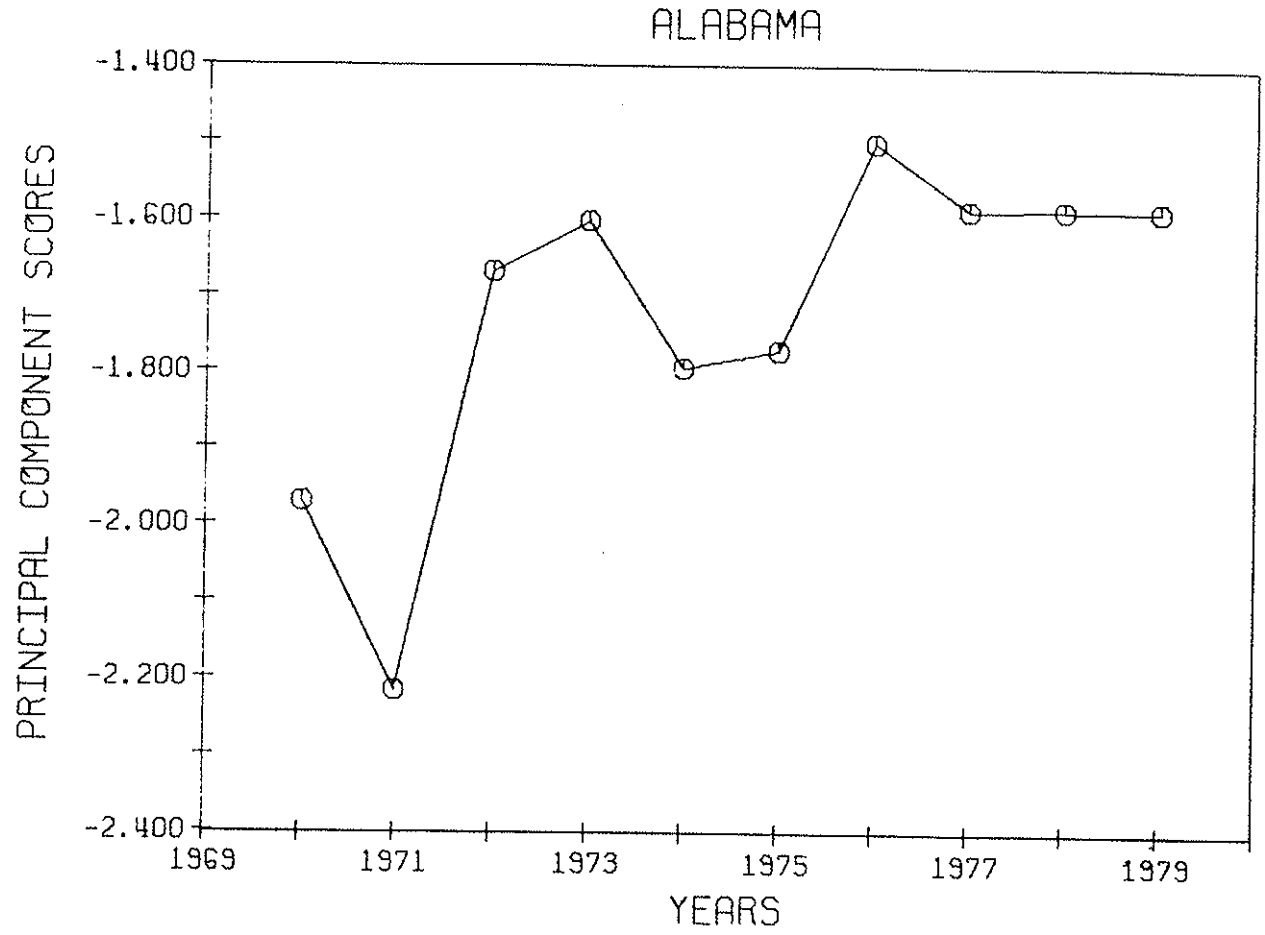
The scores indicate a library's position in the ARL relative to the other 74 ARL libraries in the sample. Positive scores are above the ARL median, and negative scores, below. We can roughly describe the percentiles into which scores fall, as follows:

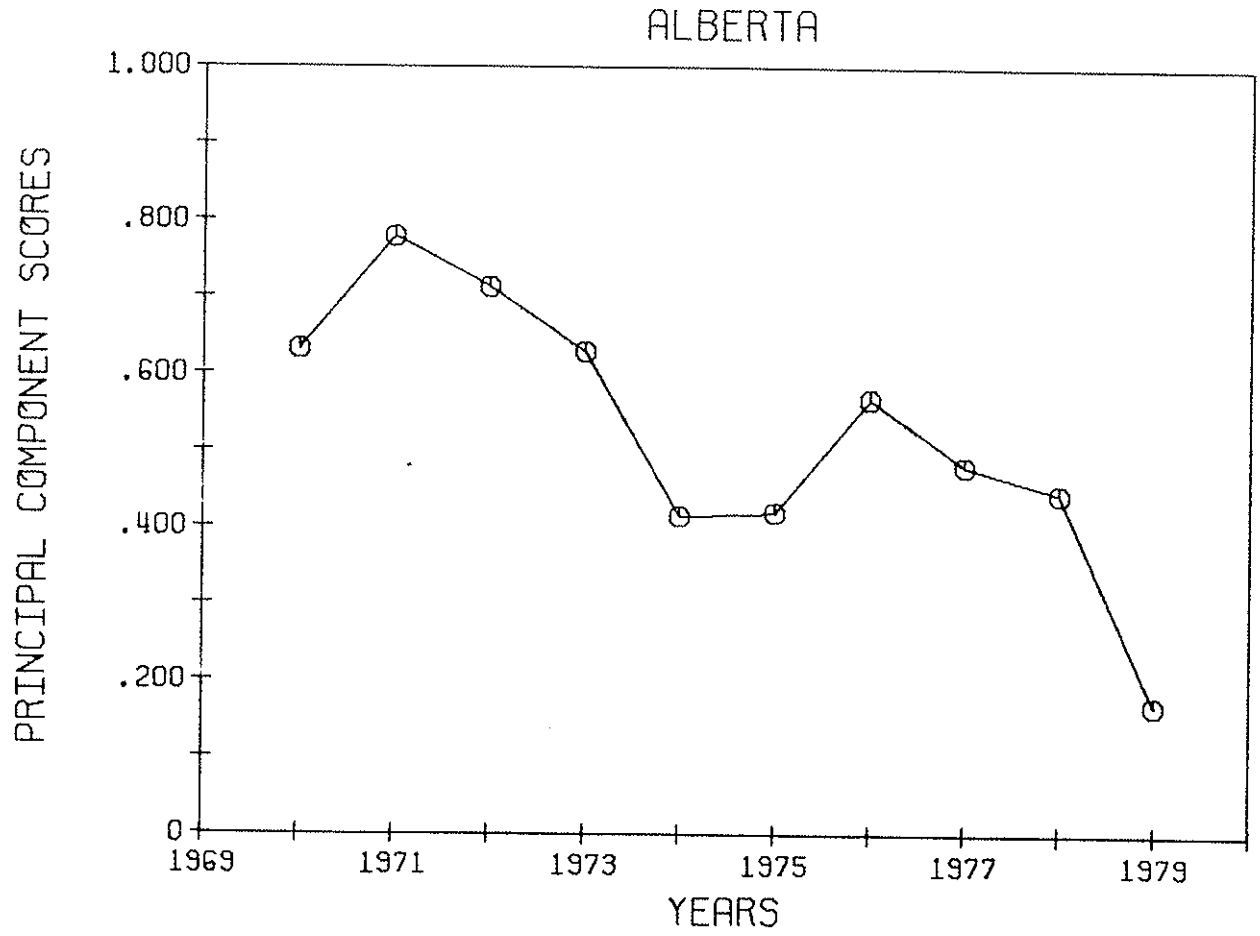
Component Score	Percentile
above 3	upper .1%
above 2	upper 2%
above 1	upper 16%
above .5	upper 31%
above 0	upper 50%
below 0	lower 50%
below -.5	lower 31%
below -1	lower 16%
below -2	lower 2%
below -3	lower .1%

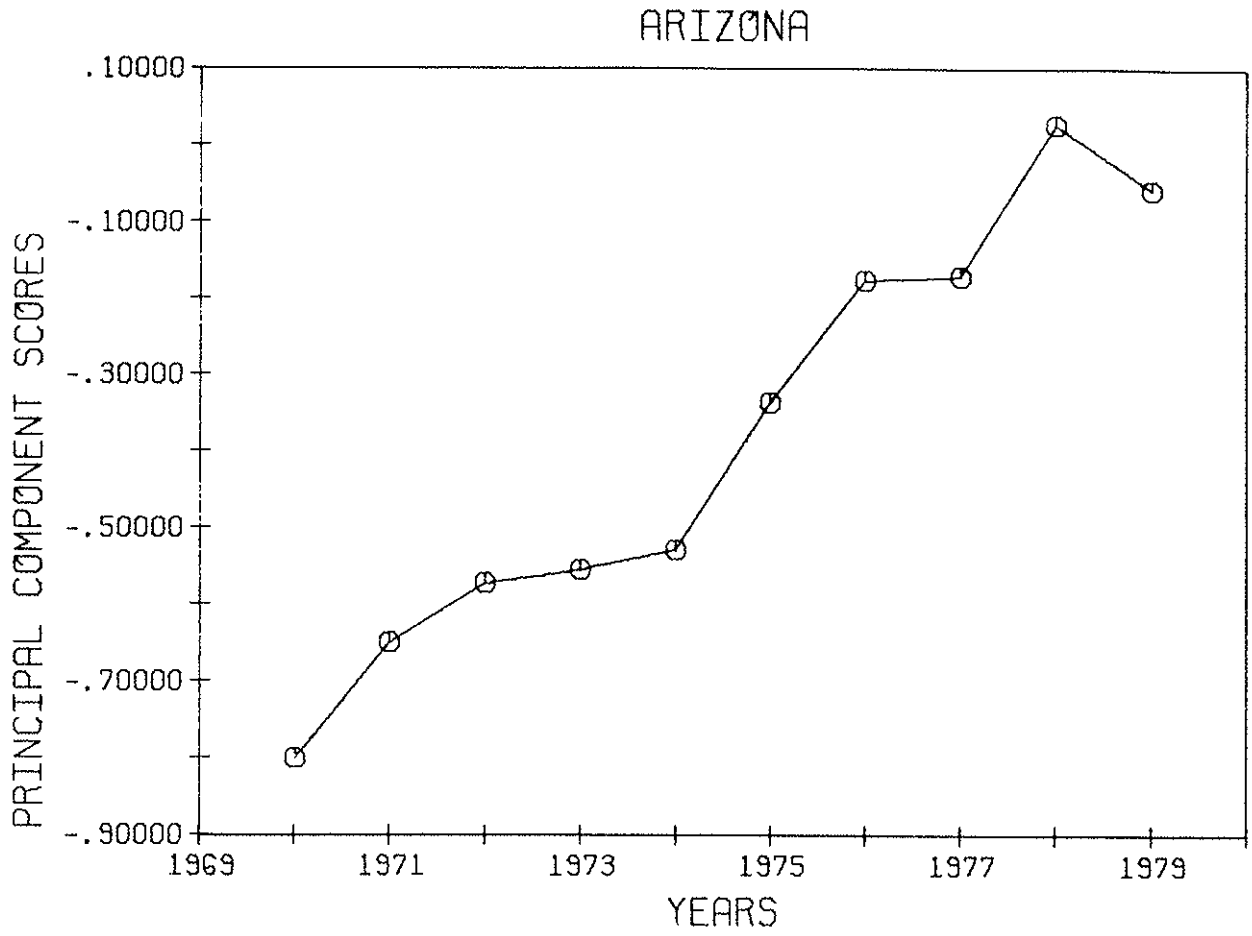
Thus, a score of .7 is in the upper 31% but below the upper 16%. In other words, approximately 15% of the ARL libraries in any year will have scores in the .5 to +1 range.

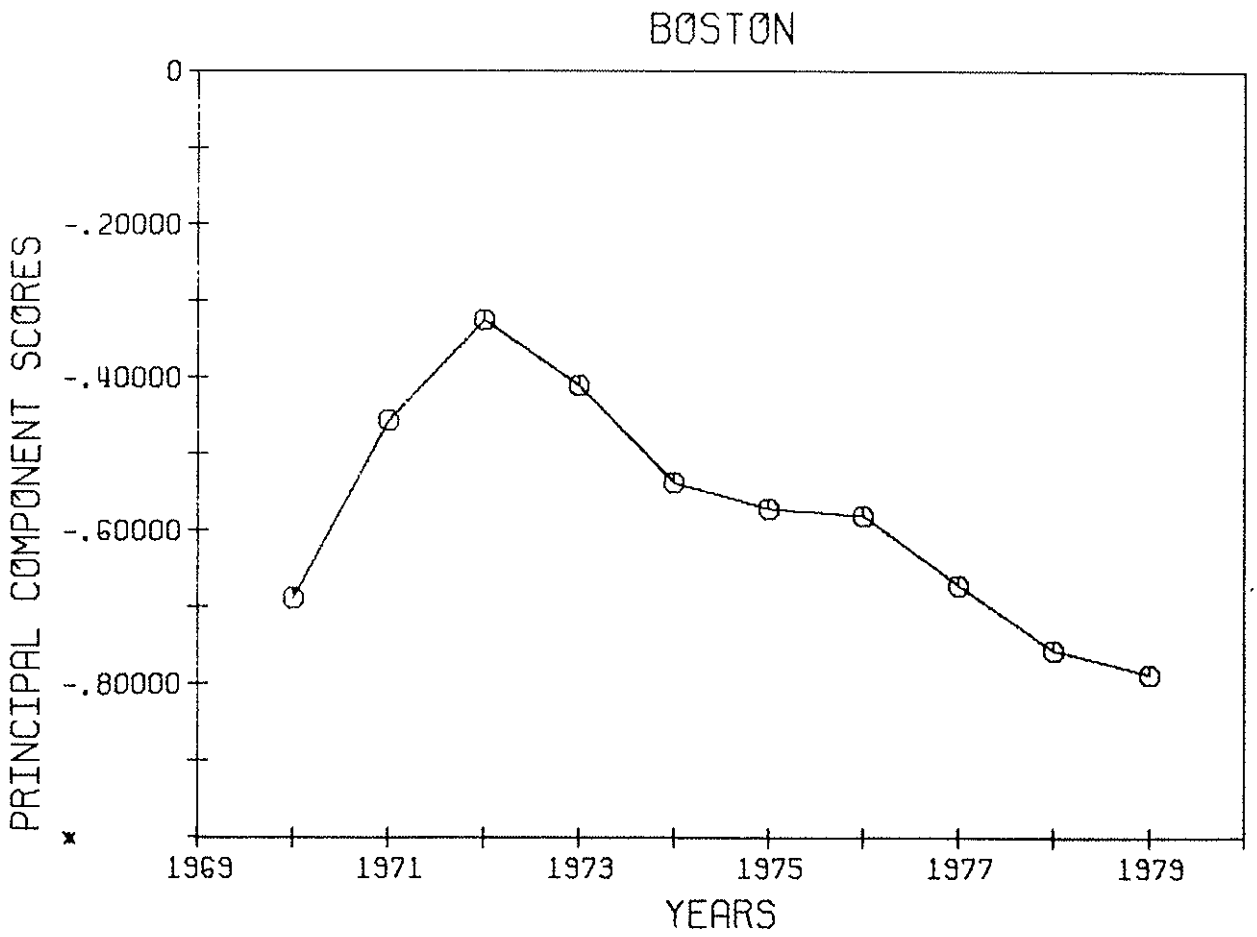
It should be reiterated that the scores measure library size in collections, expenditures, and staffing relative to the other ARL libraries. An increase or decrease in a score from one year to the next indicates only a change in relation to the other libraries. It does not necessarily indicate an absolute rise or fall in a library's collections, expenditures, or staffing. For example, if one

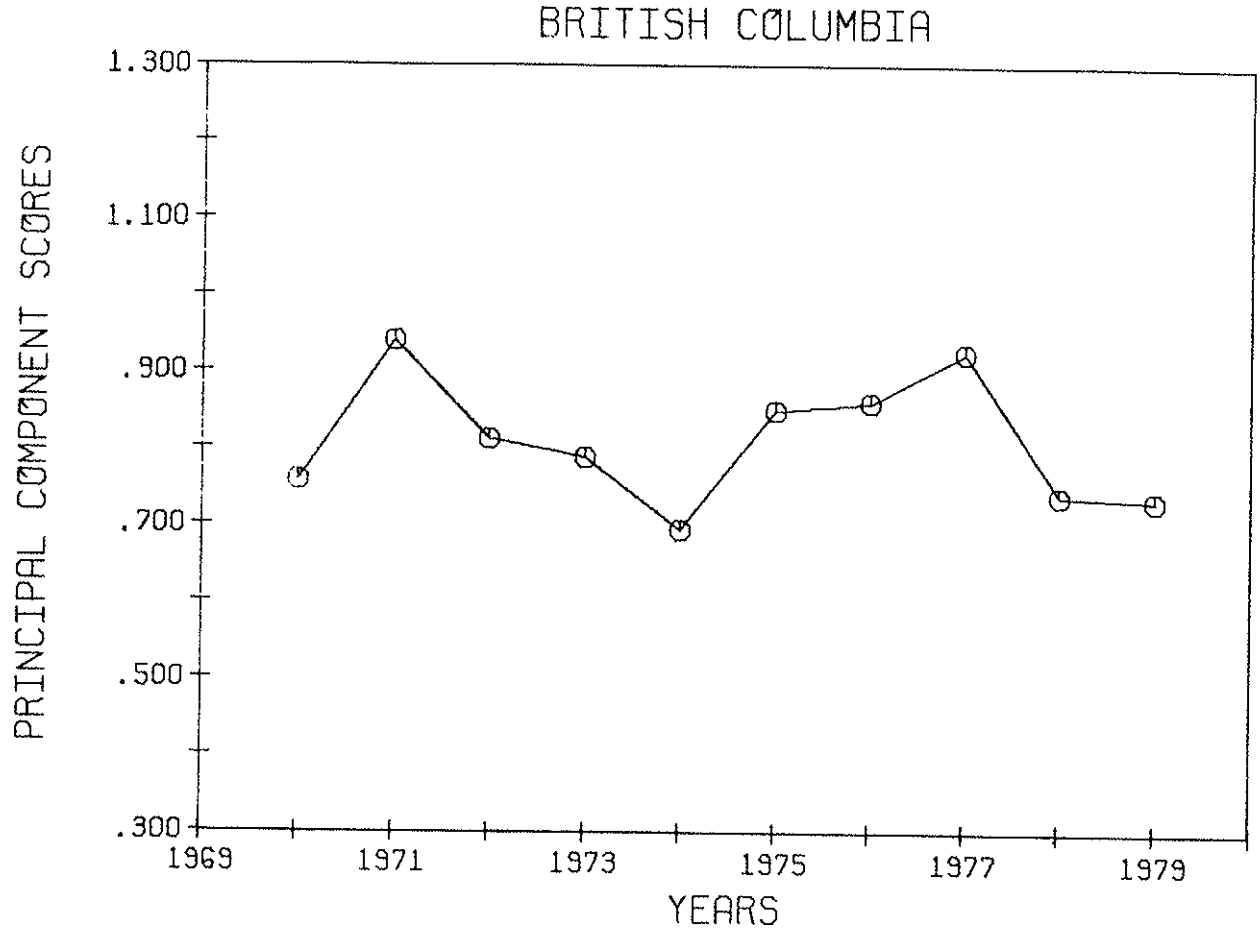
library has increases of 2-3% in various categories of expenditure, whereas most of the other libraries show increases of 6-7%, then the score for that library is likely to be lowered. The scores are merely a way of displaying a member's relationship in library size to its peers in the ARL.

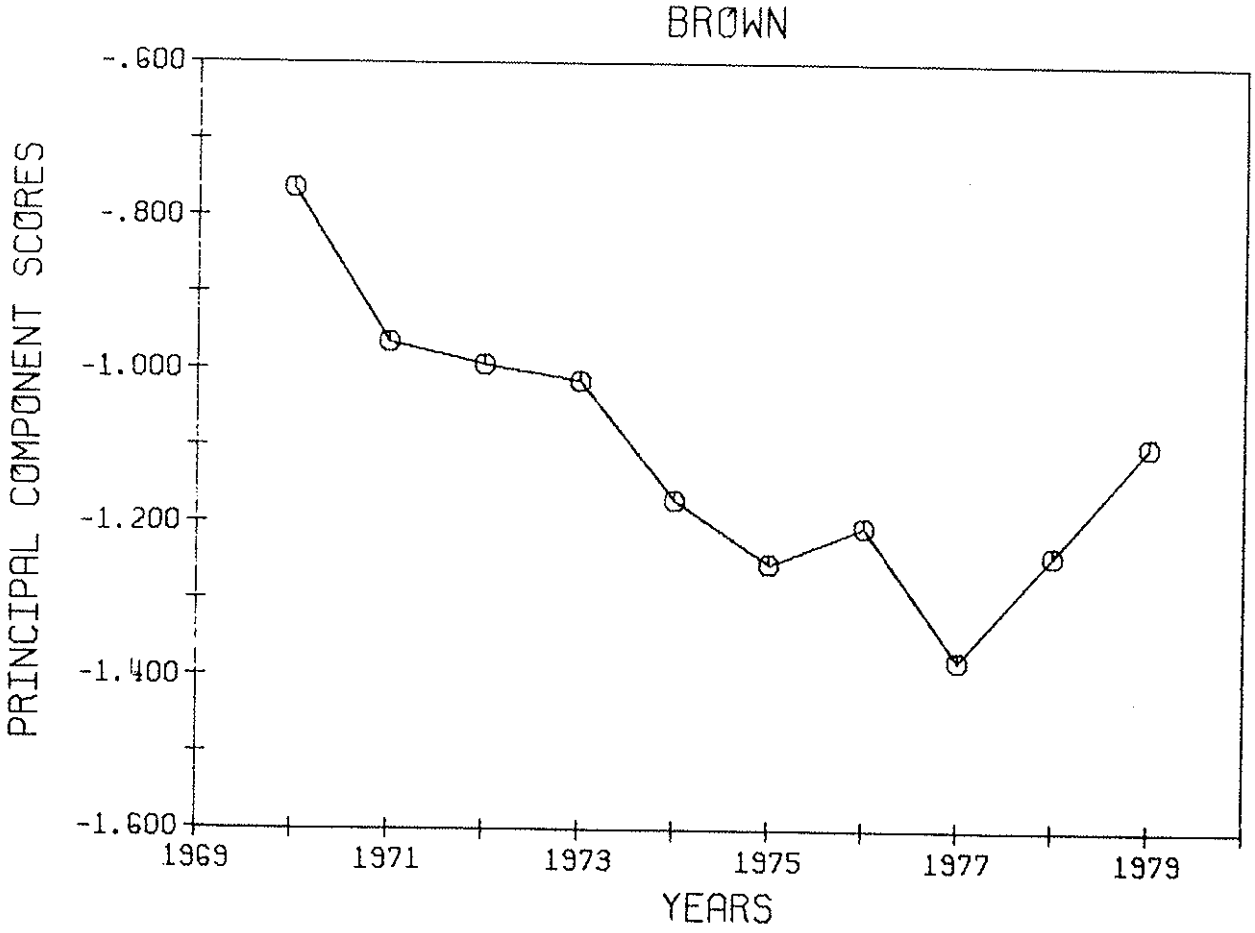

ALABAMA
PRINCIPAL COMPONENT SCORES
-1.400
-1.600
-1.800
-2.000
-2.200
-2.400
1969
1971
1973
1975
1977
1979
YEARS

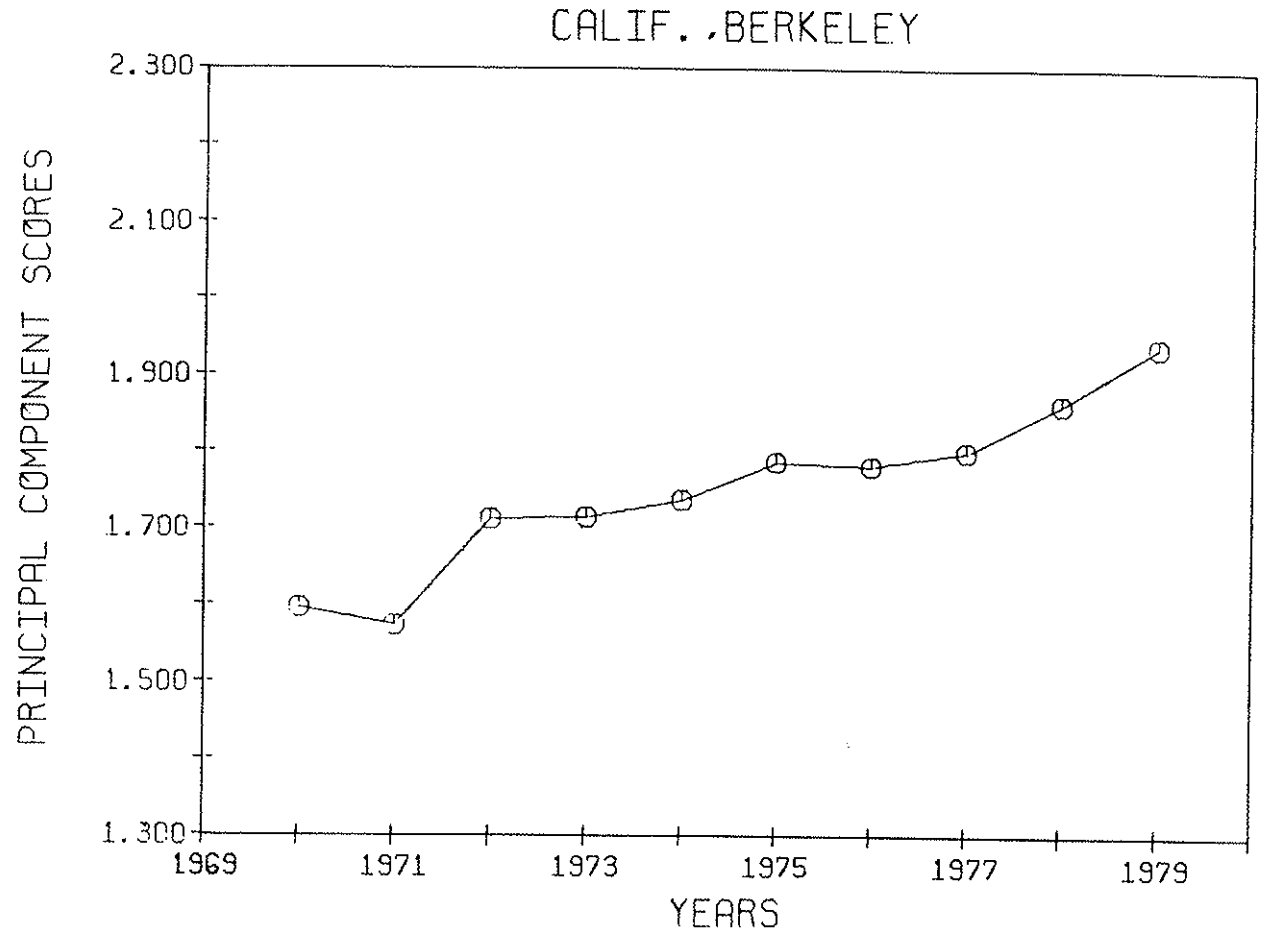

ALBERTA
PRINCIPAL COMPONENT SCORES
1.000
.800
.600
.400
.200
0
1969
1971
1973
1975
1977
1979
YEARS

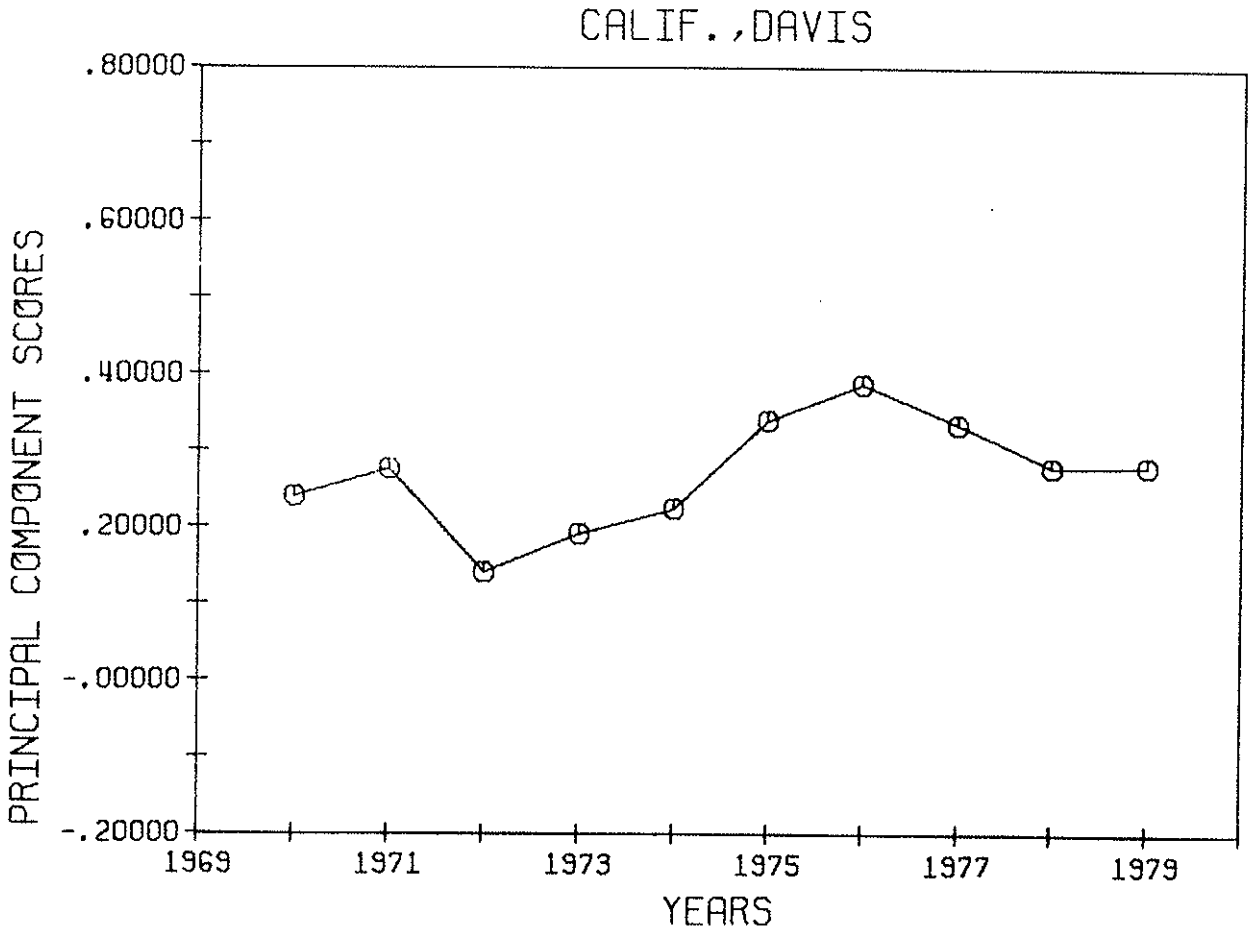

ARIZONA
.10000
-.10000
-.30000
-.50000
-.70000
-.90000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

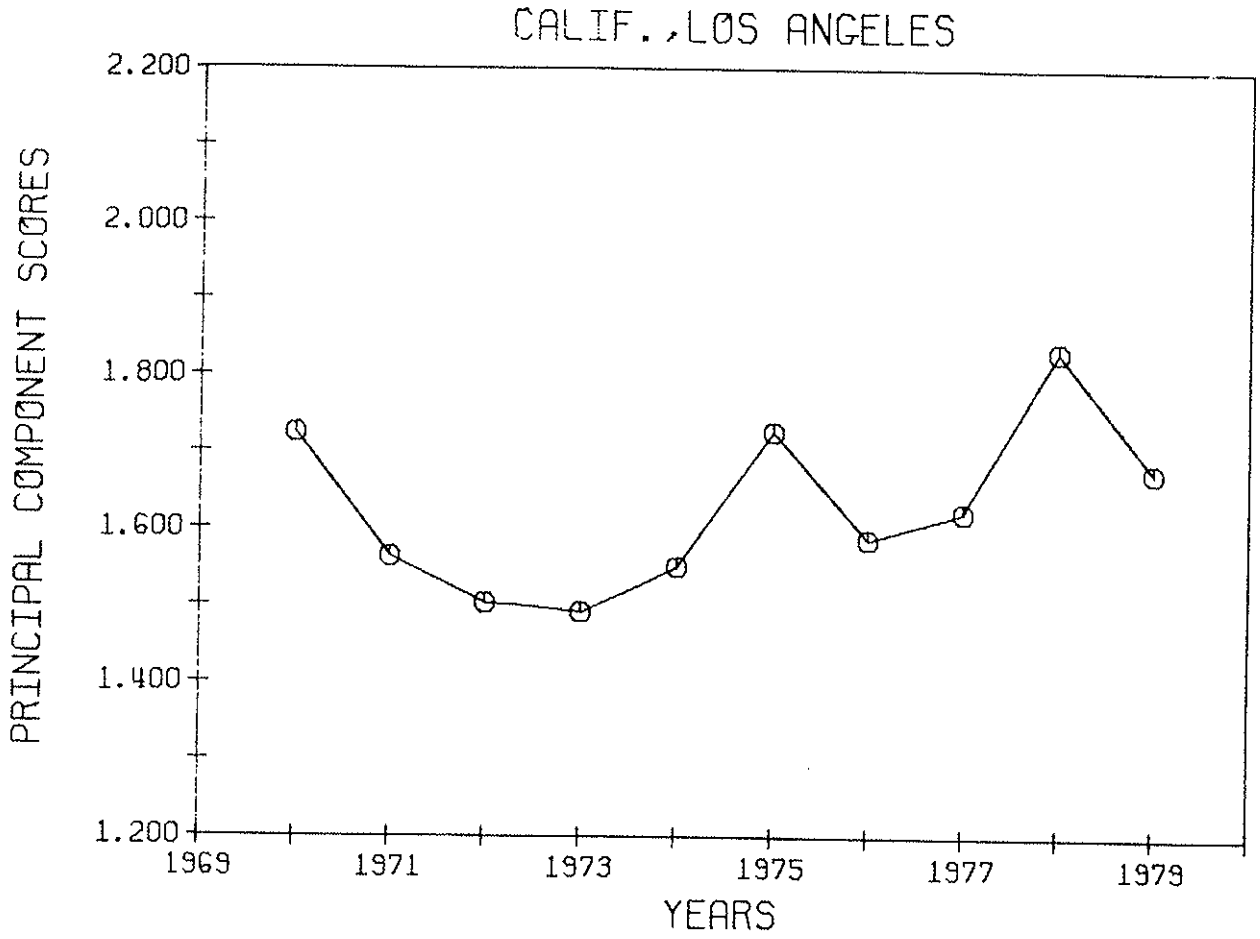

BOSTON
0
-.20000
-.40000
-.60000
-.80000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

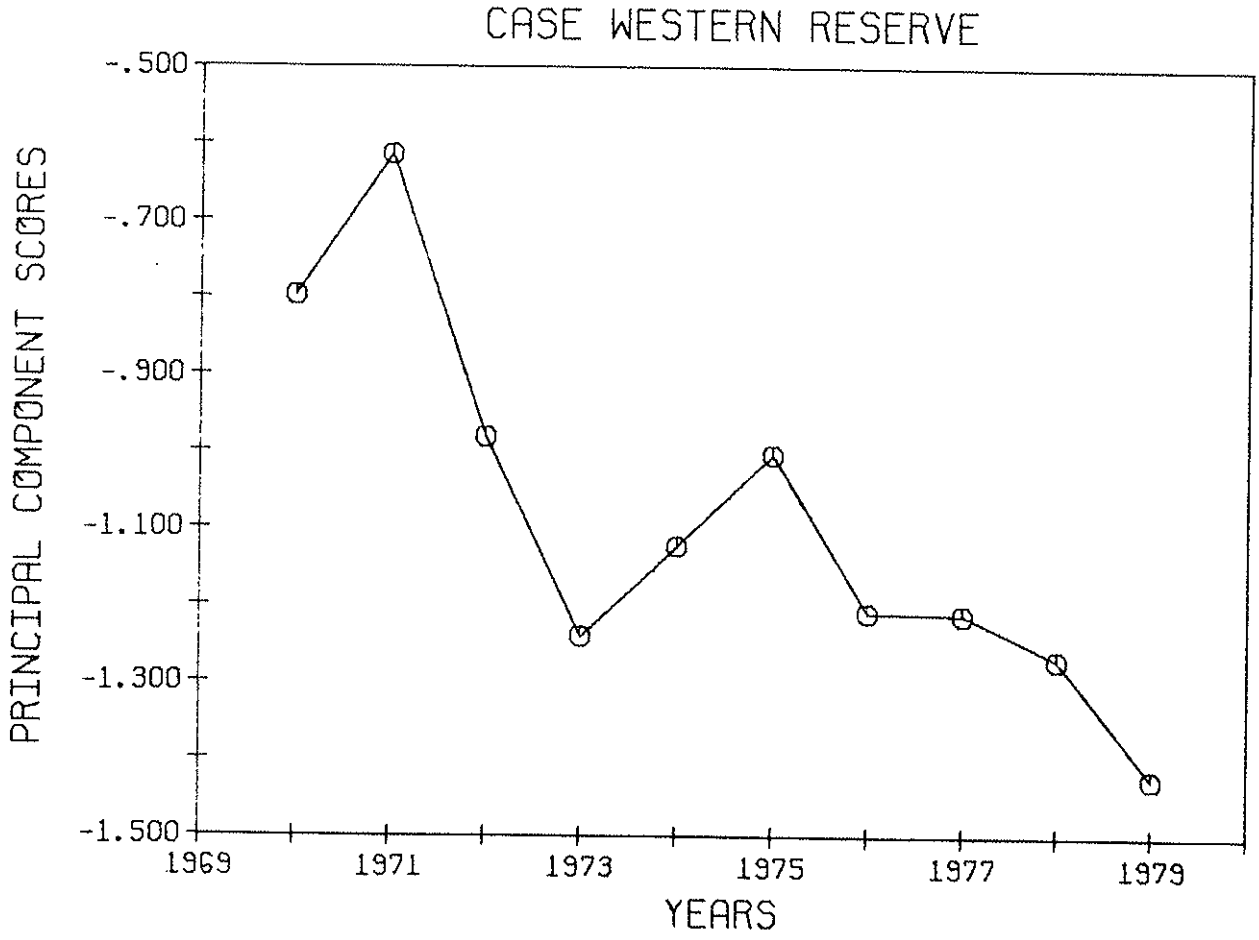

BRITISH COLUMBIA
PRINCIPAL COMPONENT SCORES
1.300
1.100
.900
.700
.500
.300
1969
1971
1973
1975
1977
1979
YEARS

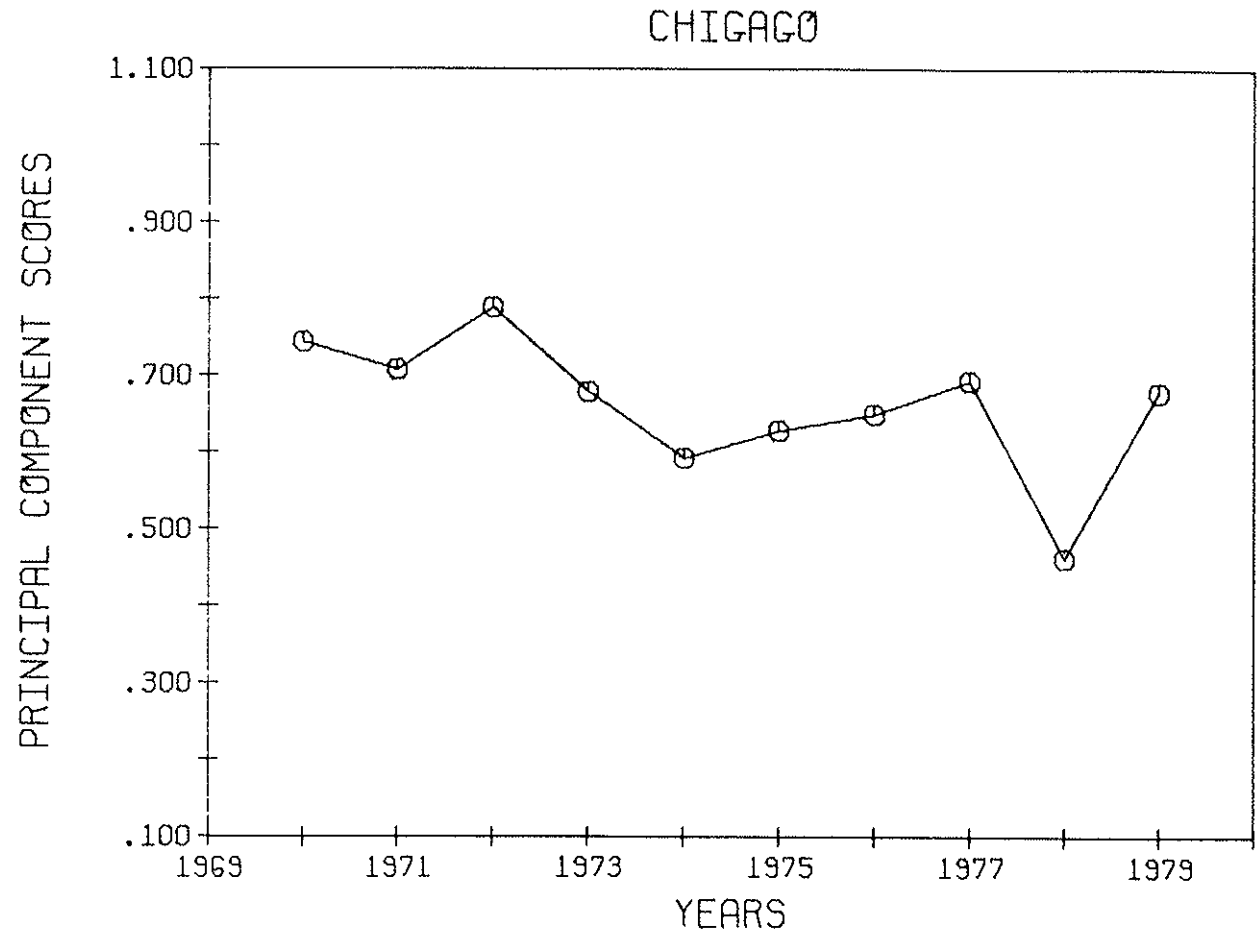

BROWN
PRINCIPAL COMPONENT SCORES
-.600
-.800
-1.000
-1.200
-1.400
-1.600
1969
1971
1973
1975
1977
1979
YEARS

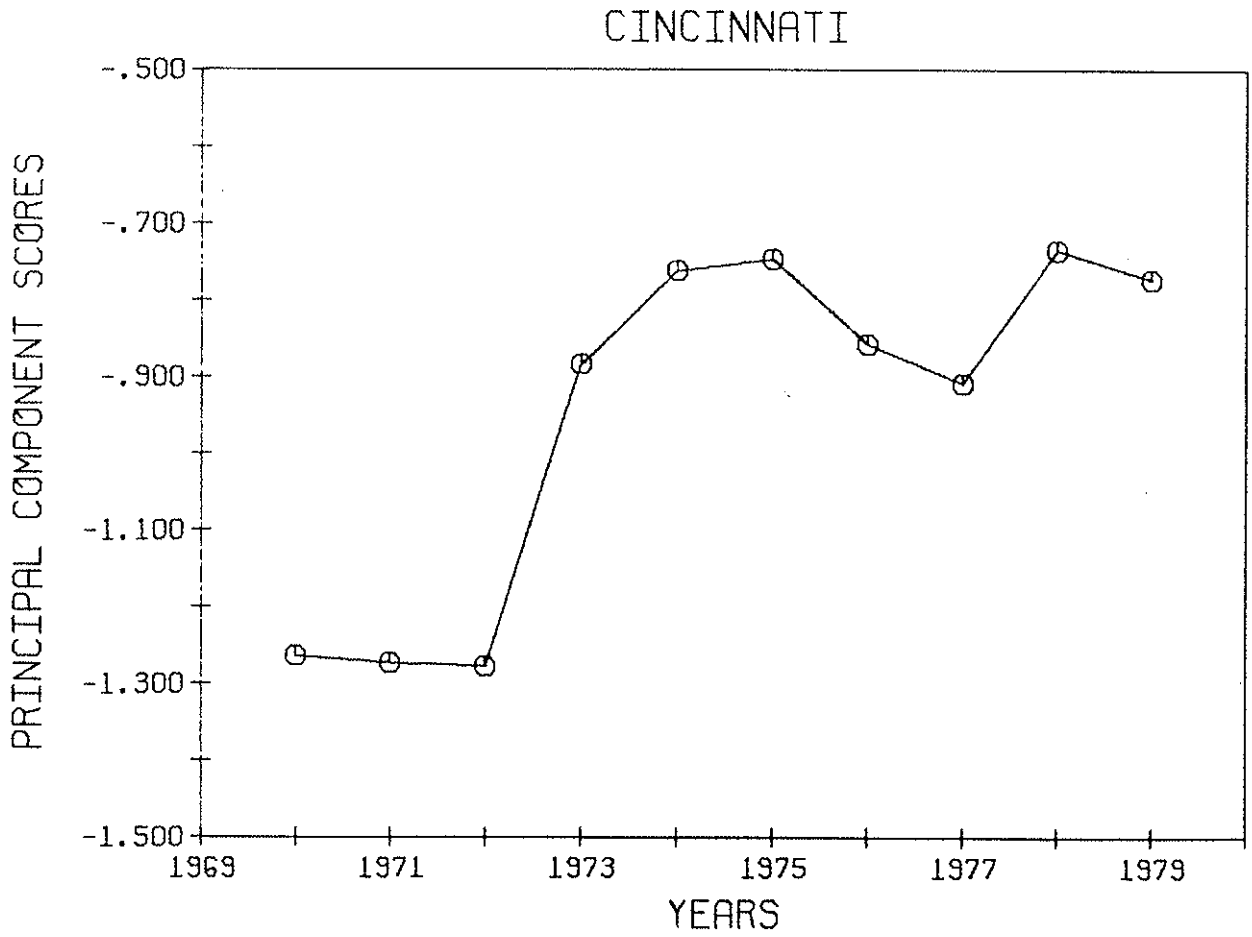

CALIF.,BERKELEY
PRINCIPAL COMPONENT SCORES
2.300
2.100
1.900
1.700
1.500
1.300
1969
1971
1973
1975
1977
1979
YEARS

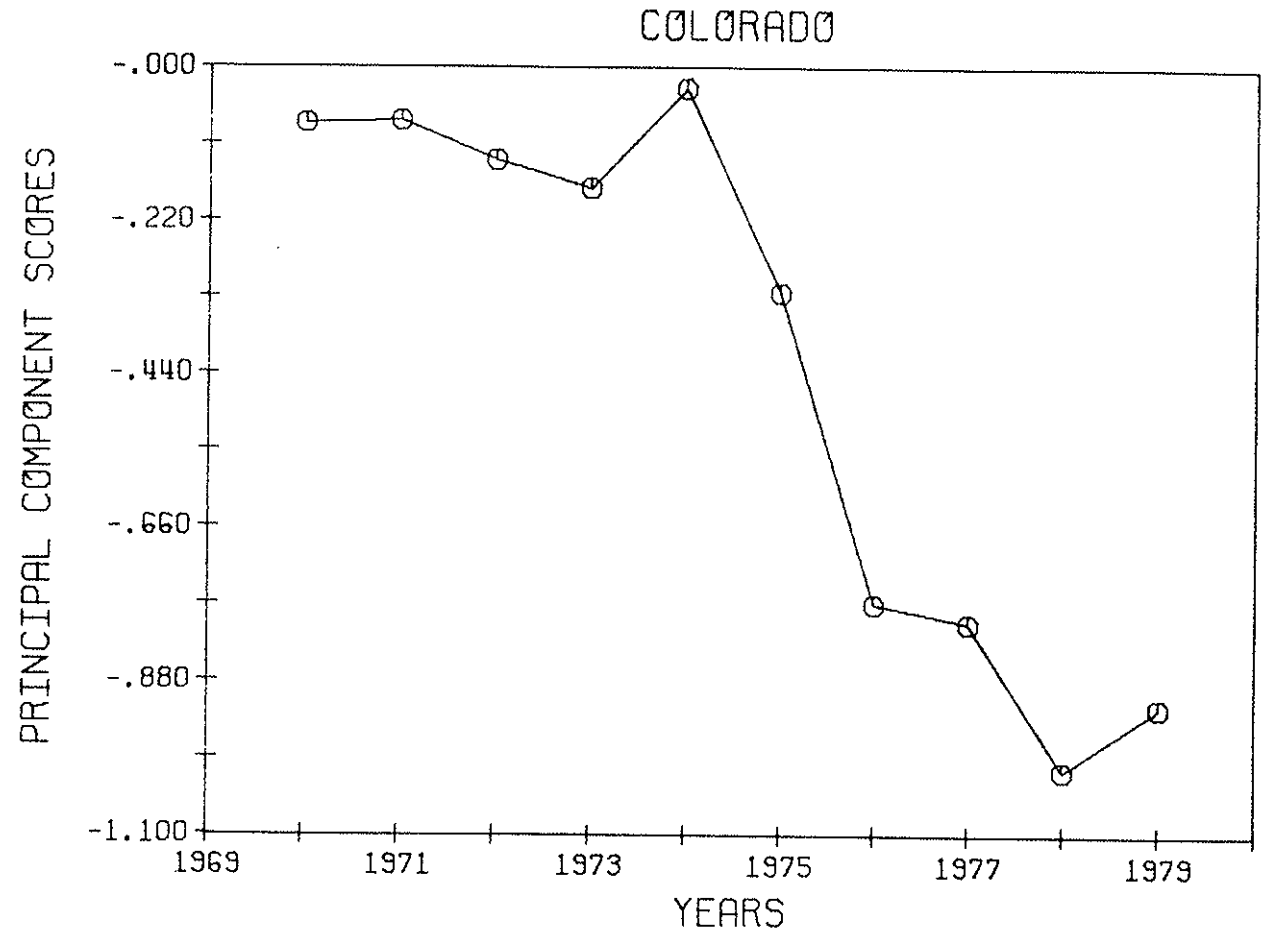

CALIF.,DAVIS
PRINCIPAL COMPONENT SCORES
.80000
.60000
.40000
.20000
-.00000
-.20000
1969
1971
1973
1975
1977
1979
YEARS

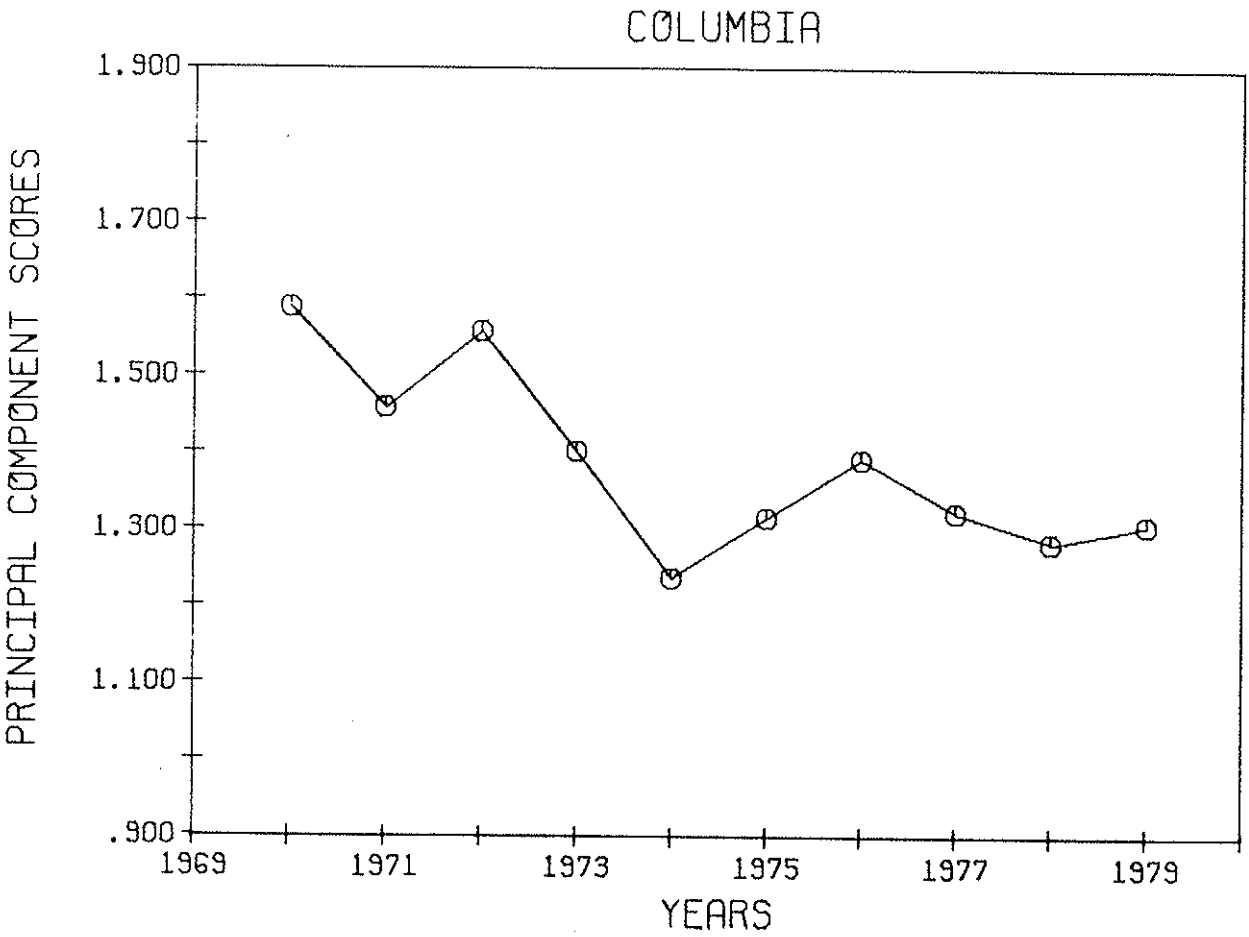

CALIF., LOS ANGELES
PRINCIPAL COMPONENT SCORES
2.200
2.000
1.800
1.600
1.400
1.200
1969
1971
1973
1975
1977
1979
YEARS

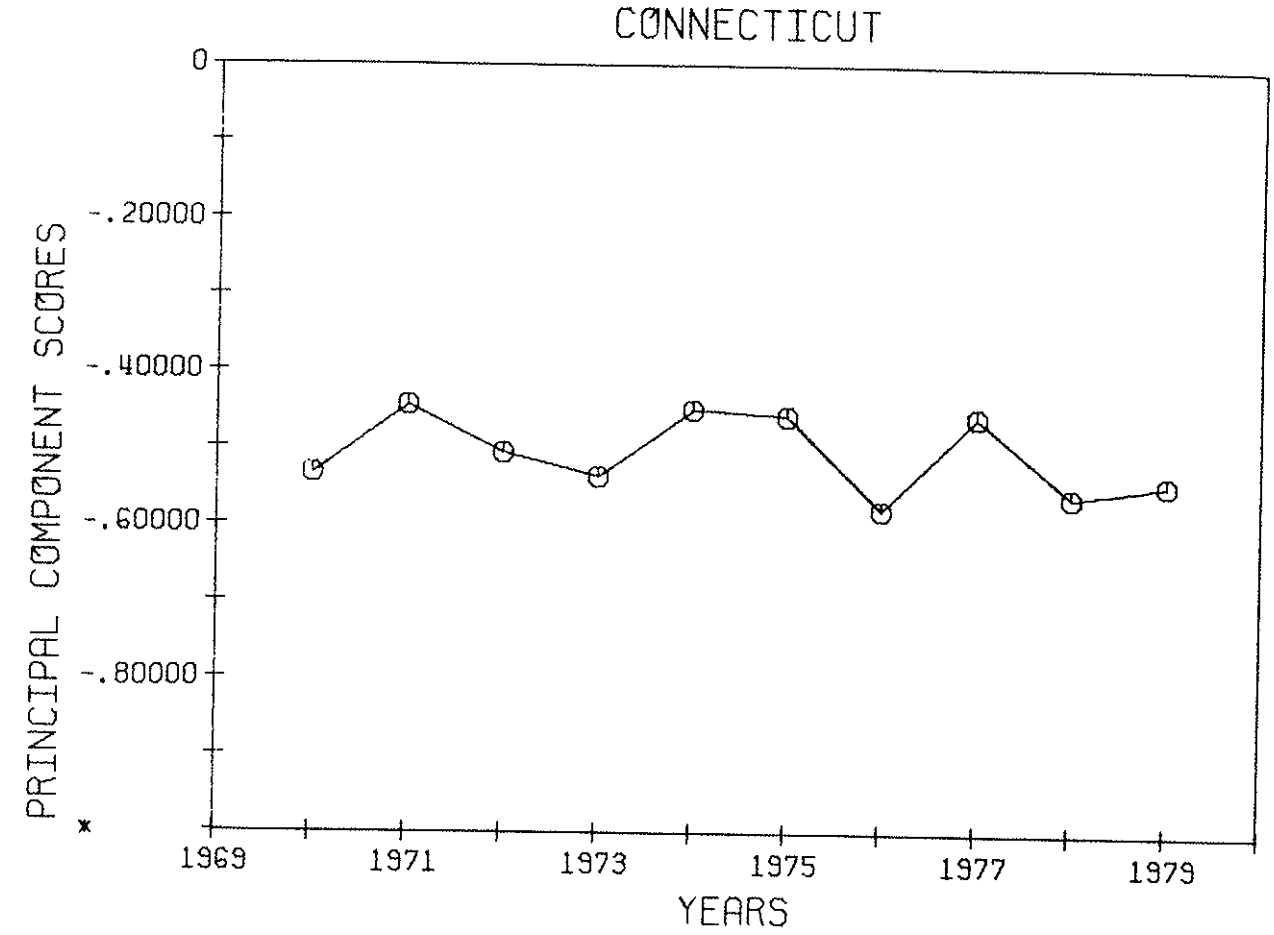

CASE WESTERN RESERVE
PRINCIPAL COMPONENT SCORES
-.500
-.700
-.900
-1.100
-1.300
-1.500
1969
1971
1973
1975
1977
1979
YEARS

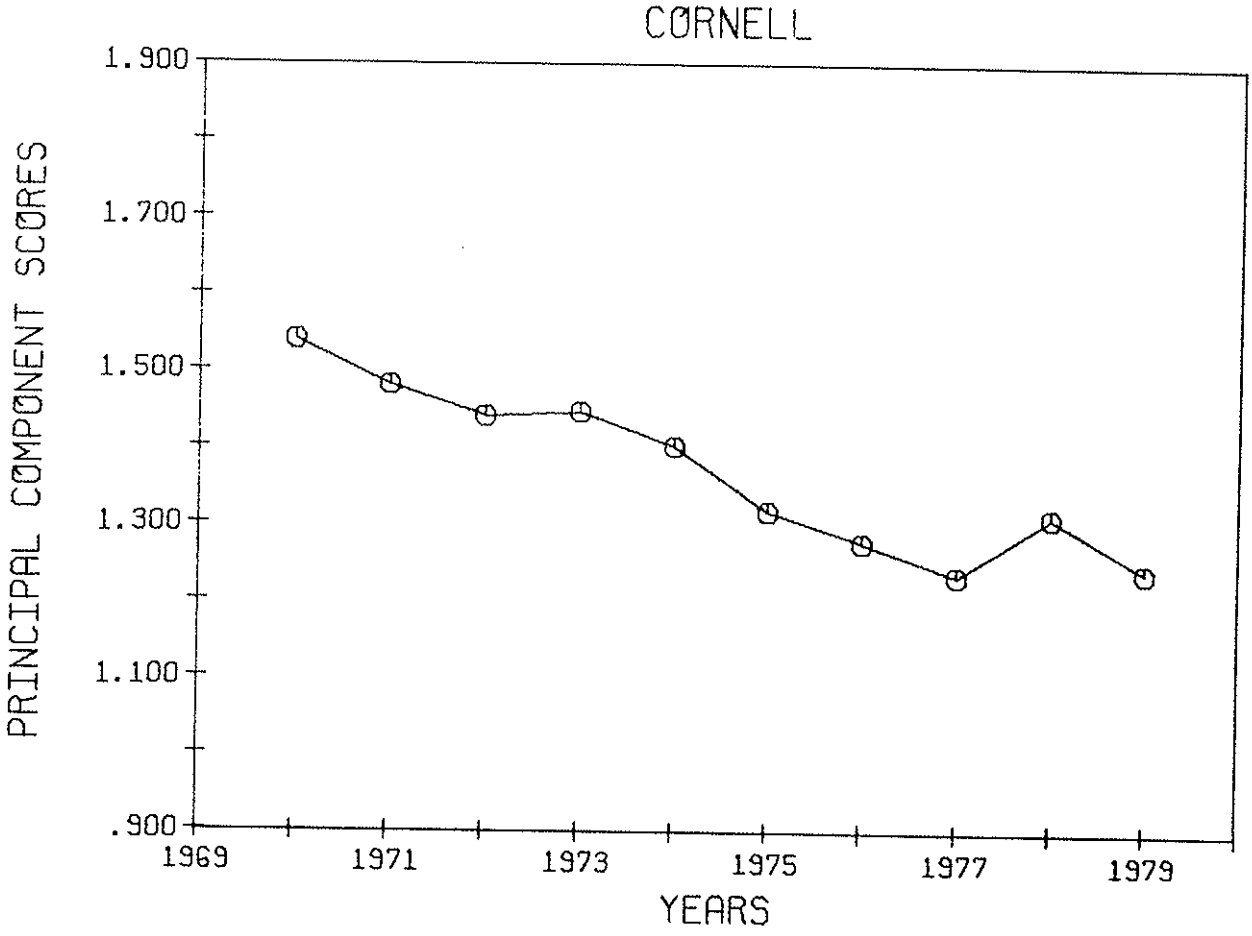

CHIGAGØ
PRINCIPAL CØMPØNENT SCØRES
1.100
.900
.700
.500
.300
.100
1969
1971
1973
1975
1977
1979
YEARS

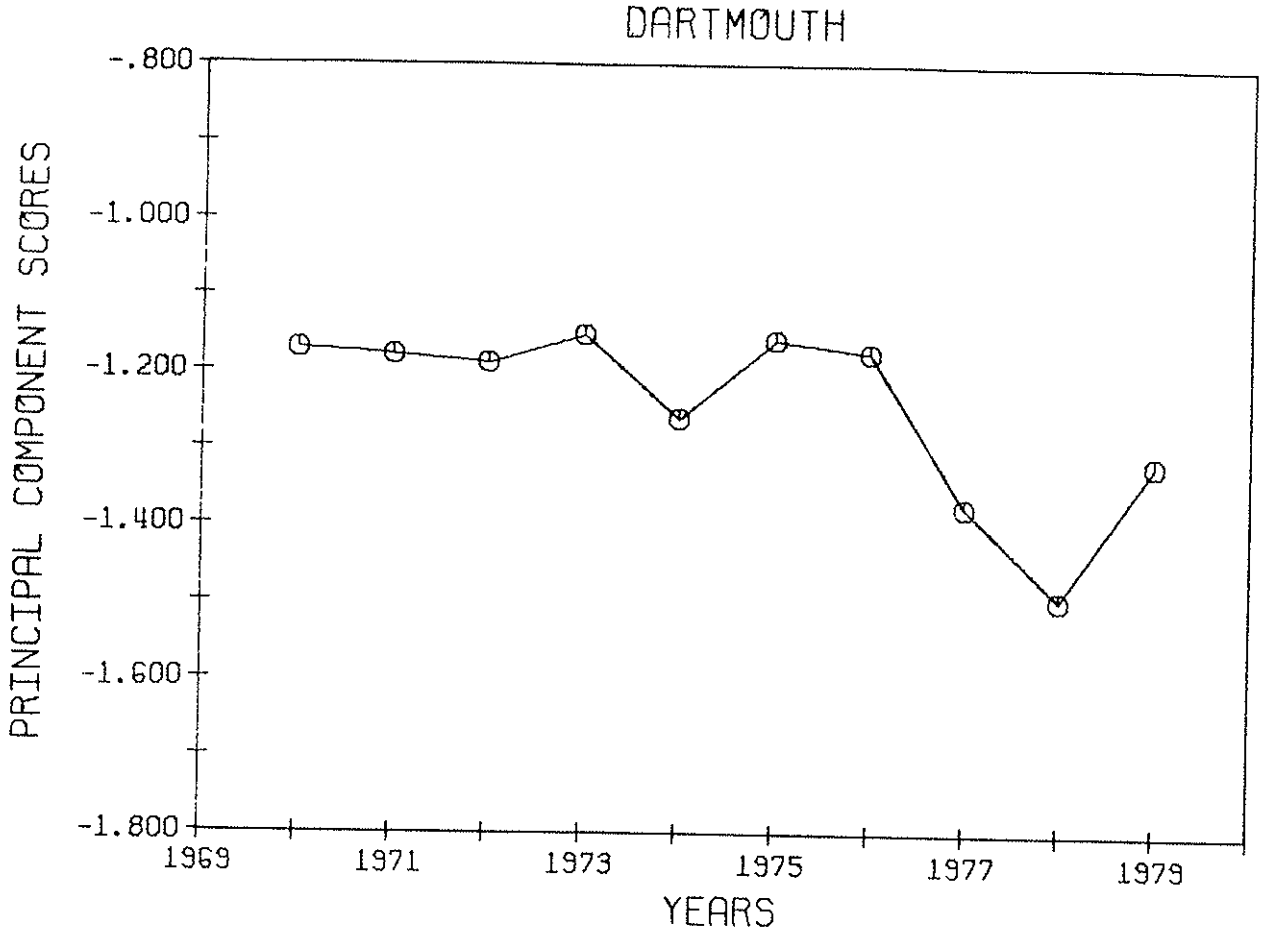

CINCINNATI
PRINCIPAL CØMPØNENT SCØRES
-.500
-.700
-.900
-1.100
-1.300
-1.500
1969
1971
1973
1975
1977
1979
YEARS

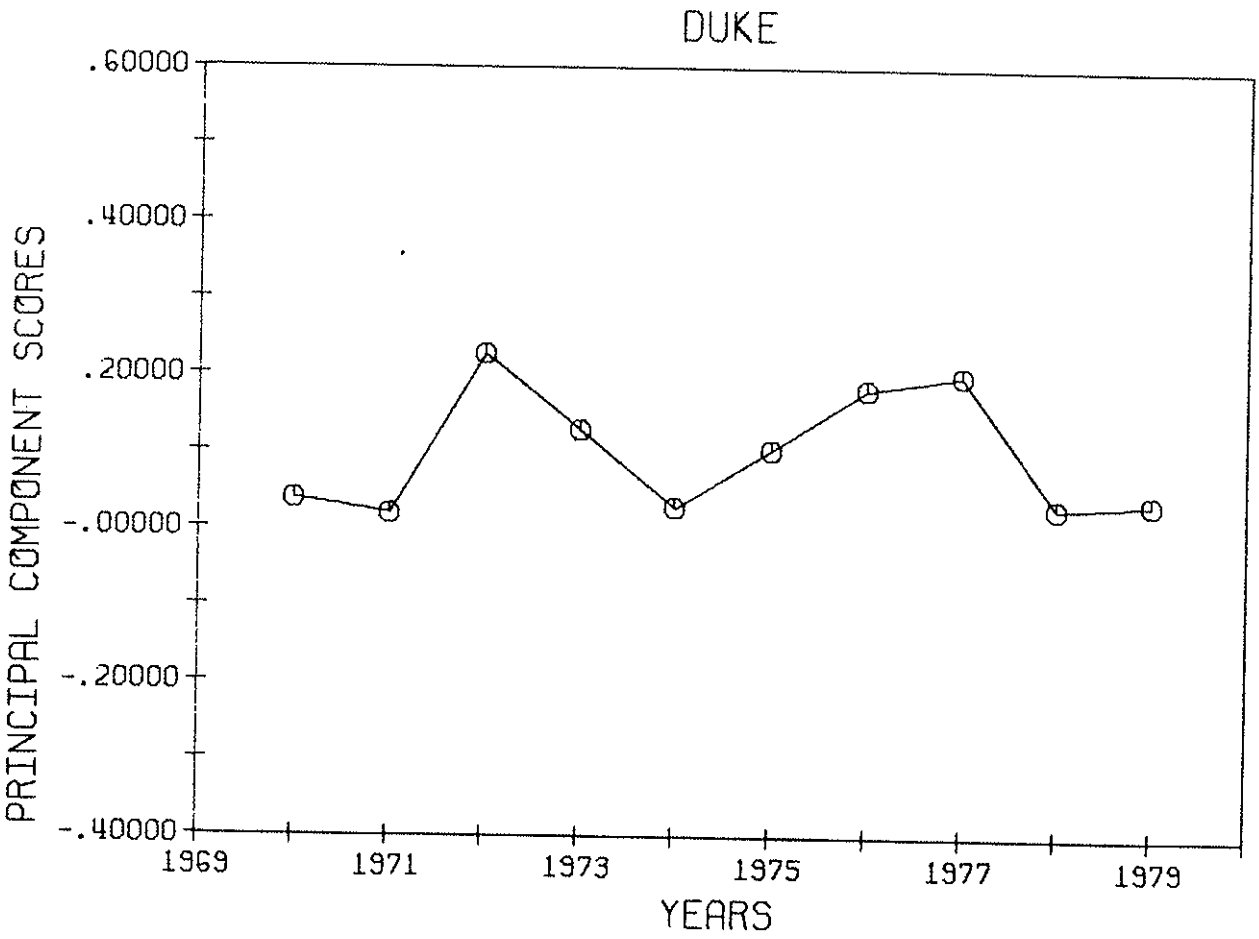

COLORADO
-.000
-.220
-.440
-.660
-.880
-1.100
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

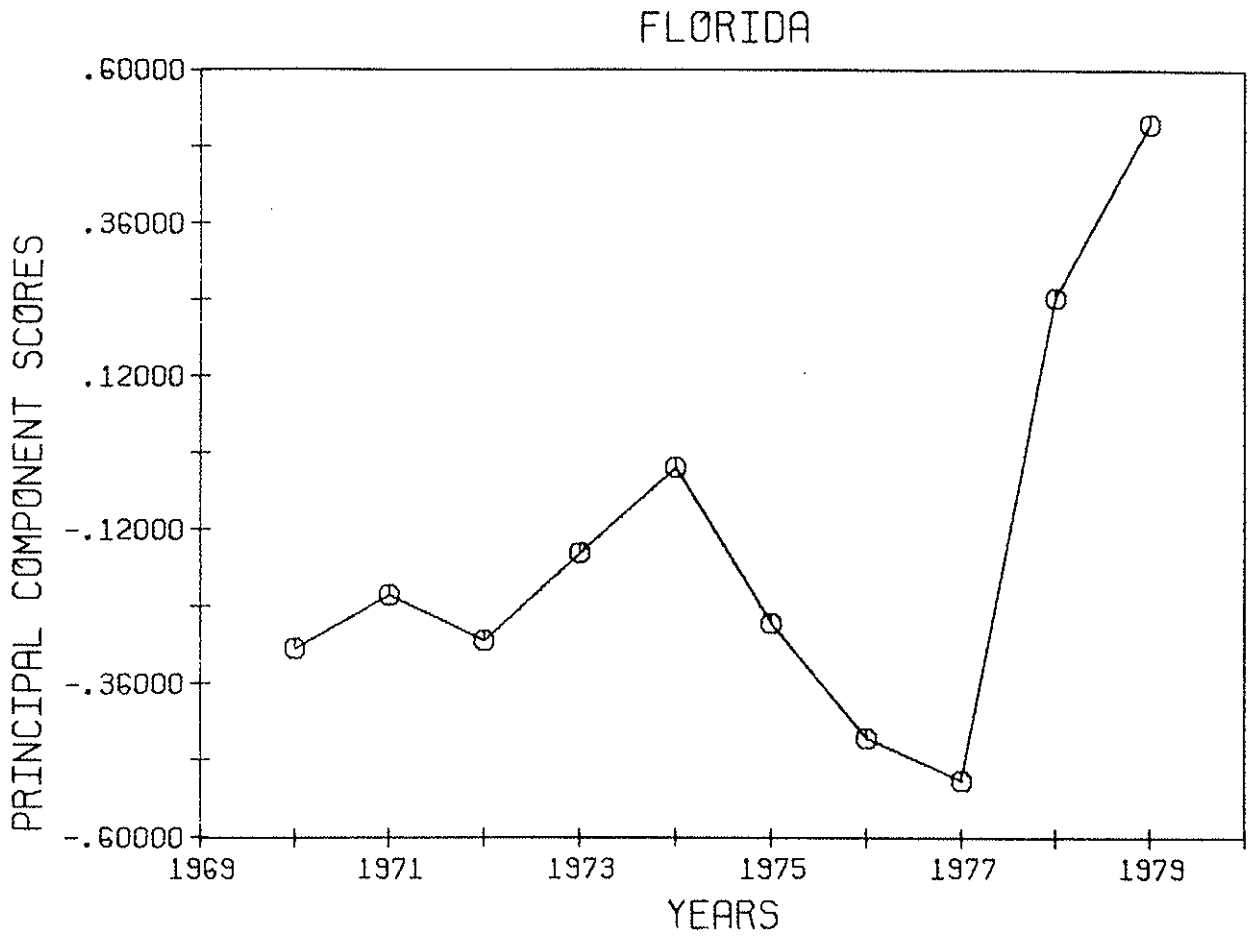

COLUMBIA
1.900
1.700
1.500
1.300
1.100
.900
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

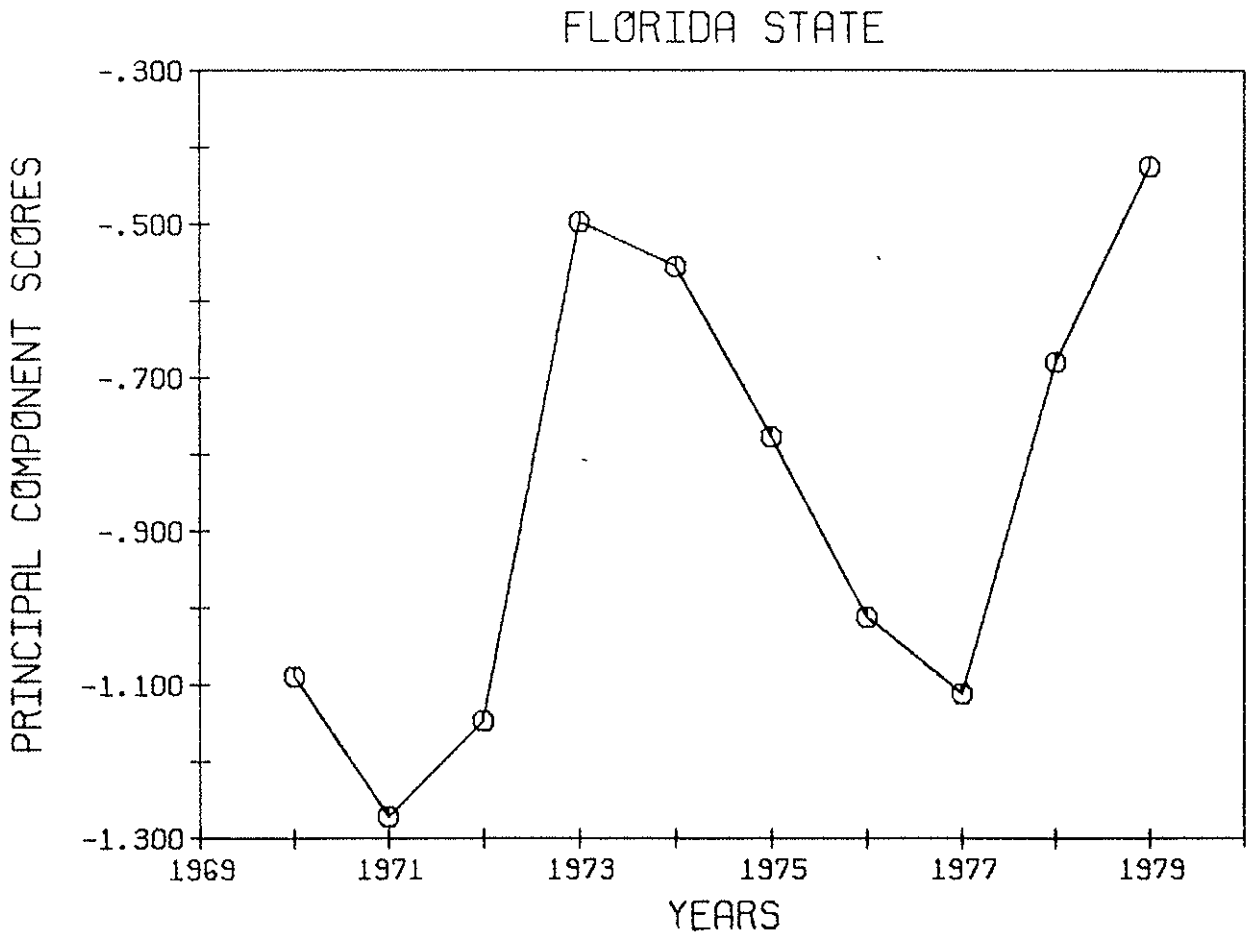

CONNECTICUT
0
-.20000
-.40000
-.60000
-.80000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

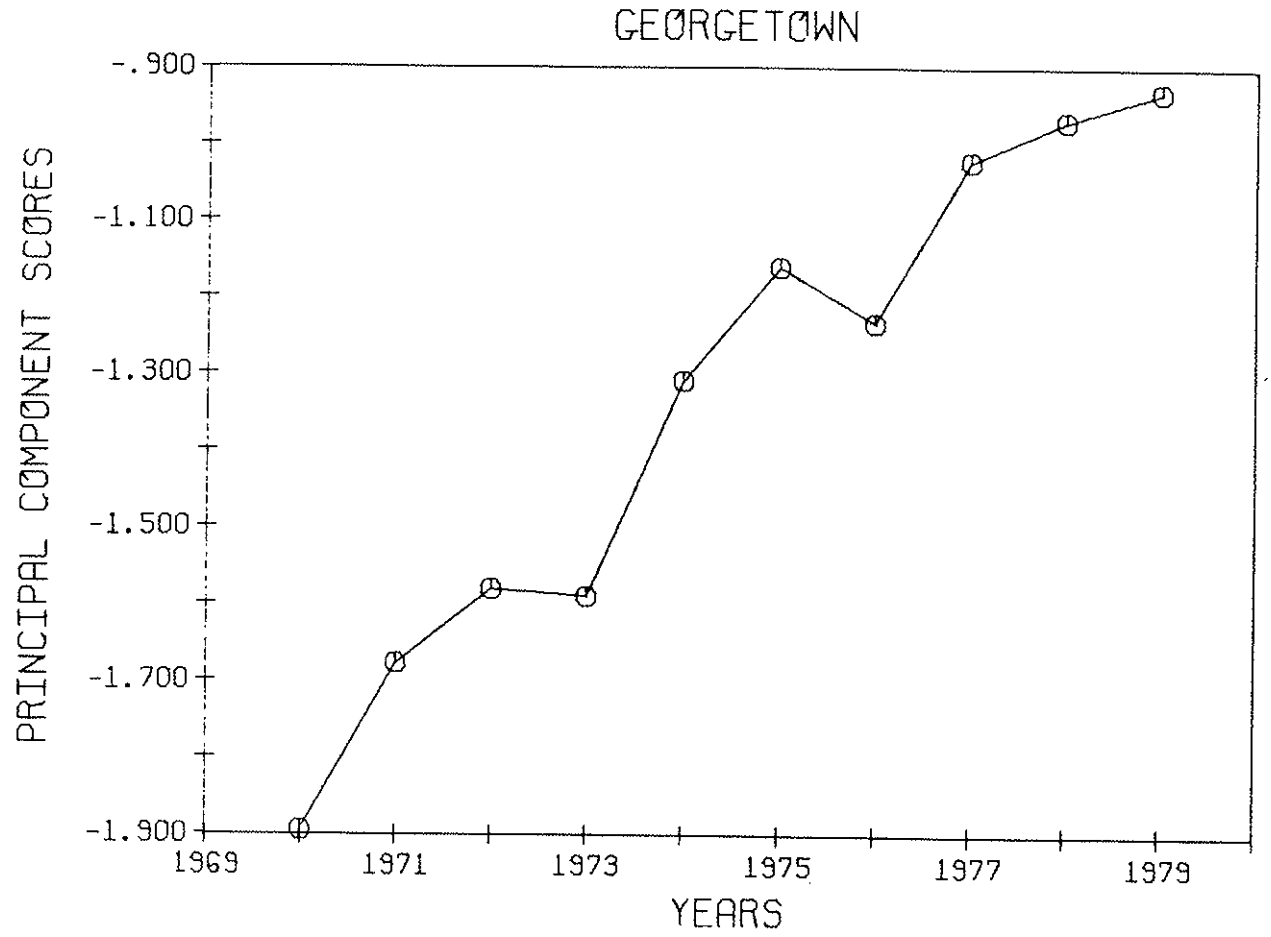

CORNELL
1.900
1.700
1.500
1.300
1.100
.900
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

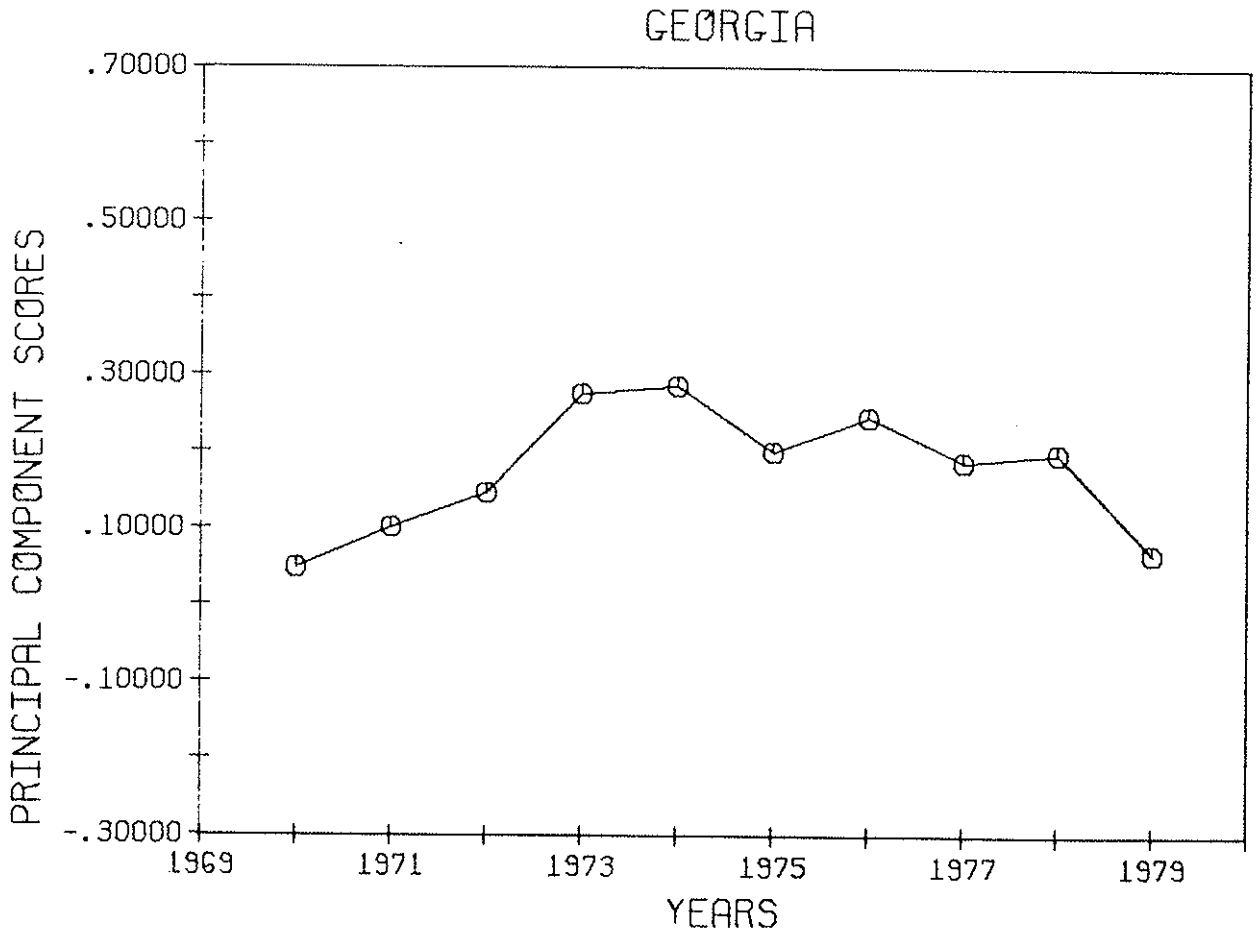

DARTMOUTH
-.800
-1.000
-1.200
-1.400
-1.600
-1.800
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

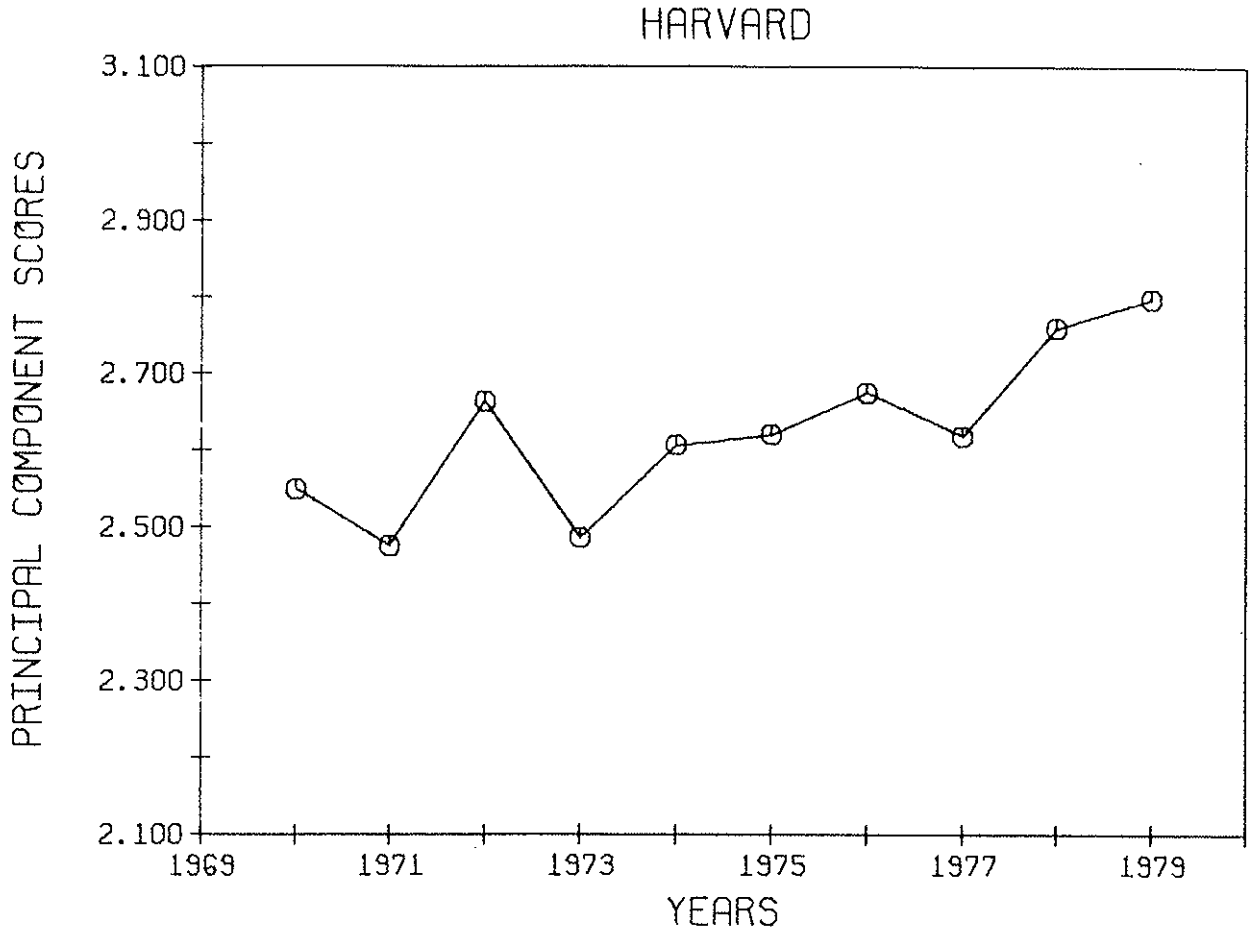

DUKE
.60000
.40000
.20000
-.00000
-.20000
-.40000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

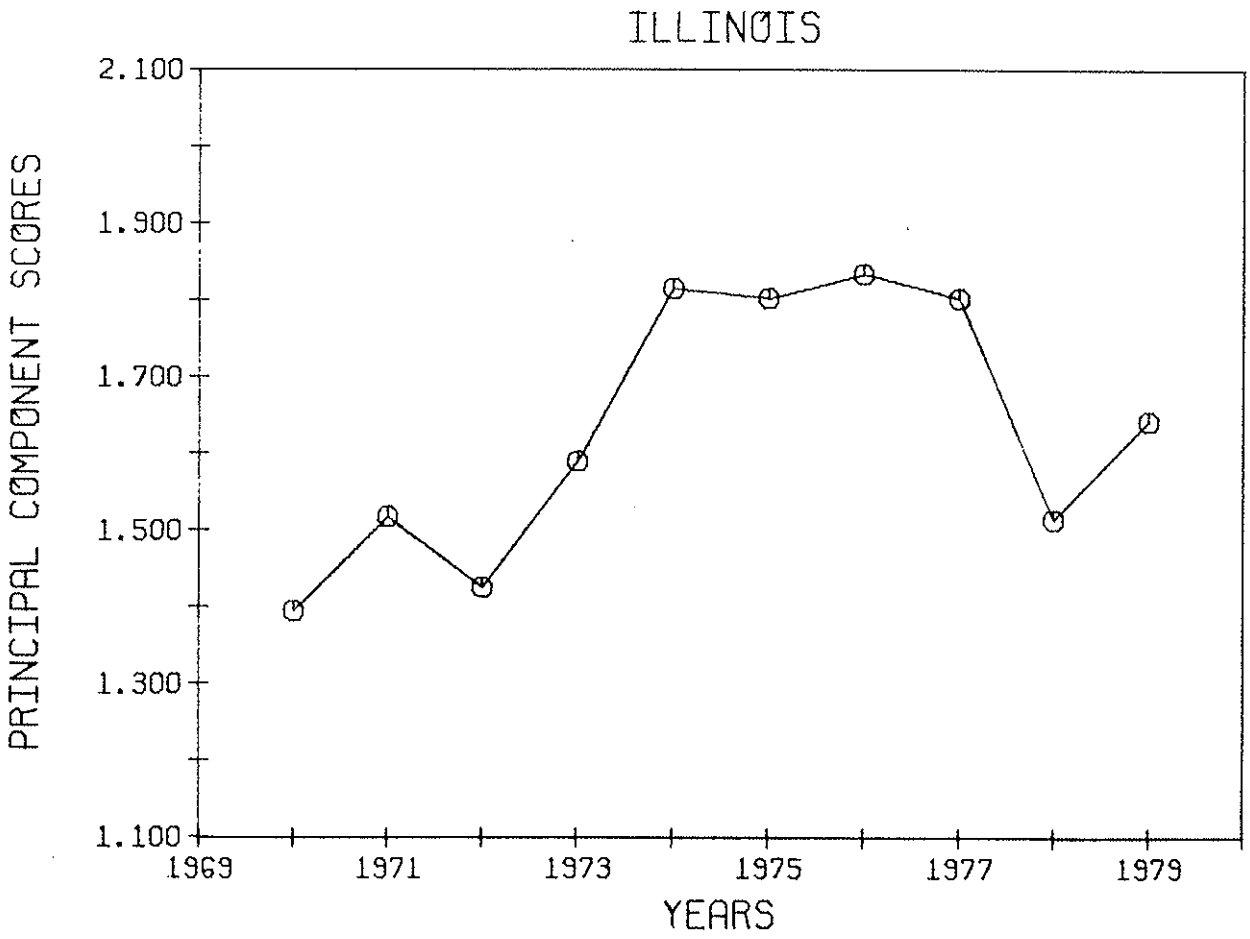

FLØRIDA
.60000
.36000
.12000
-.12000
-.36000
-.60000
PRINCIPAL CØMPØNENT SCØRES
1969
1971
1973
1975
1977
1979
YEARS

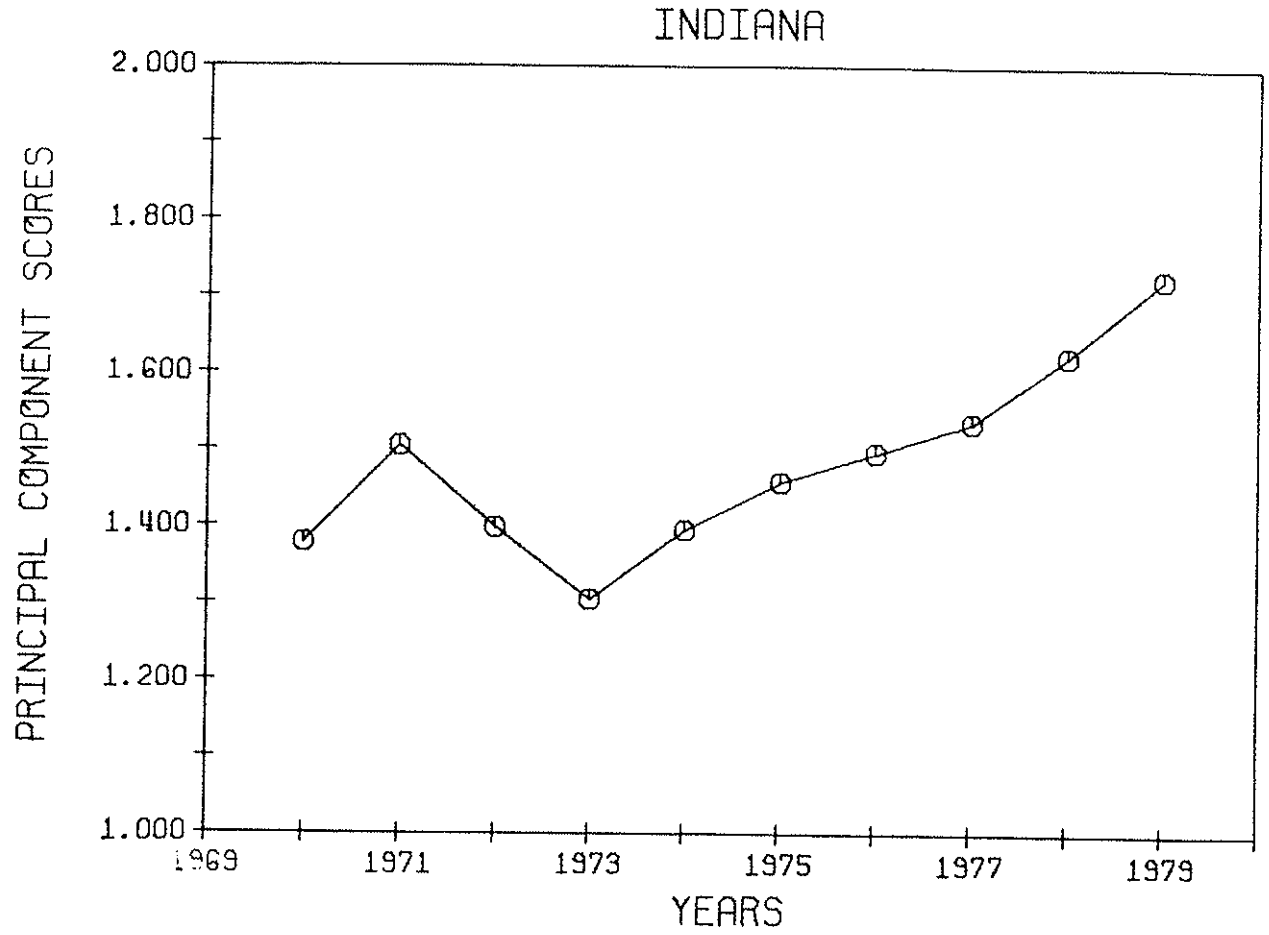

FLØRIDA STATE
-.300
-.500
-.700
-.900
-1.100
-1.300
PRINCIPAL CØMPØNENT SCØRES
1969
1971
1973
1975
1977
1979
YEARS

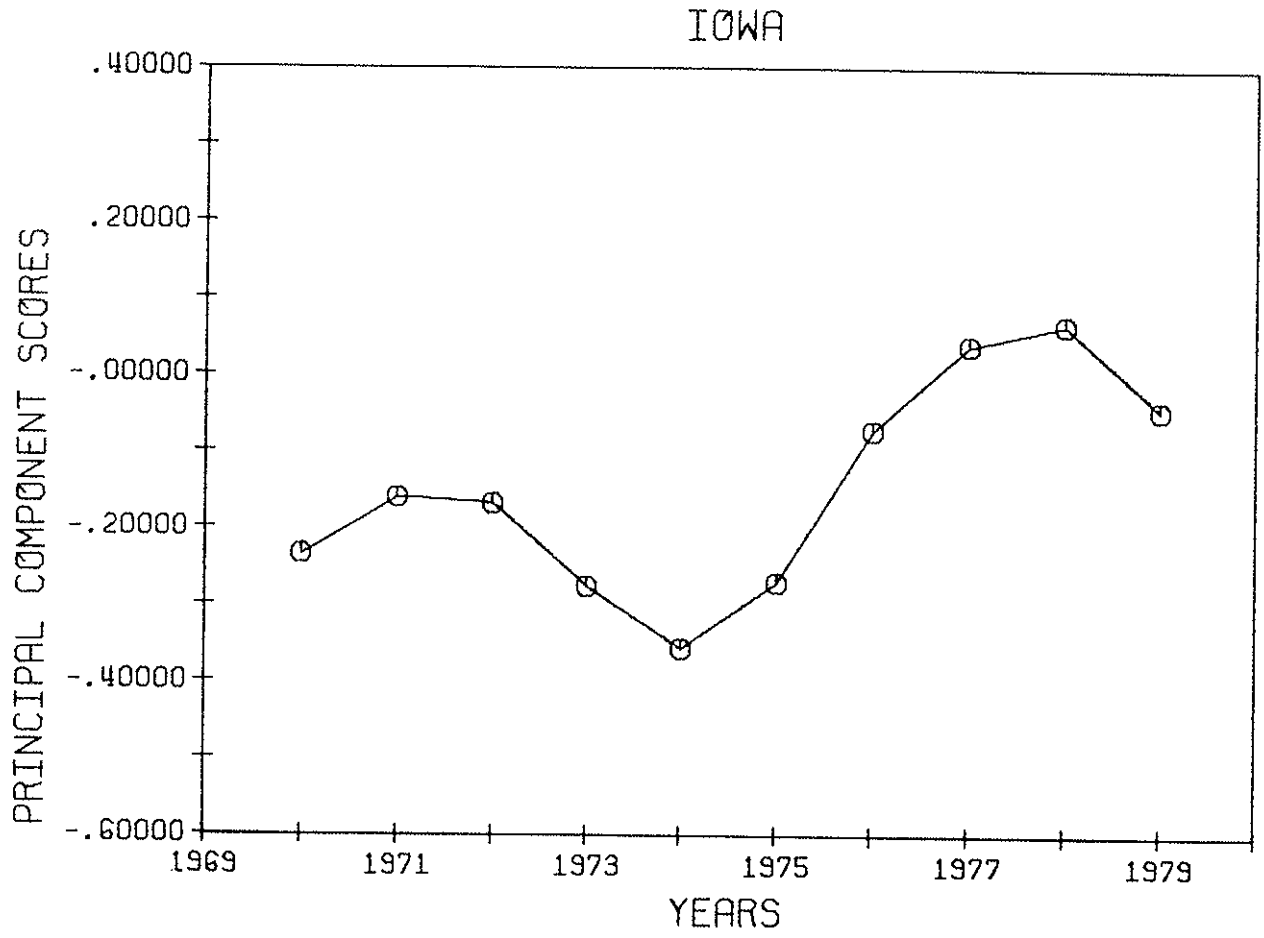

GEORGETOWN
PRINCIPAL COMPONENT SCORES
-.900
-1.100
-1.300
-1.500
-1.700
-1.900
1969
1971
1973
1975
1977
1979
YEARS

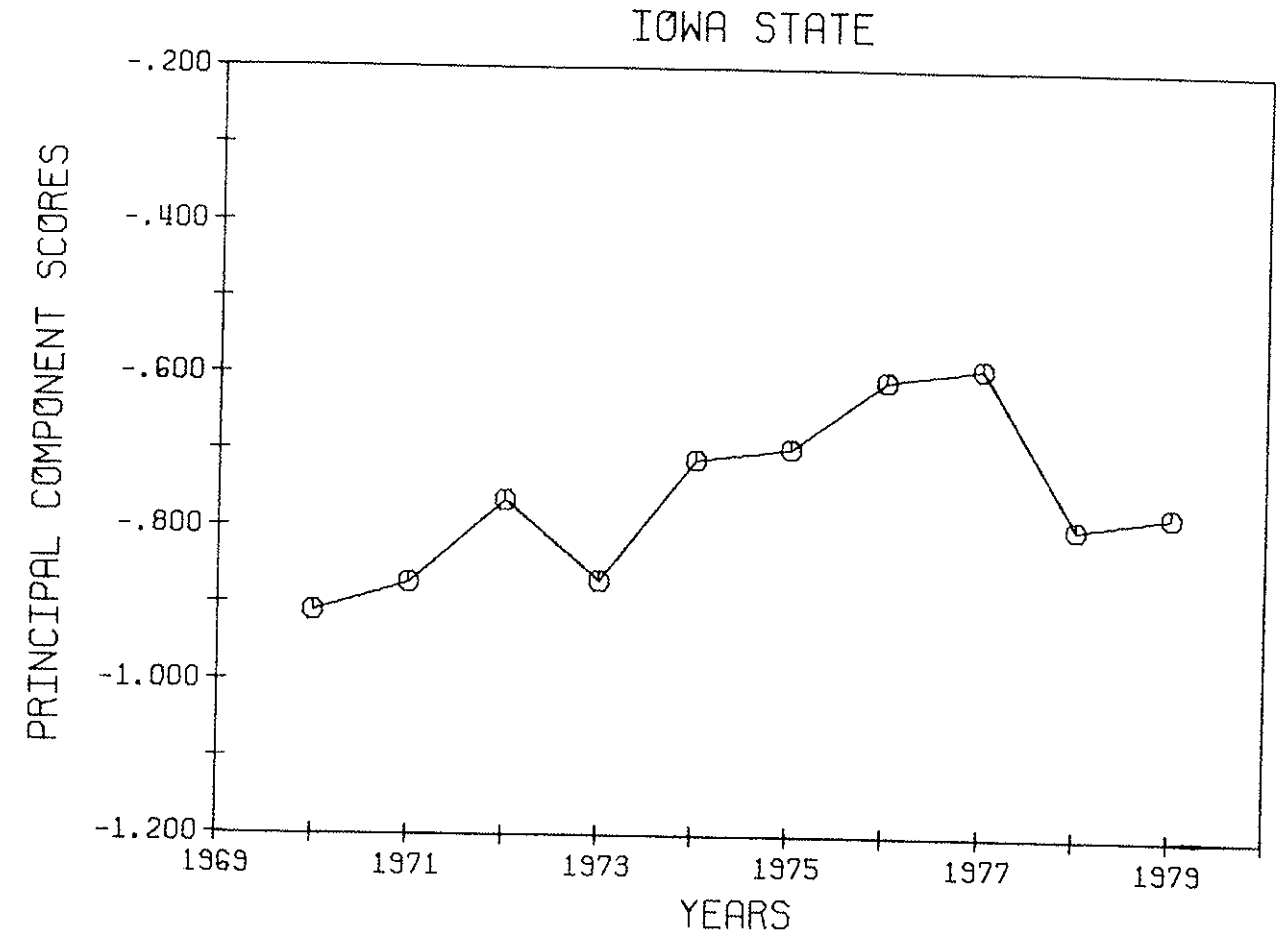

GEORGIA
PRINCIPAL COMPONENT SCORES
.70000
.50000
.30000
.10000
-.10000
-.30000
1969
1971
1973
1975
1977
1979
YEARS

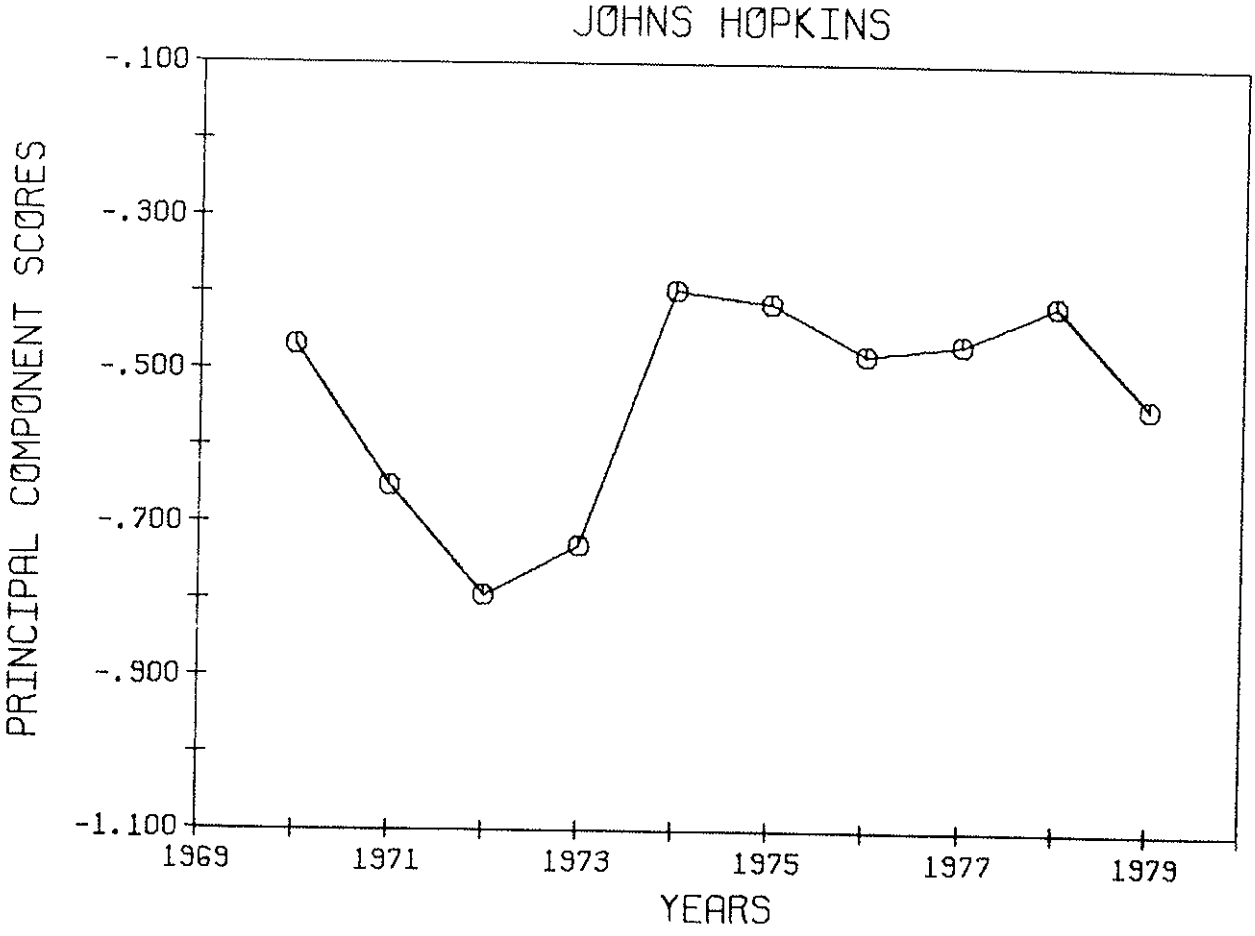

HARVARD
PRINCIPAL COMPONENT SCORES
3.100
2.900
2.700
2.500
2.300
2.100
1969
1971
1973
1975
1977
1979
YEARS

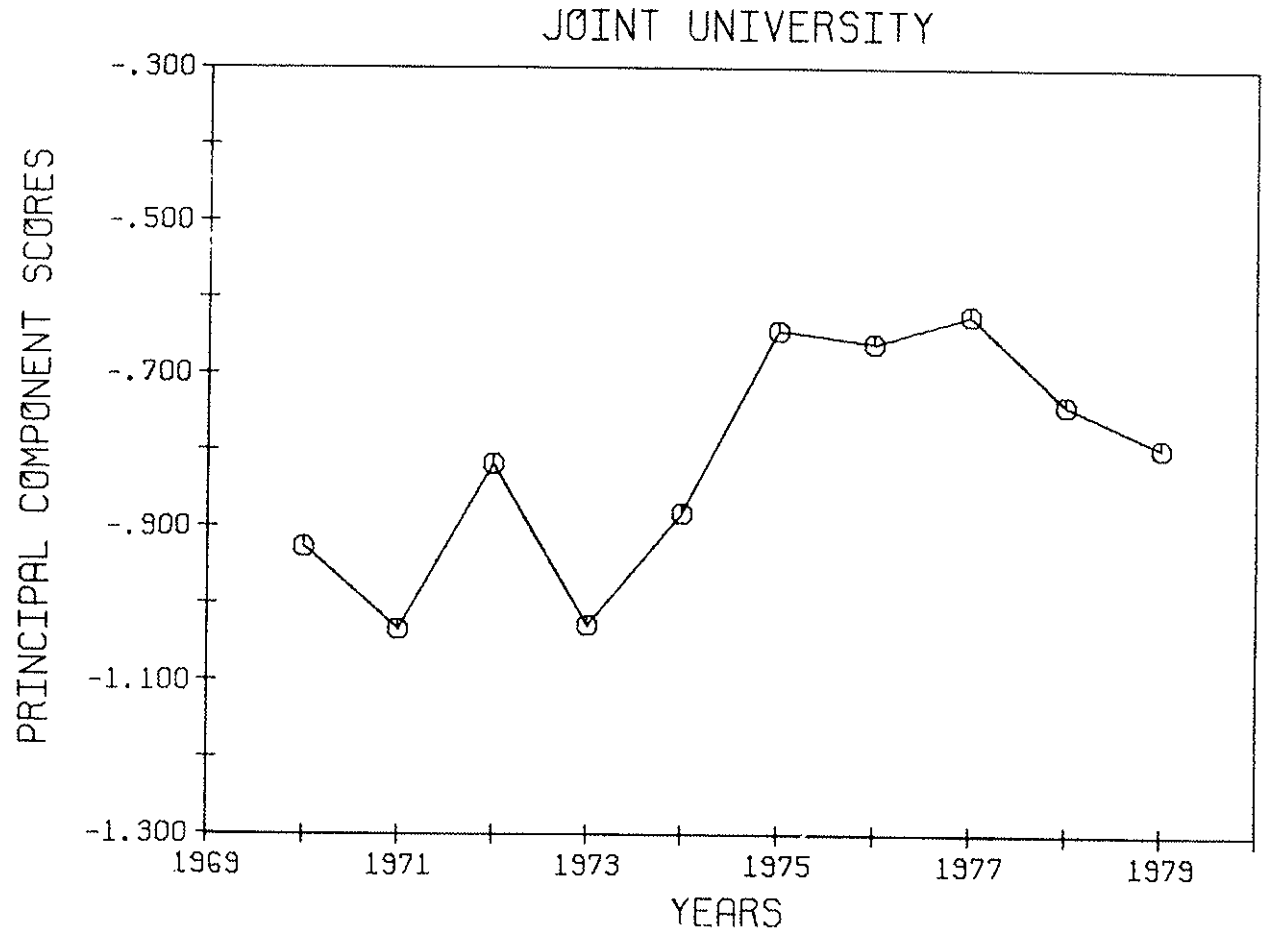

ILLINOIS
PRINCIPAL COMPONENT SCORES
2.100
1.900
1.700
1.500
1.300
1.100
1969
1971
1973
1975
1977
1979
YEARS

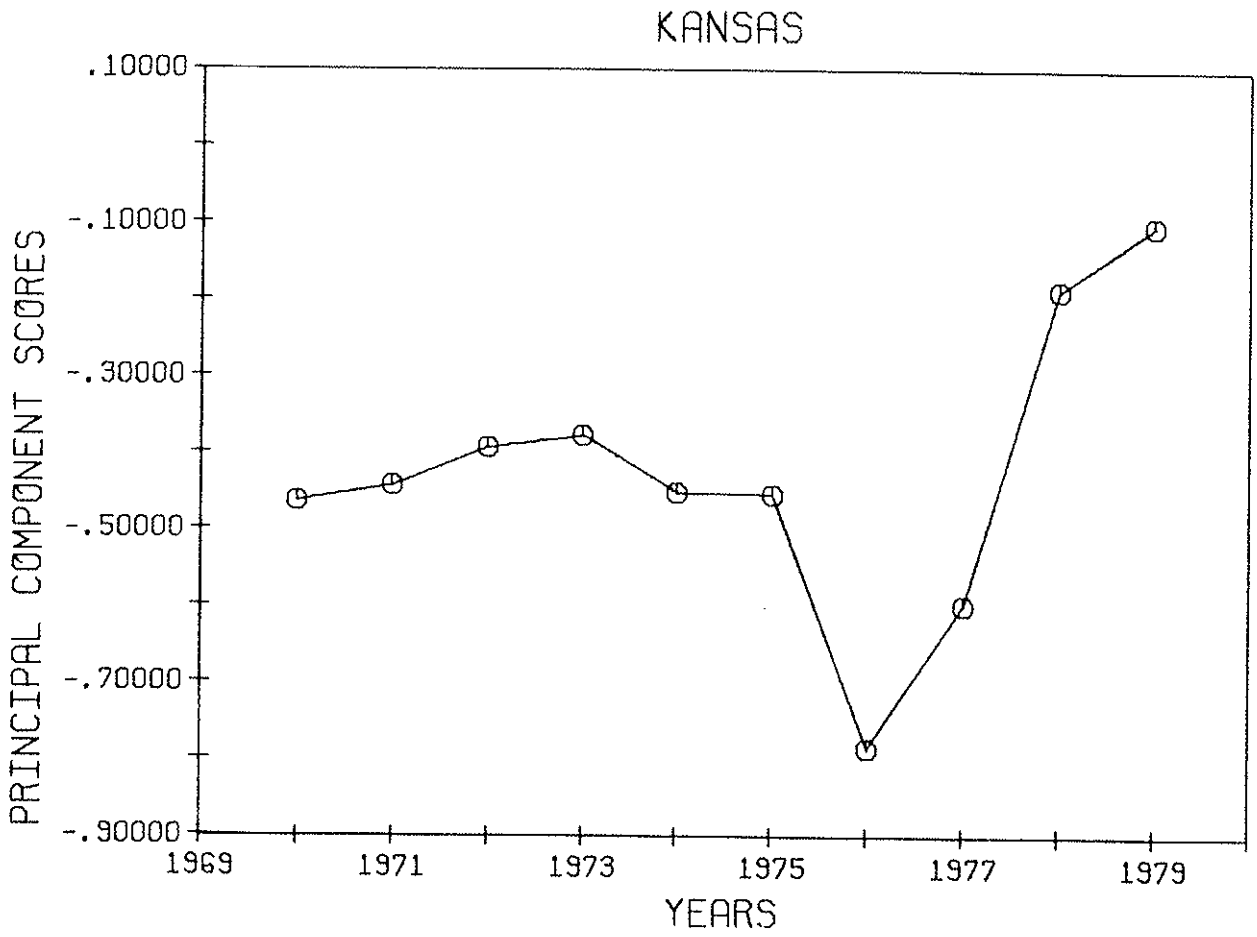

INDIANA
2.000
1.800
1.600
1.400
1.200
1.000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

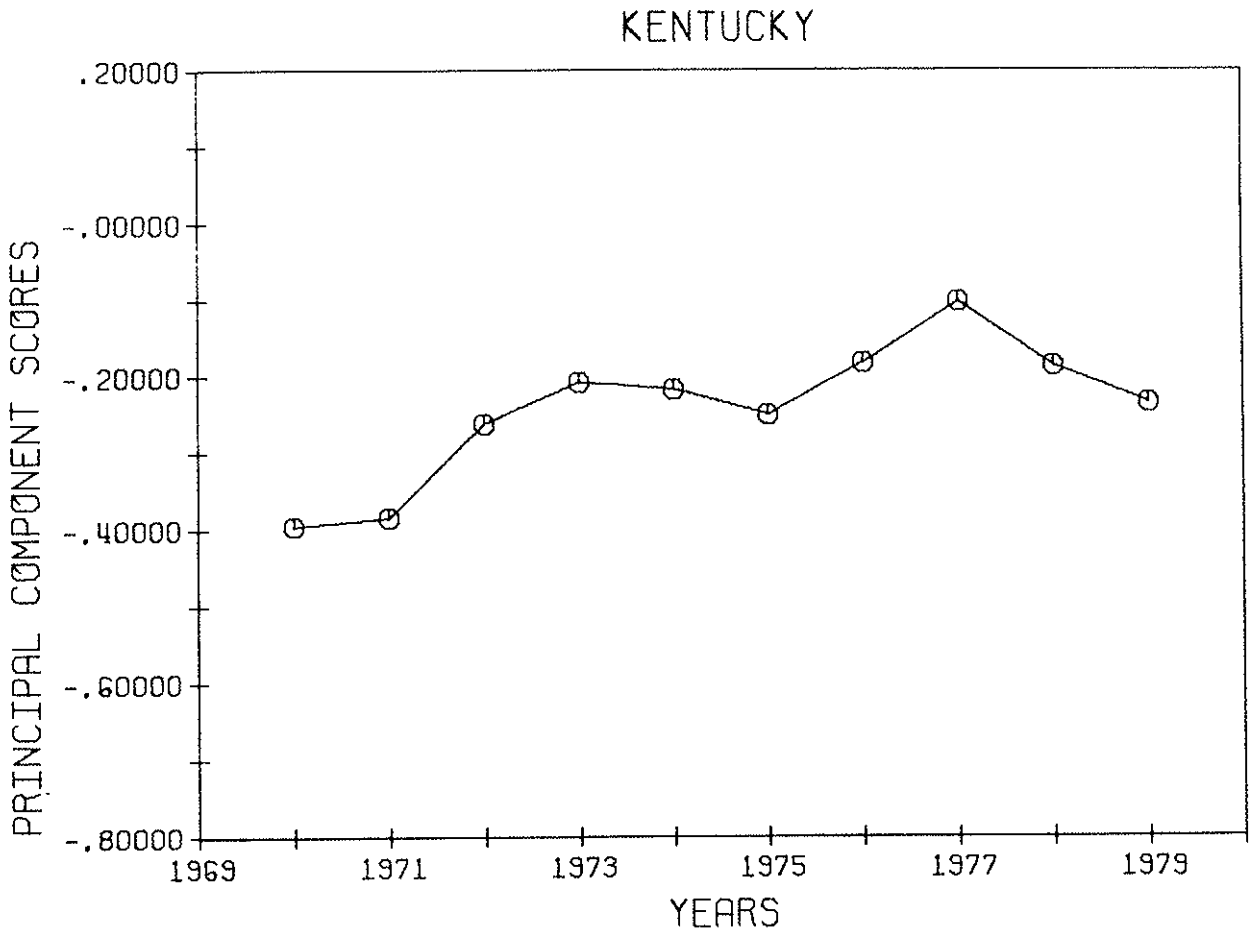

IOWA
.40000
.20000
-.00000
-.20000
-.40000
-.60000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

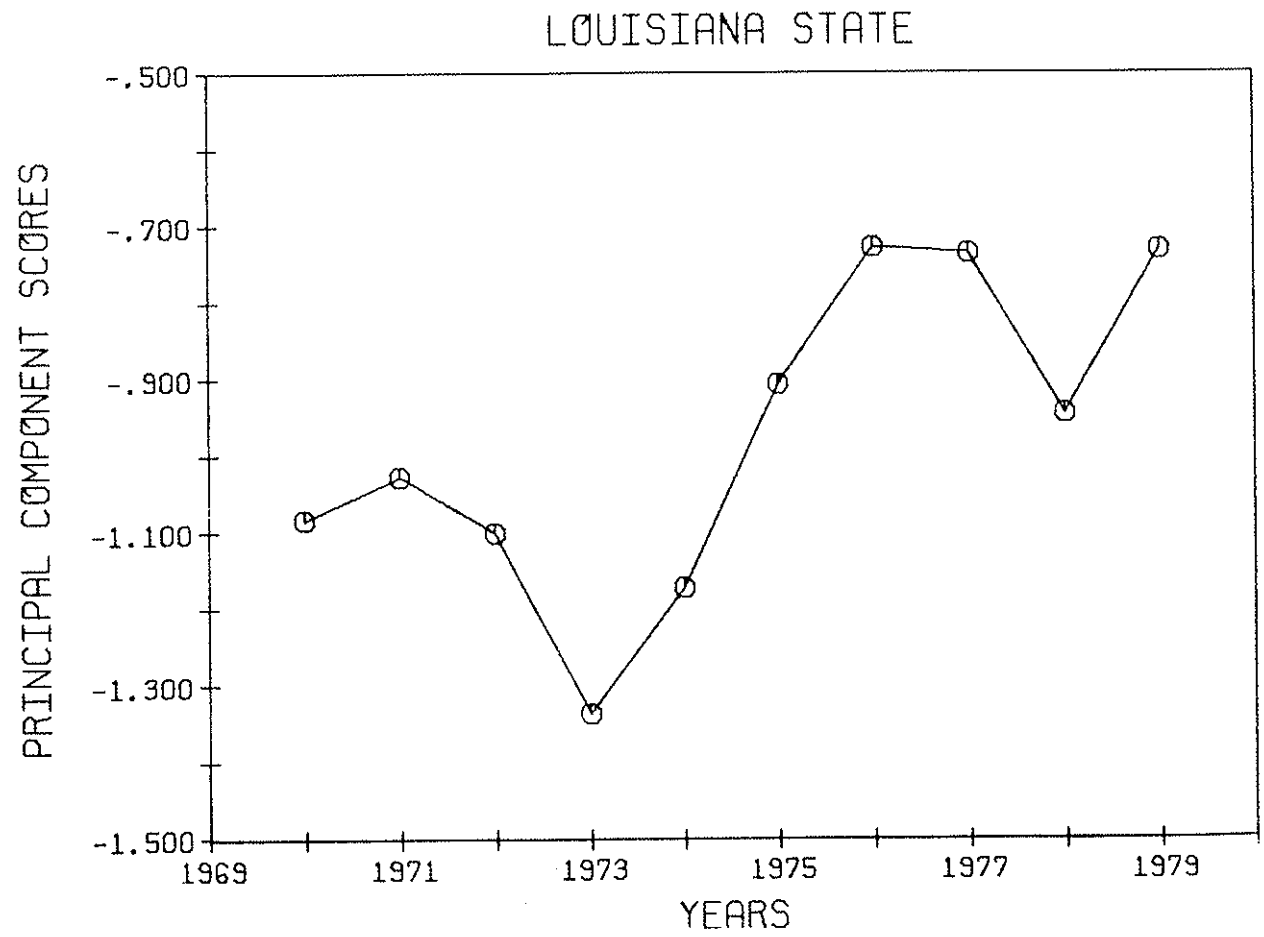

IOWA STATE
PRINCIPAL COMPONENT SCORES
-.200
-.400
-.600
-.800
-1.000
-1.200
1969
1971
1973
1975
1977
1979
YEARS

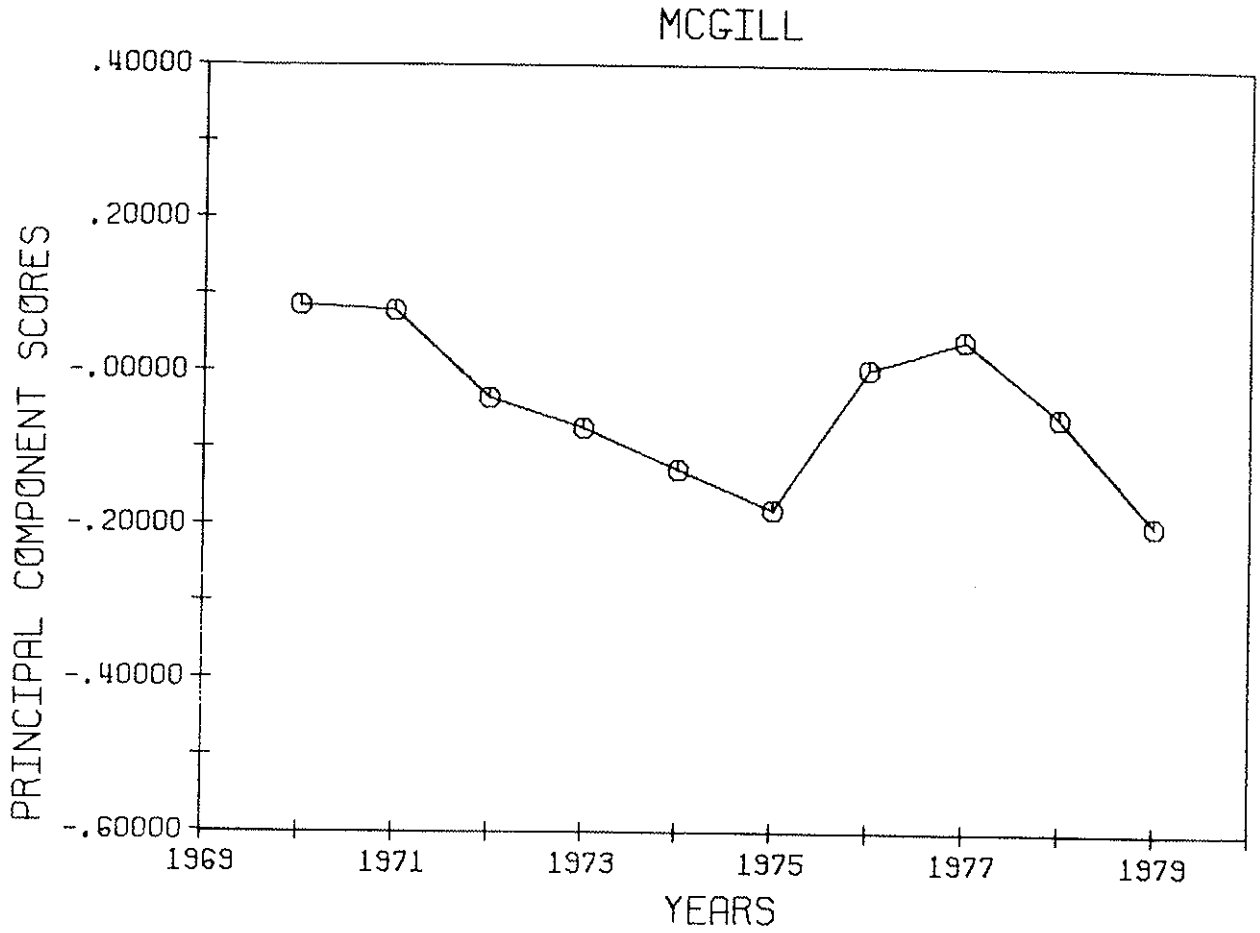

JOHNS HOPKINS
PRINCIPAL COMPONENT SCORES
-.100
-.300
-.500
-.700
-.900
-1.100
1969
1971
1973
1975
1977
1979
YEARS

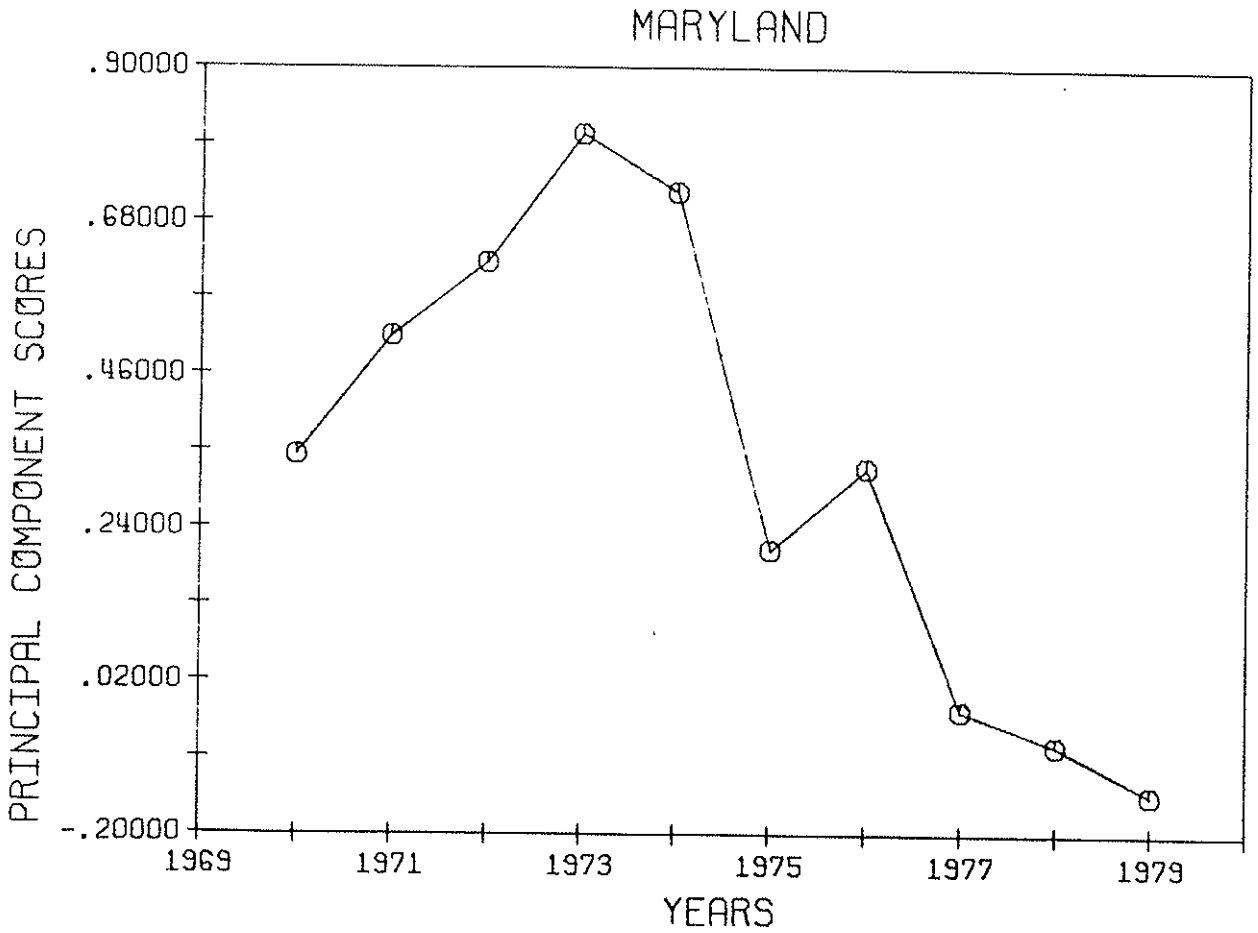

JOINT UNIVERSITY
PRINCIPAL COMPONENT SCORES
-.300
-.500
-.700
-.900
-1.100
-1.300
1969
1971
1973
1975
1977
1979
YEARS

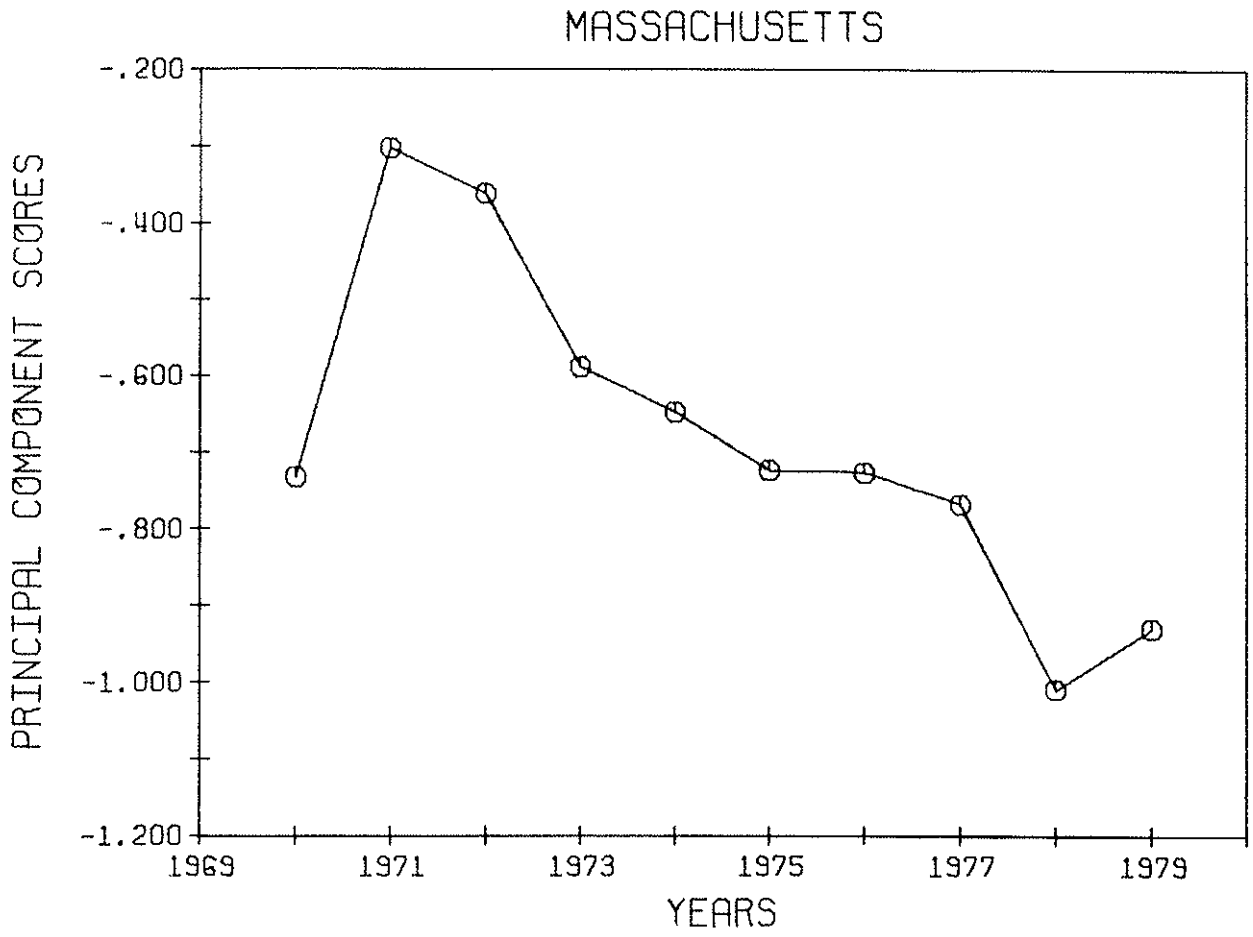

KANSAS
PRINCIPAL COMPONENT SCORES
.10000
-.10000
-.30000
-.50000
-.70000
-.90000
1969
1971
1973
1975
1977
1979
YEARS

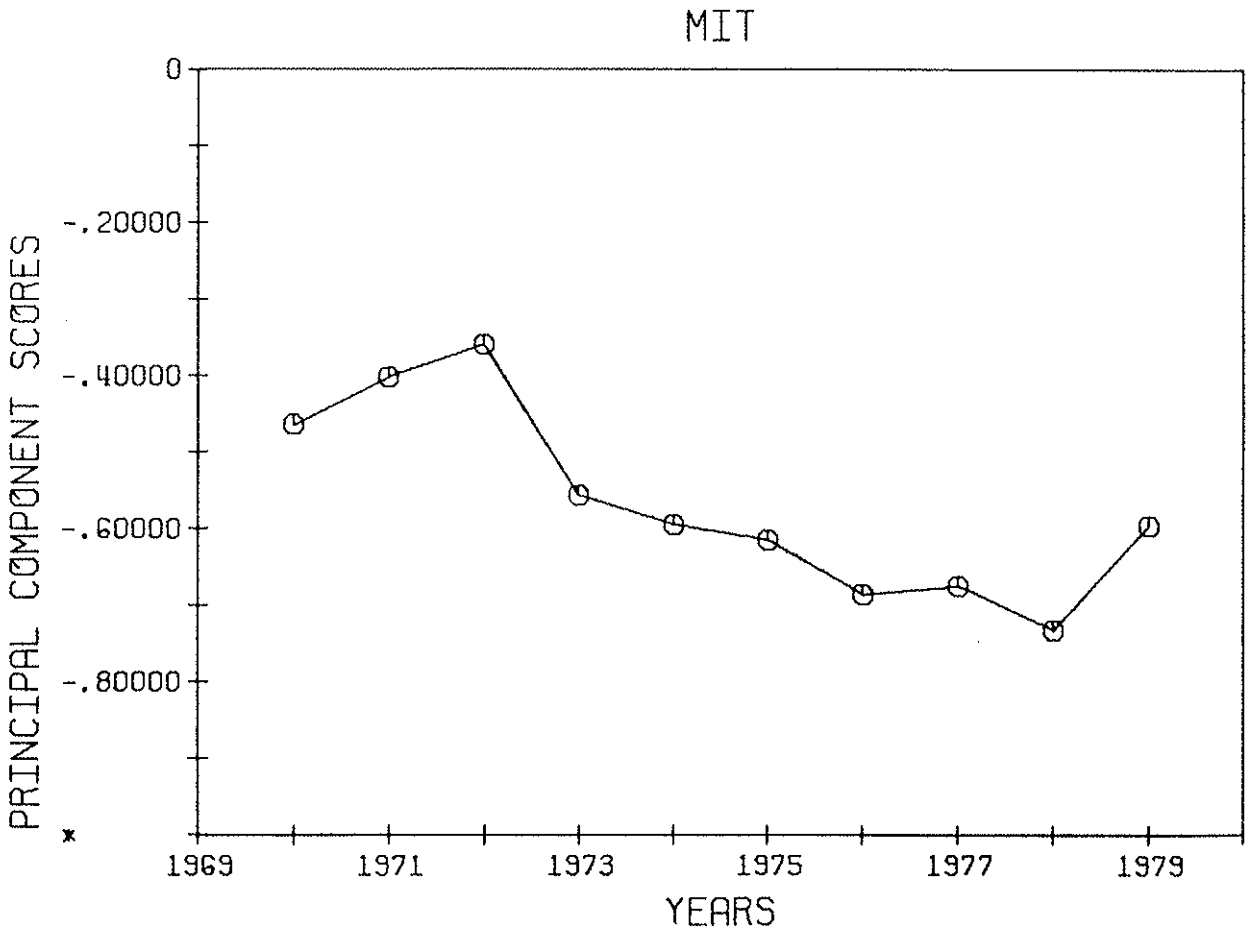

KENTUCKY
.20000
-.00000
-.20000
-.40000
-.60000
-.80000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

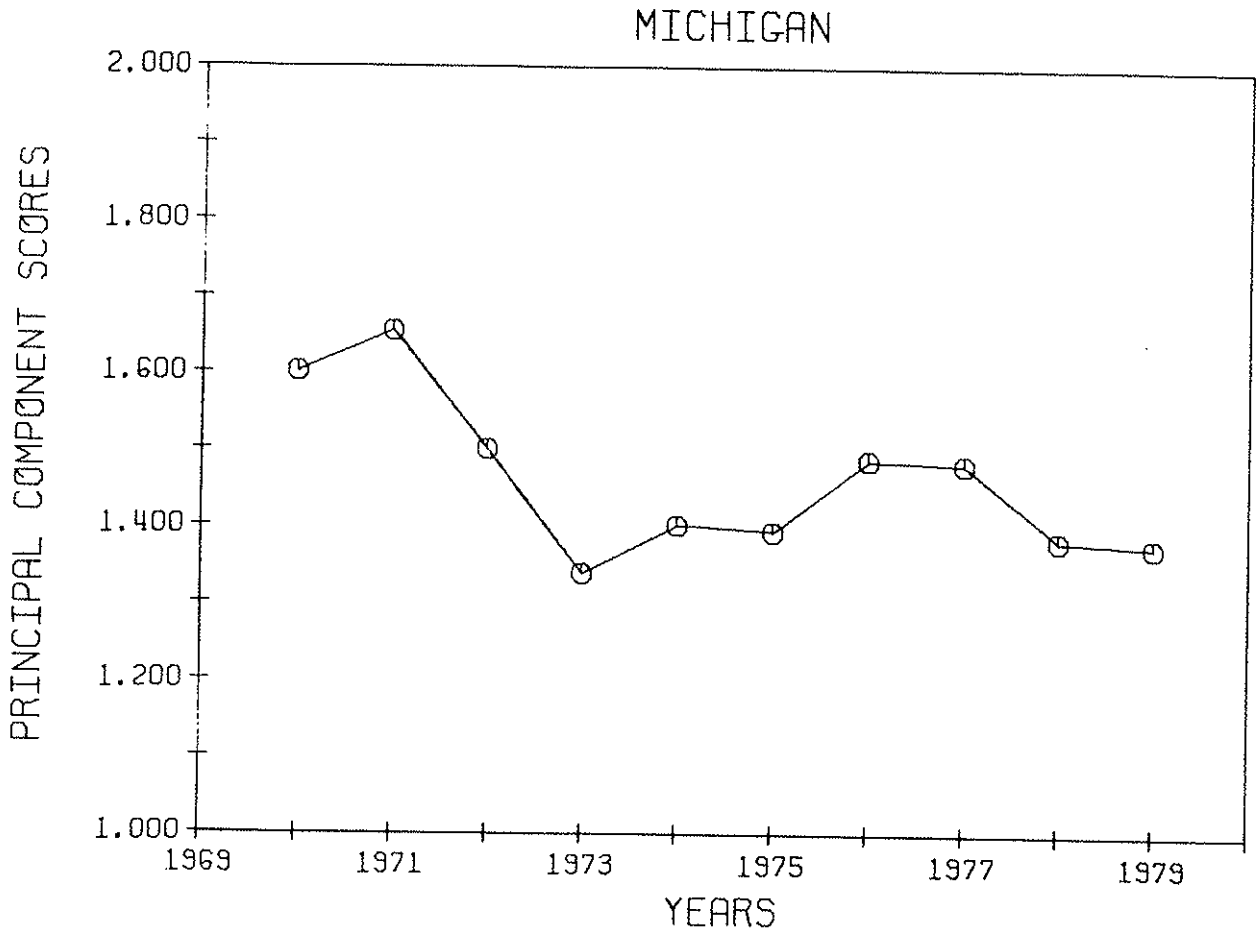

LOUISIANA STATE
-.500
-.700
-.900
-1.100
-1.300
-1.500
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

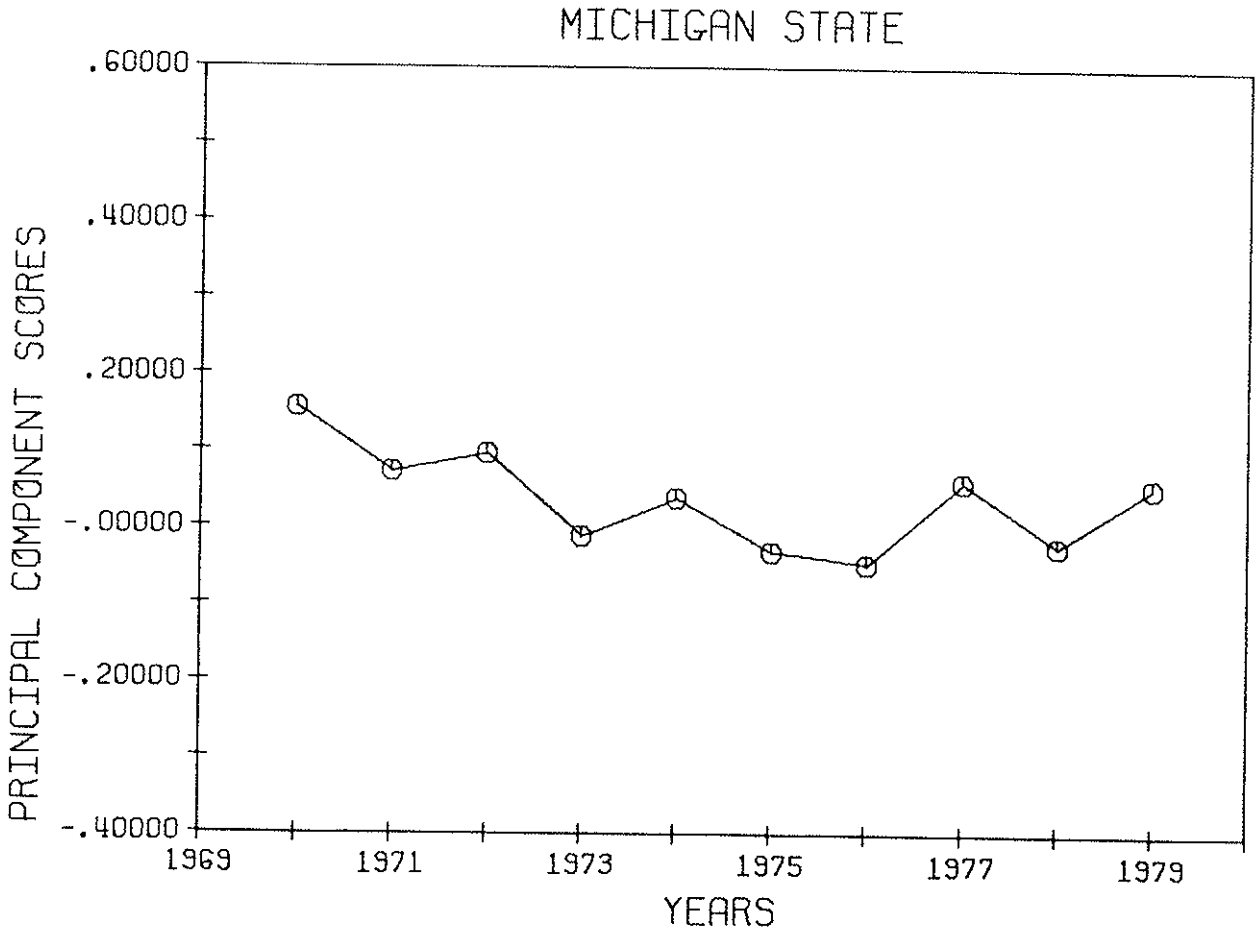

MCGILL
.40000
.20000
-.00000
-.20000
-.40000
-.60000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

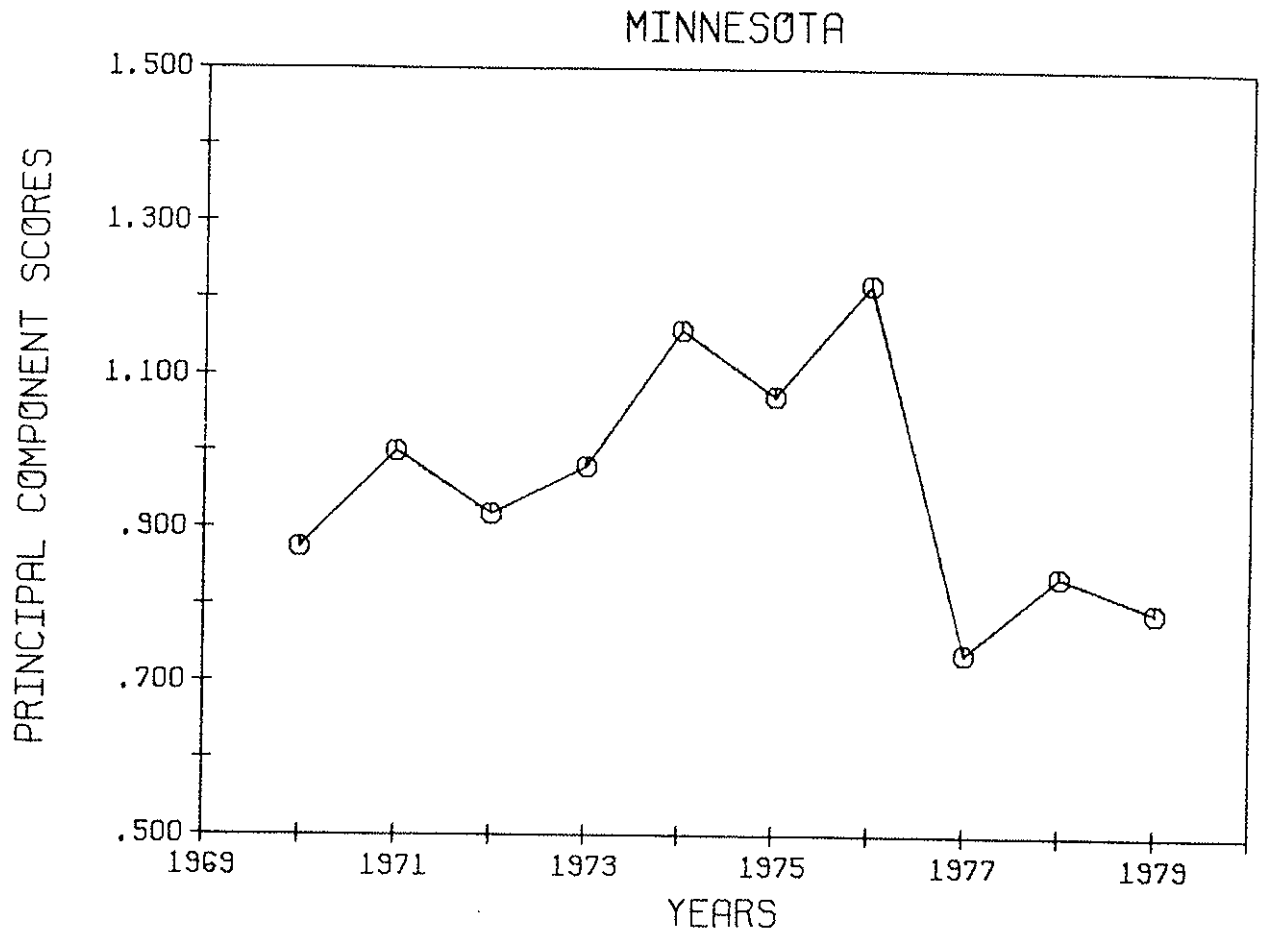

MARYLAND
.90000
.68000
.46000
.24000
.02000
-.20000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

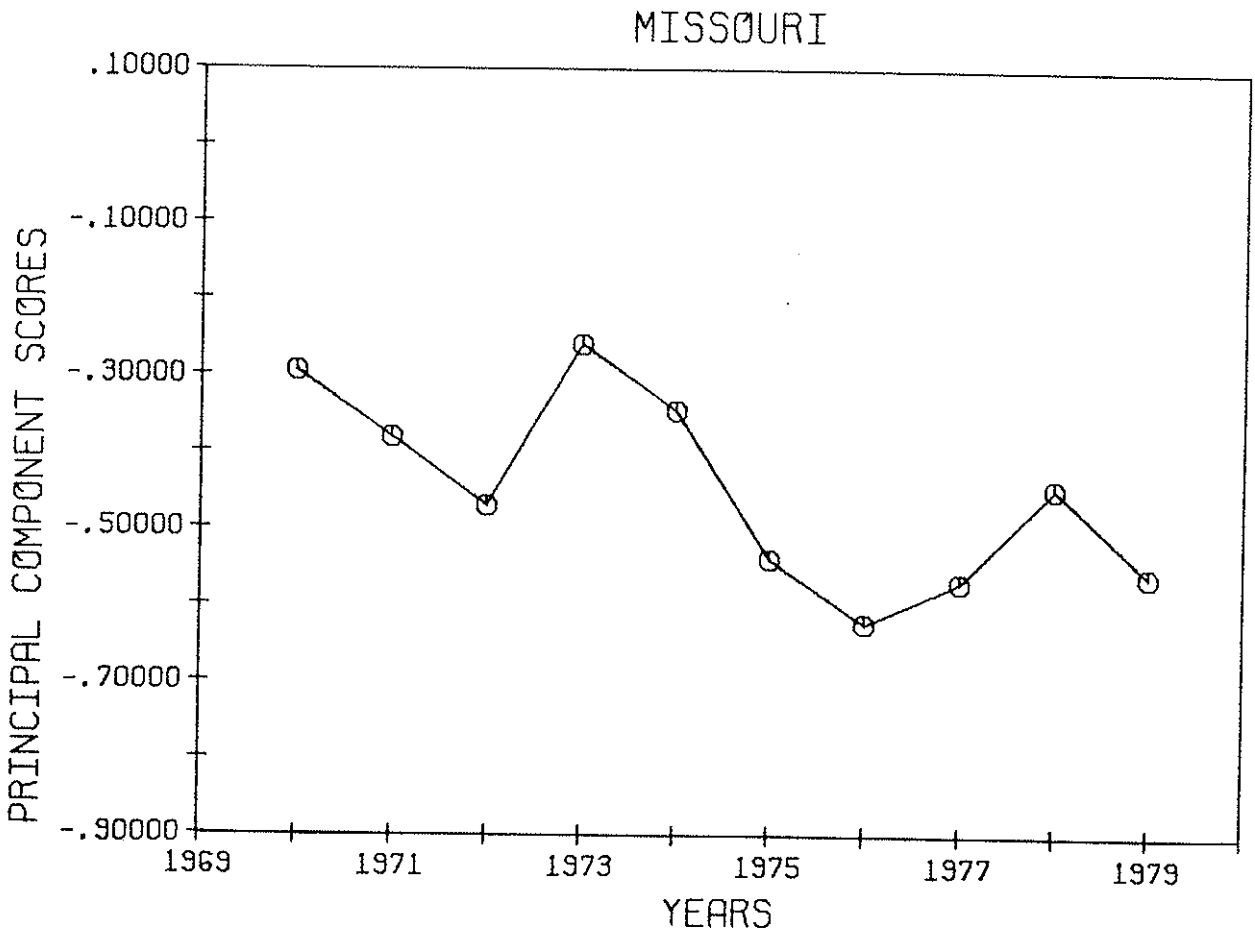

MASSACHUSETTS
-.200
-.400
-.600
-.800
-1.000
-1.200
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

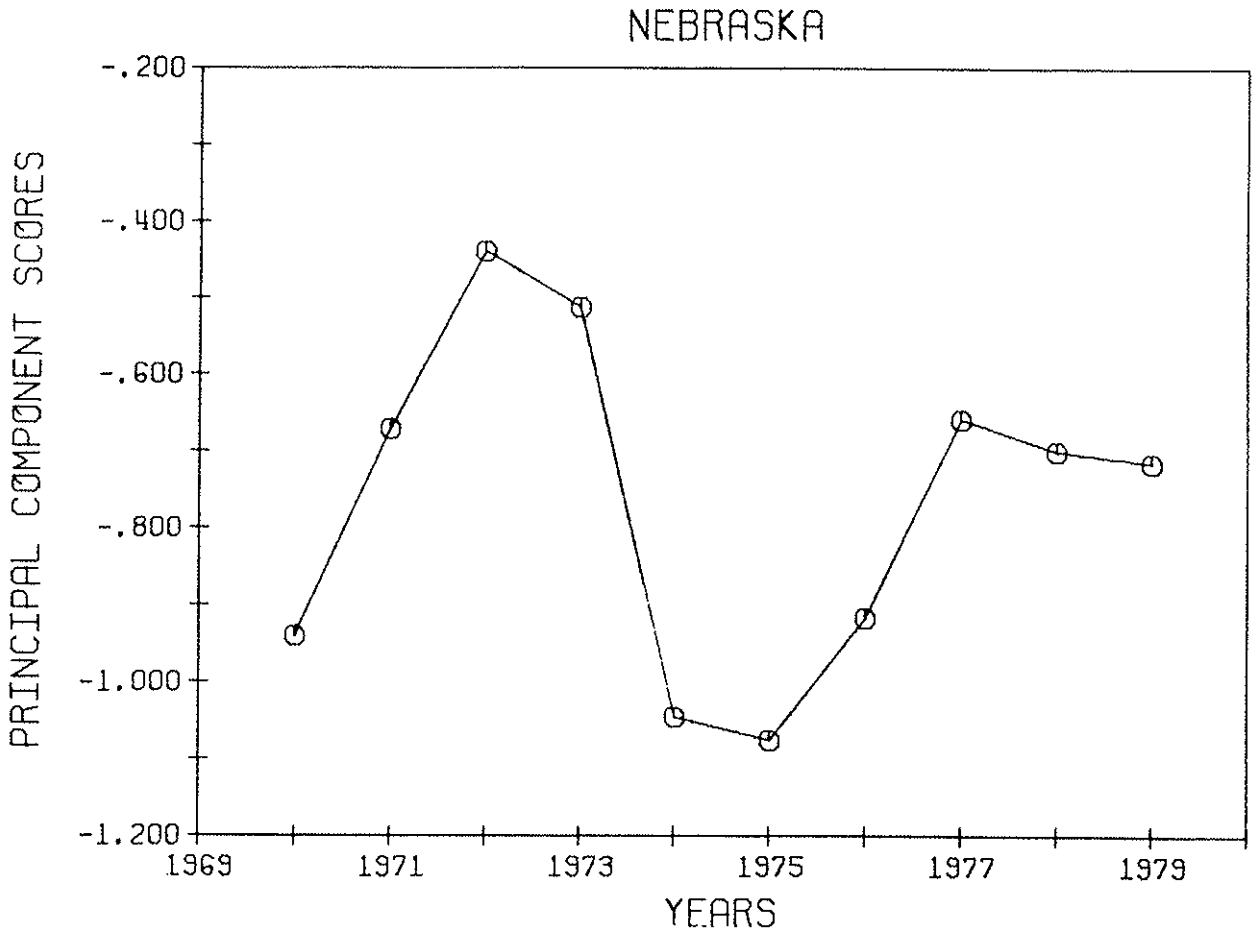

MIT
0
-.20000
-.40000
-.60000
-.80000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

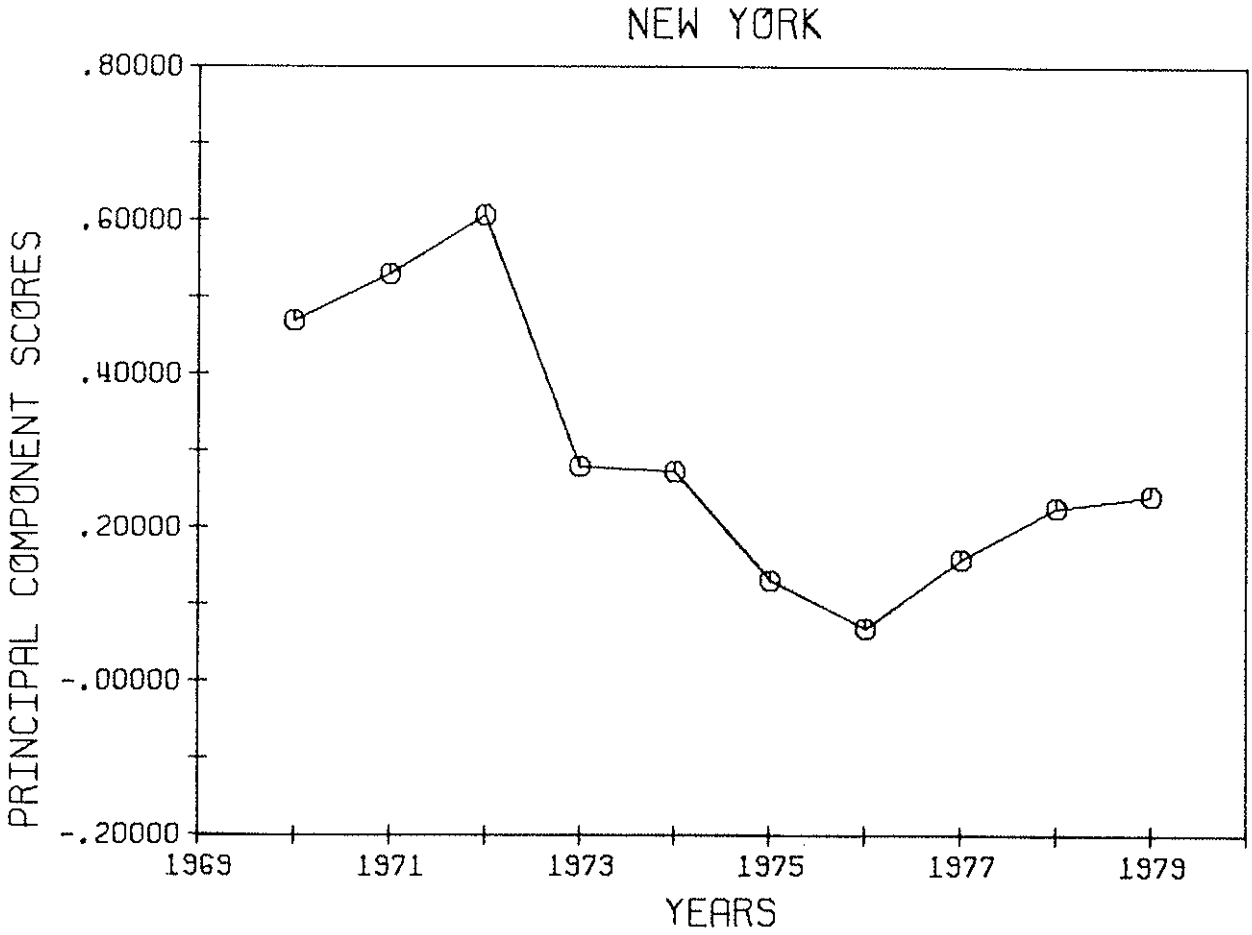

MICHIGAN
2.000
1.800
1.600
1.400
1.200
1.000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

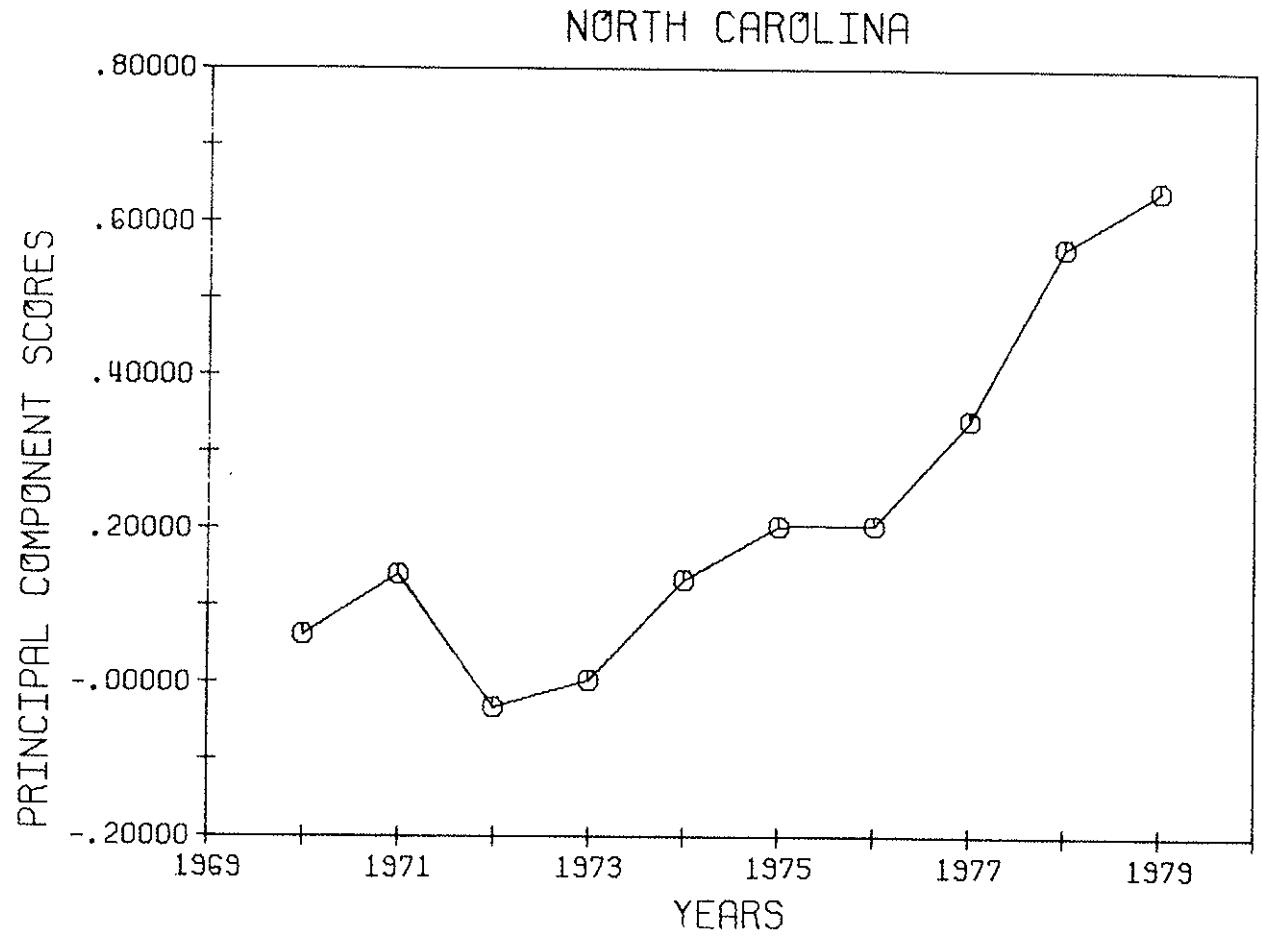

MICHIGAN STATE
.60000
.40000
.20000
-.00000
-.20000
-.40000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

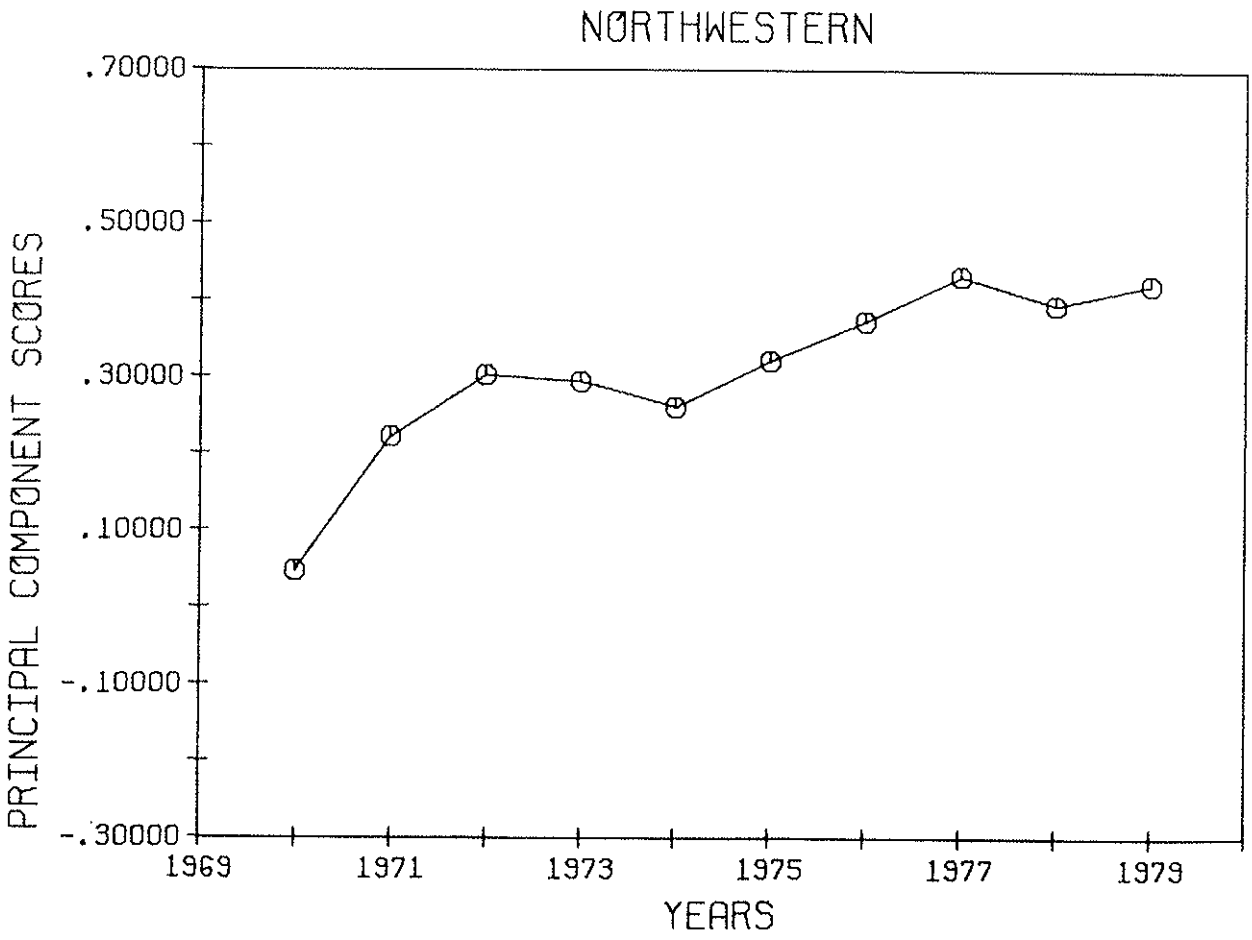

MINNESOTA
PRINCIPAL COMPONENT SCORES
1.500
1.300
1.100
.900
.700
.500
1969
1971
1973
1975
1977
1979
YEARS

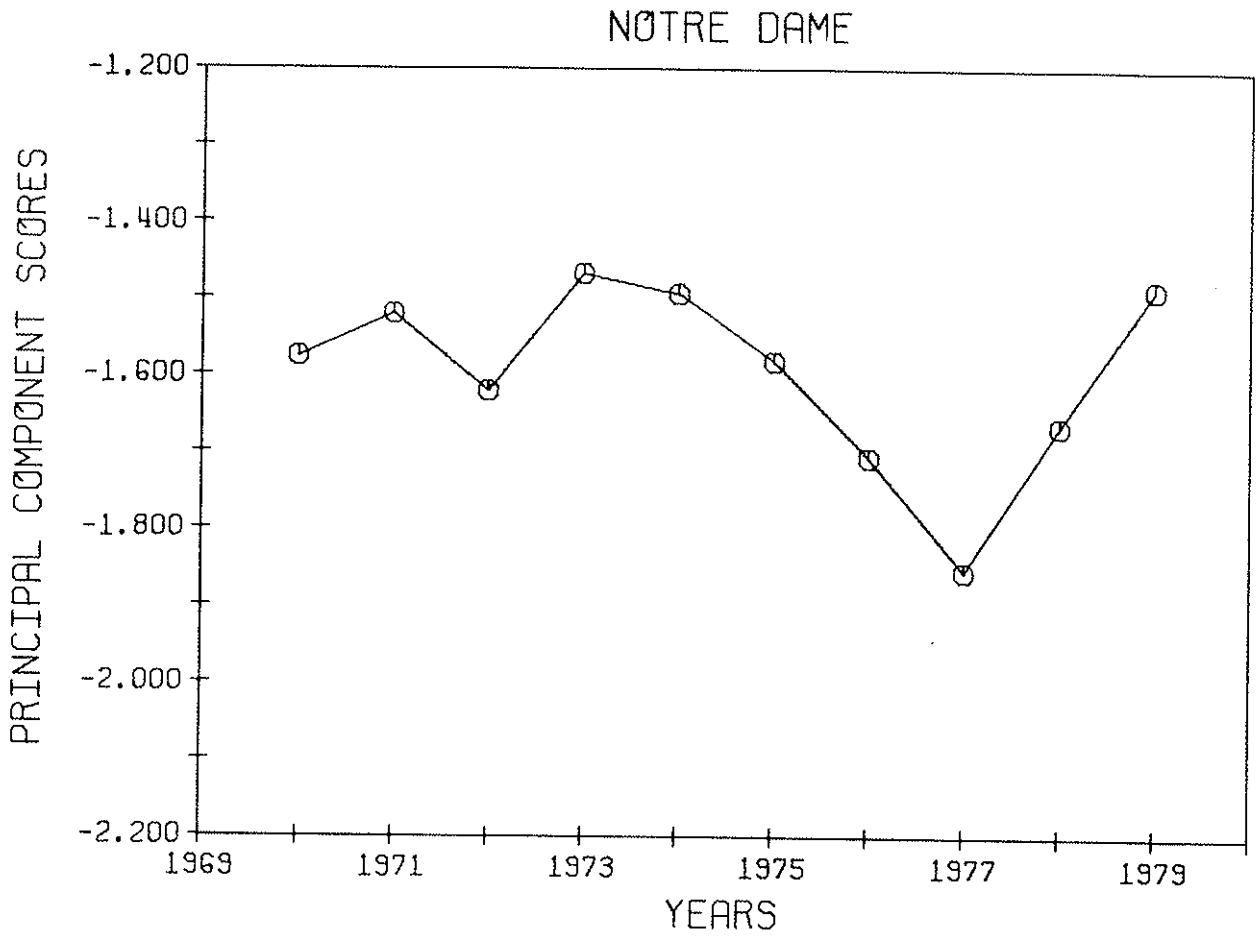

MISSOURI
PRINCIPAL COMPONENT SCORES
.10000
-.10000
-.30000
-.50000
-.70000
-.90000
1969
1971
1973
1975
1977
1979
YEARS

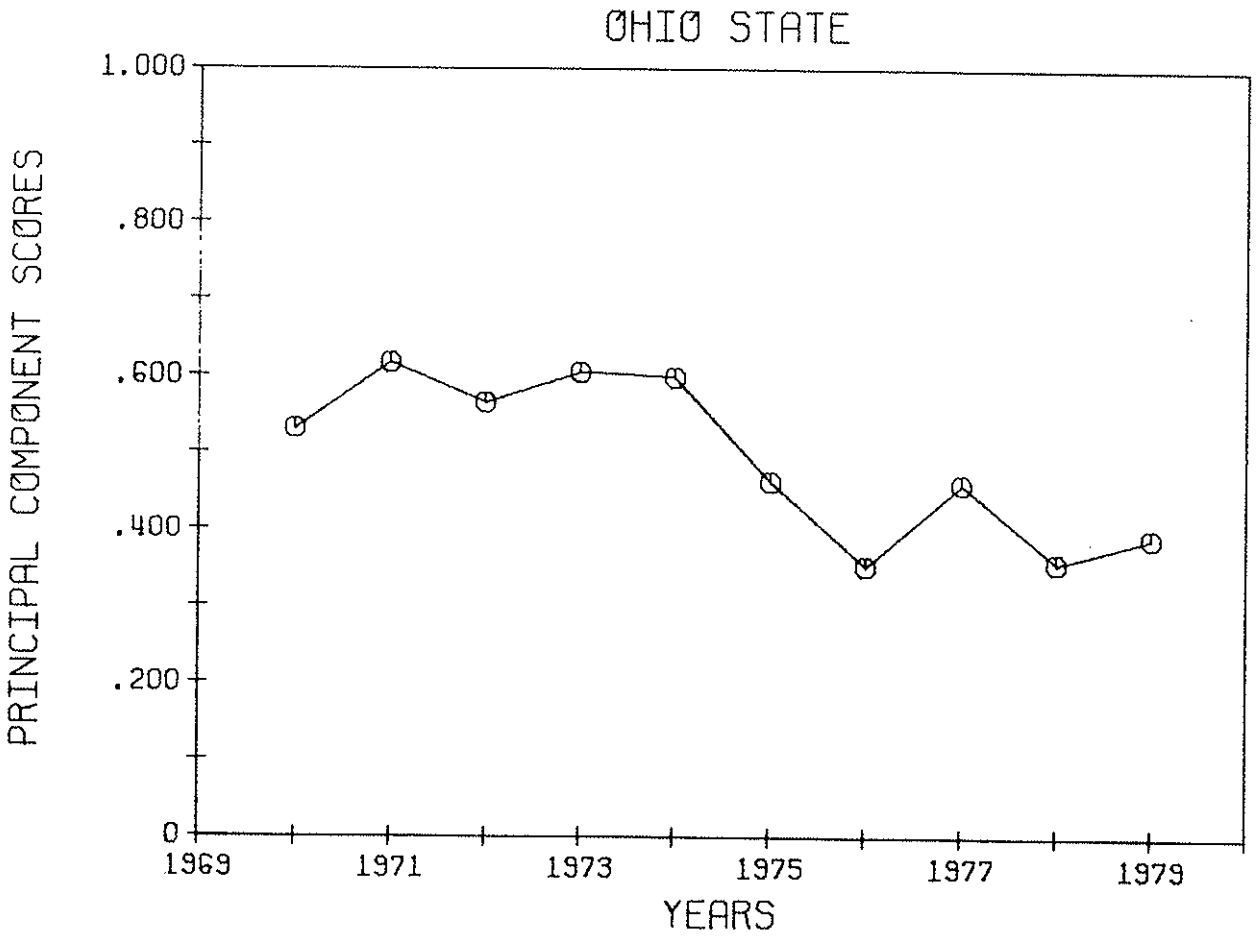

NEBRASKA
PRINCIPAL COMPONENT SCORES
-.200
-.400
-.600
-.800
-1.000
-1.200
1969
1971
1973
1975
1977
1979
YEARS

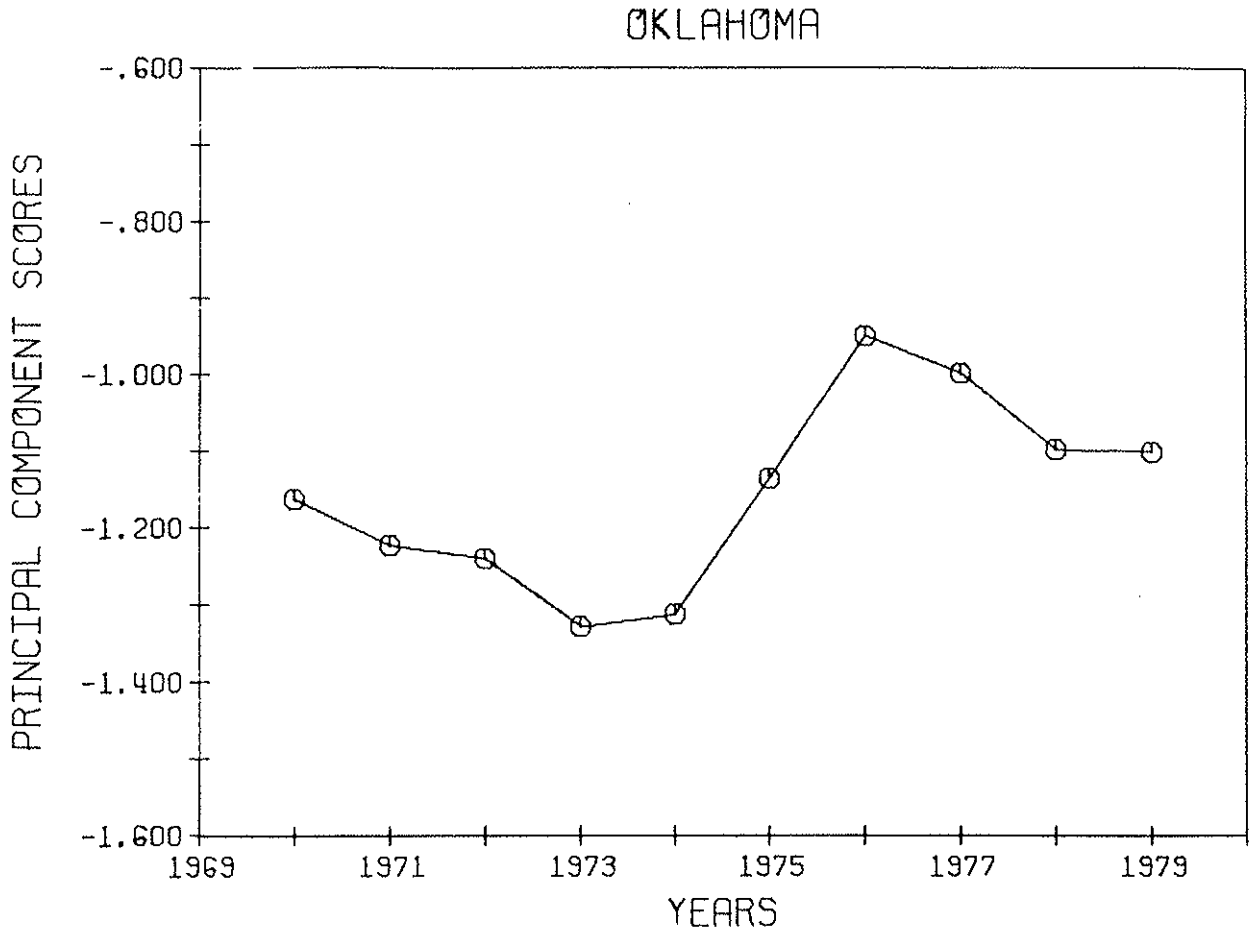

NEW YORK
PRINCIPAL COMPONENT SCORES
.80000
.60000
.40000
.20000
-.00000
-.20000
1969
1971
1973
1975
1977
1979
YEARS

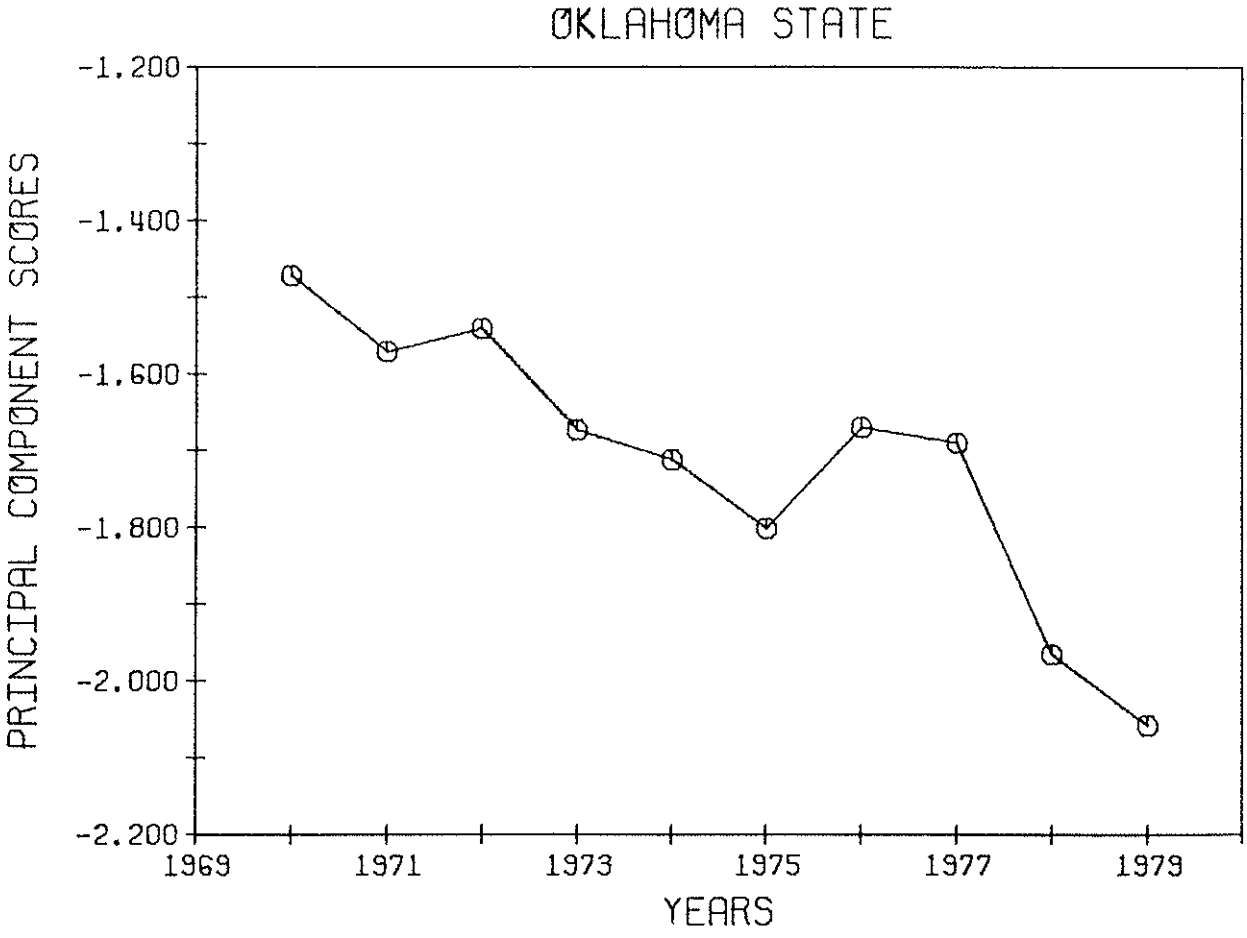

NORTH CAROLINA
.80000
.60000
.40000
.20000
-.00000
-.20000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

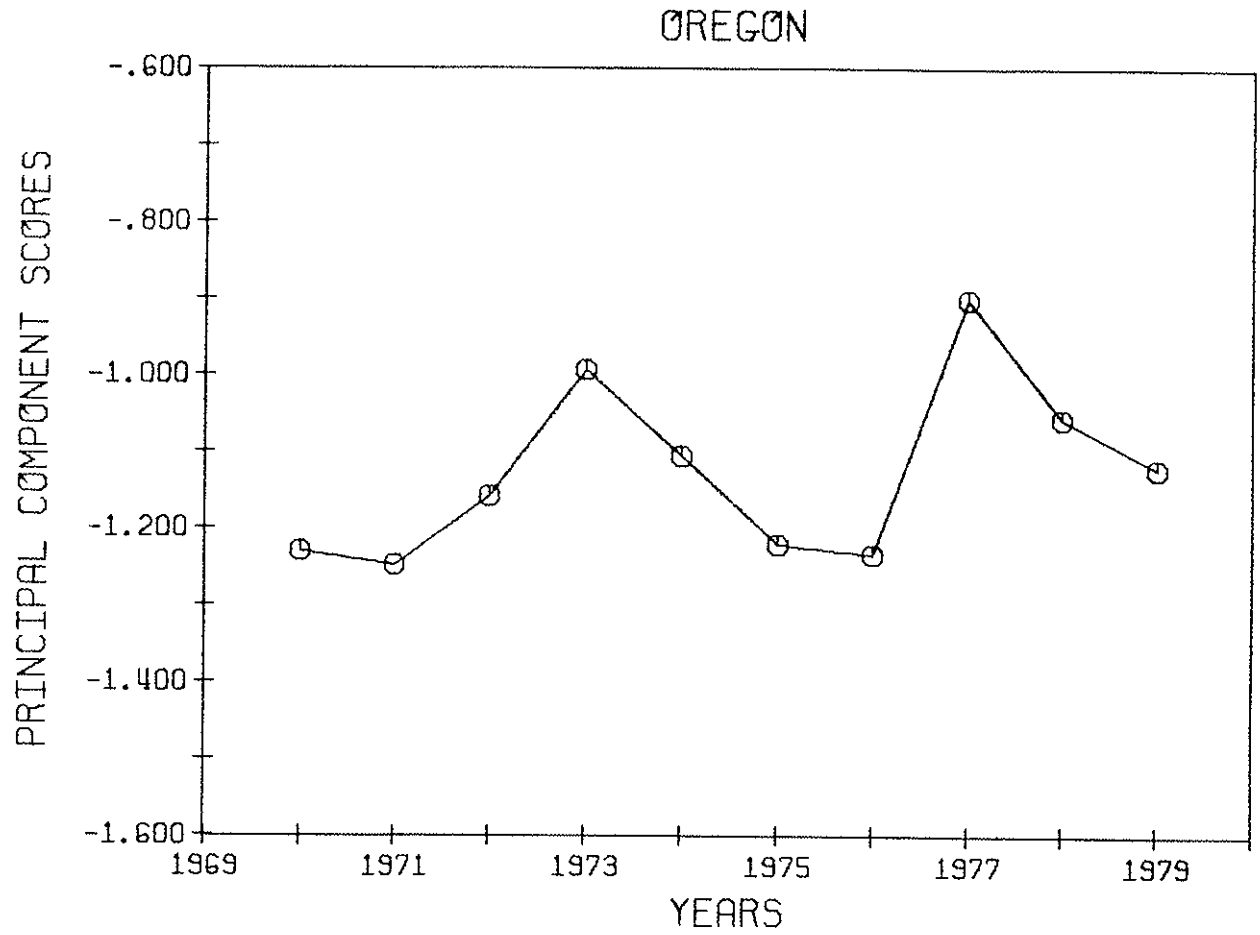

NORTHWESTERN
.70000
.50000
.30000
.10000
-.10000
-.30000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

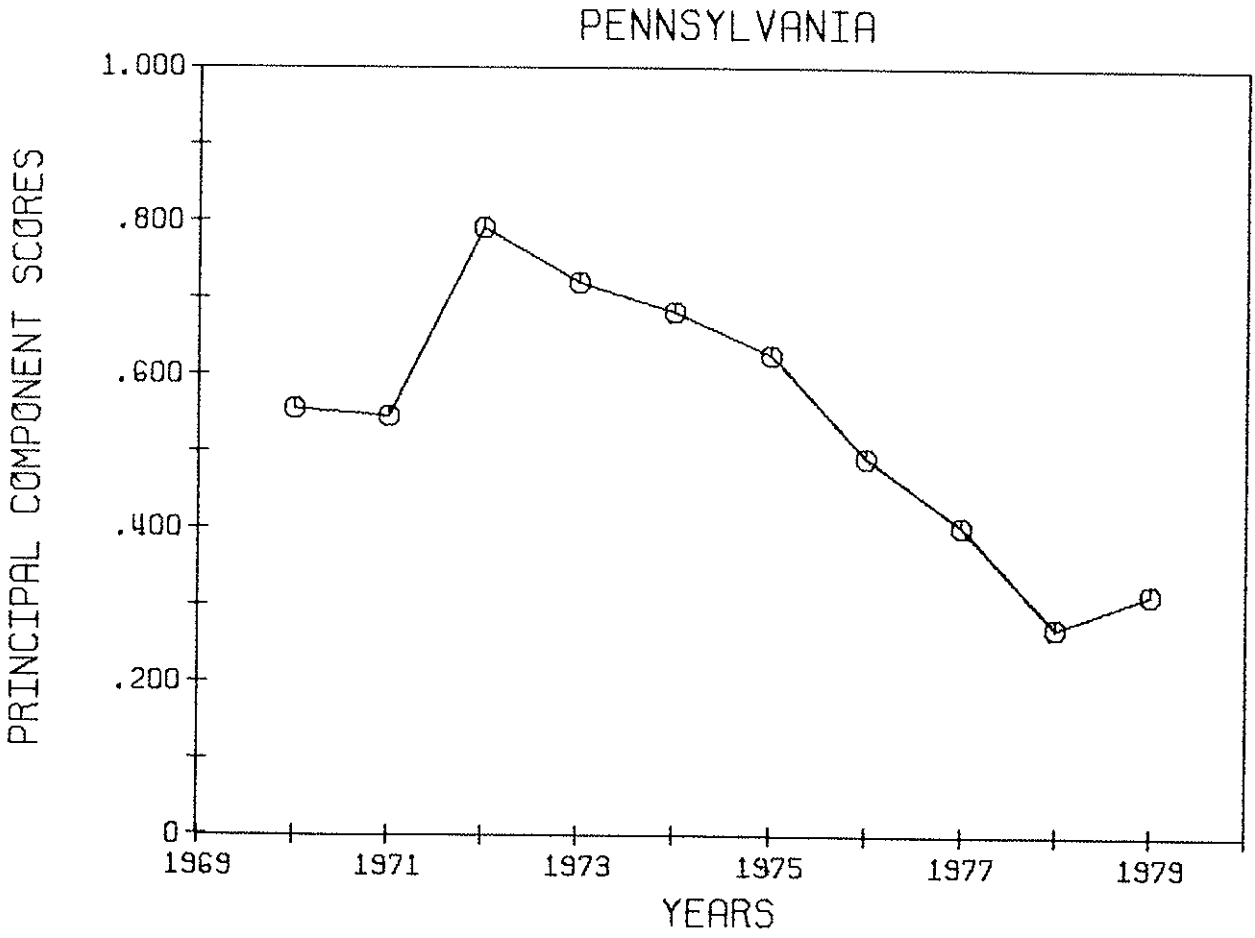

NØTRE DAME
PRINCIPAL CØMPØNENT SCØRES
-1.200
-1.400
-1.600
-1.800
-2.000
-2.200
1969
1971
1973
1975
1977
1979
YEARS

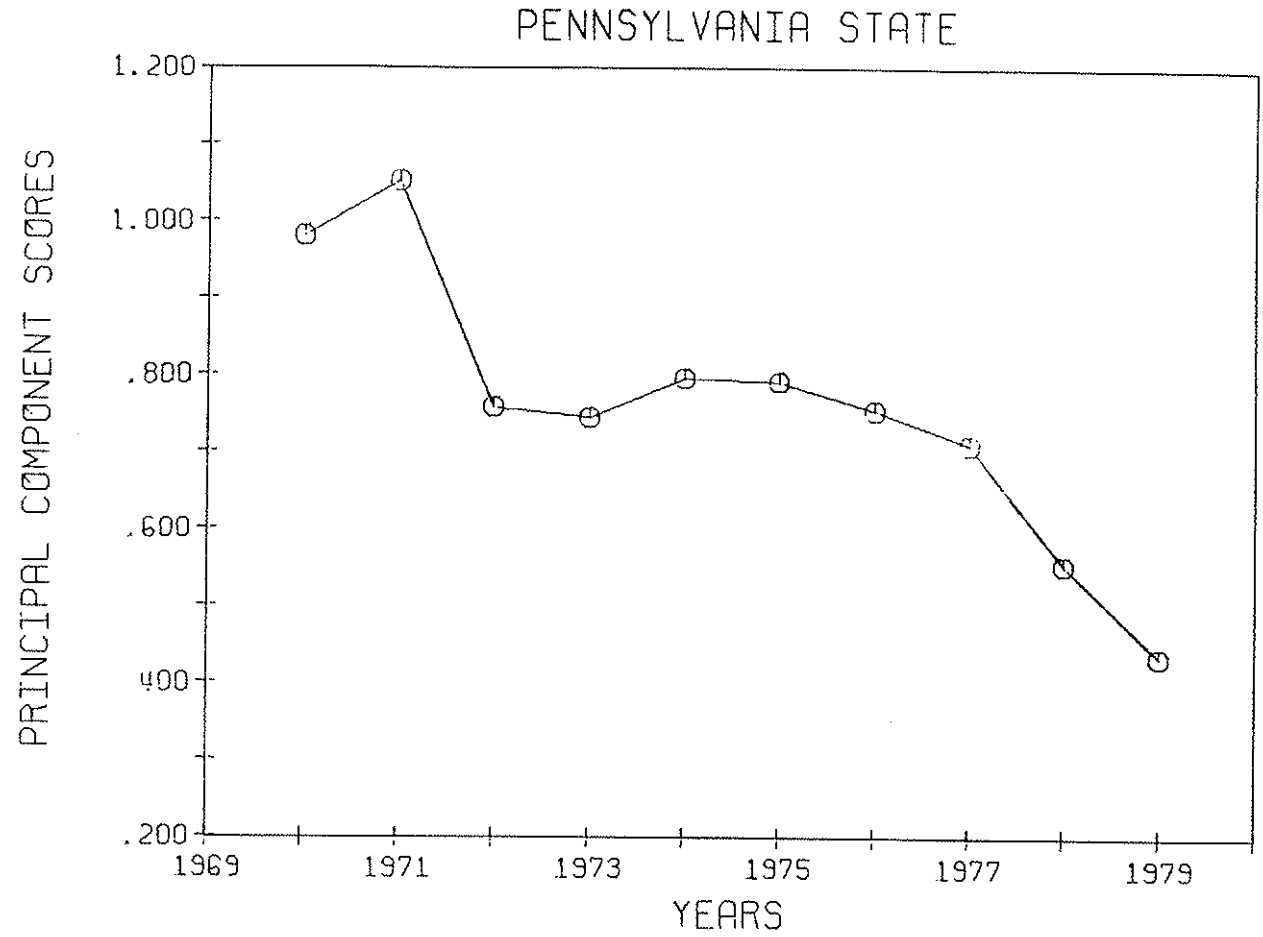

ØHIØ STATE
PRINCIPAL CØMPØNENT SCØRES
1.000
.800
.600
.400
.200
0
1969
1971
1973
1975
1977
1979
YEARS

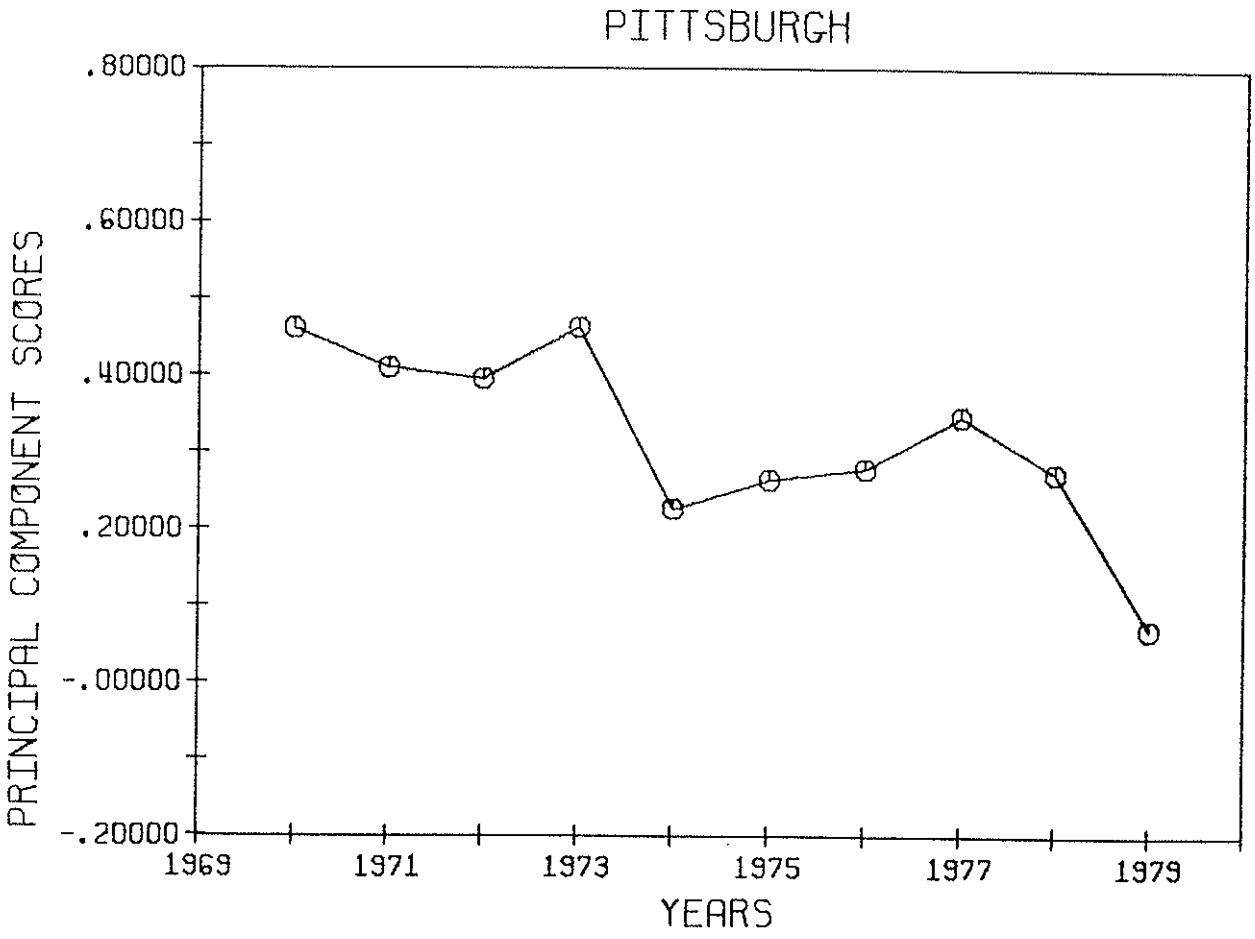

ØKLAHØMA
PRINCIPAL COMPONENT SCORES
-.600
-.800
-1.000
-1.200
-1.400
-1.600
1969
1971
1973
1975
1977
1979
YEARS

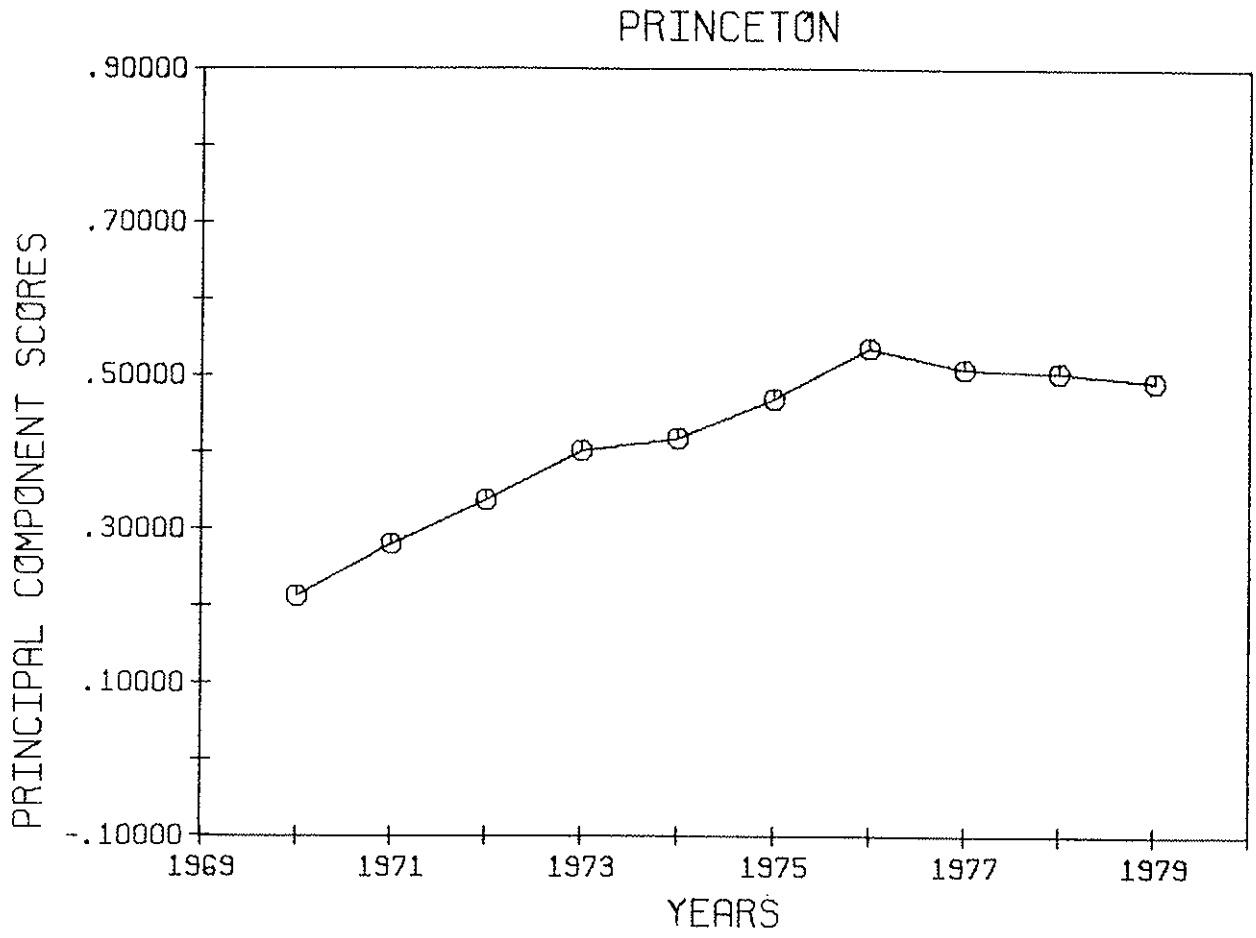

ØKLAHØMA STATE
PRINCIPAL COMPONENT SCORES
-1.200
-1.400
-1.600
-1.800
-2.000
-2.200
1969
1971
1973
1975
1977
1979
YEARS

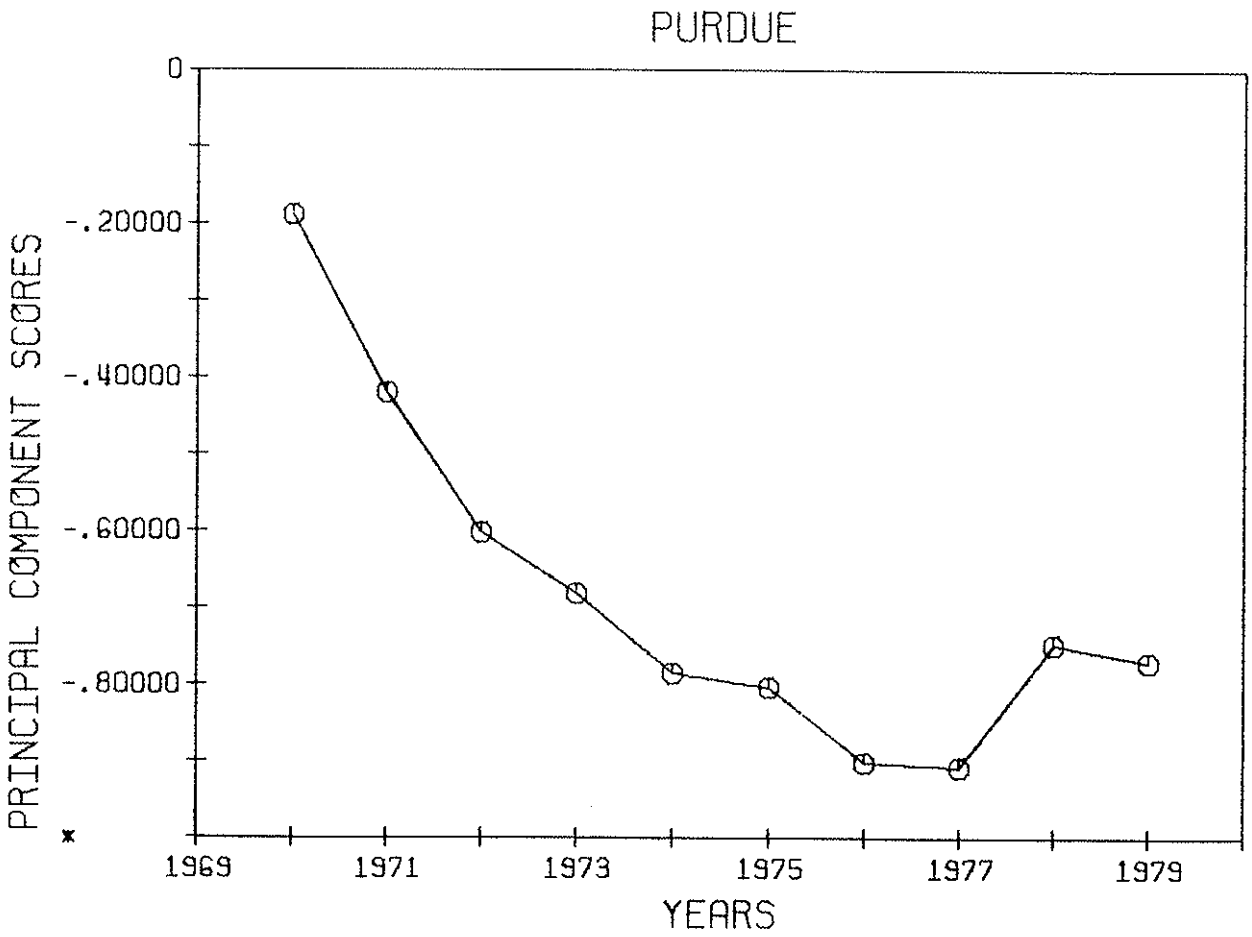

OREGON
-.600
-.800
-1.000
-1.200
-1.400
-1.600
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

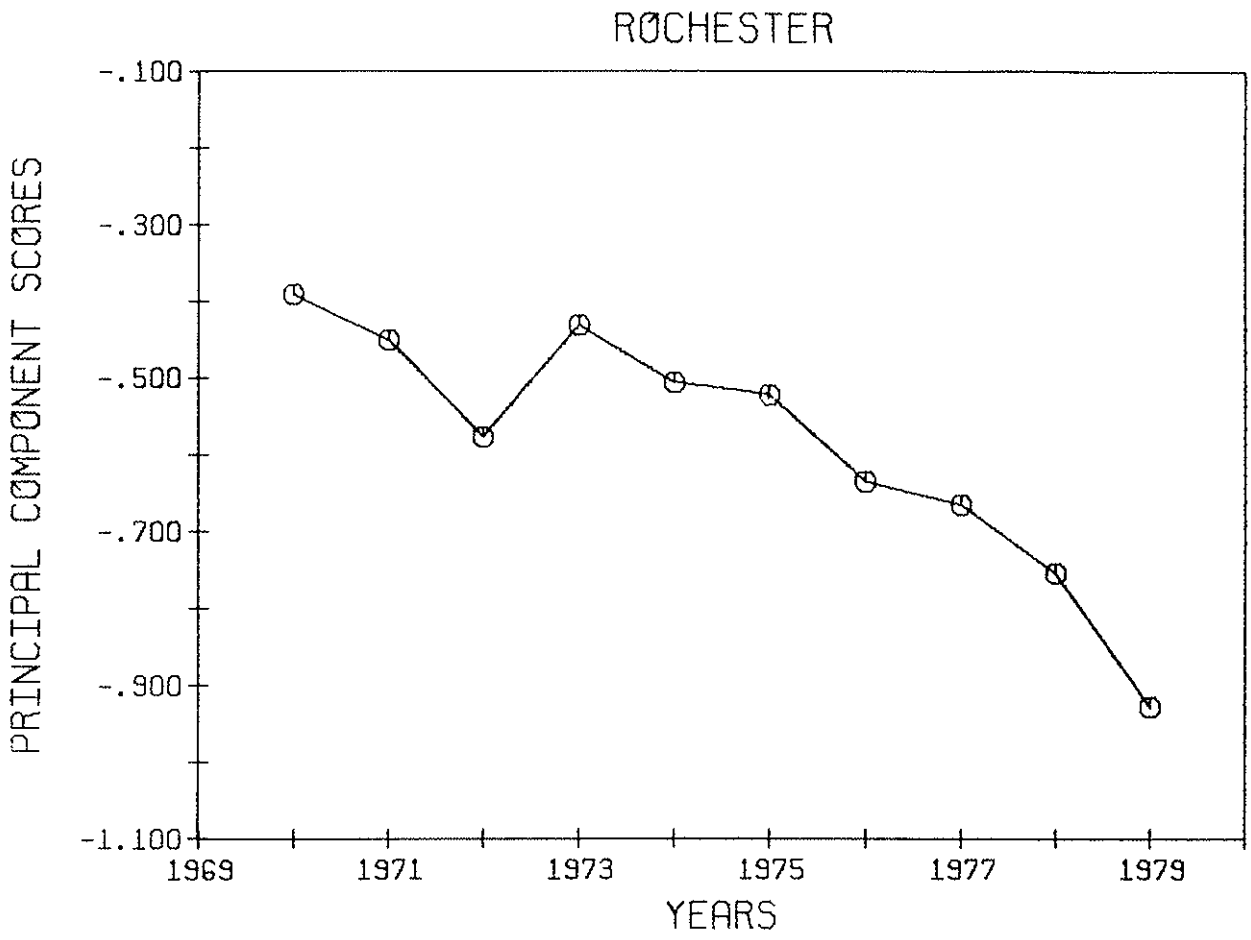

PENNSYLVANIA
1.000
.800
.600
.400
.200
0
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

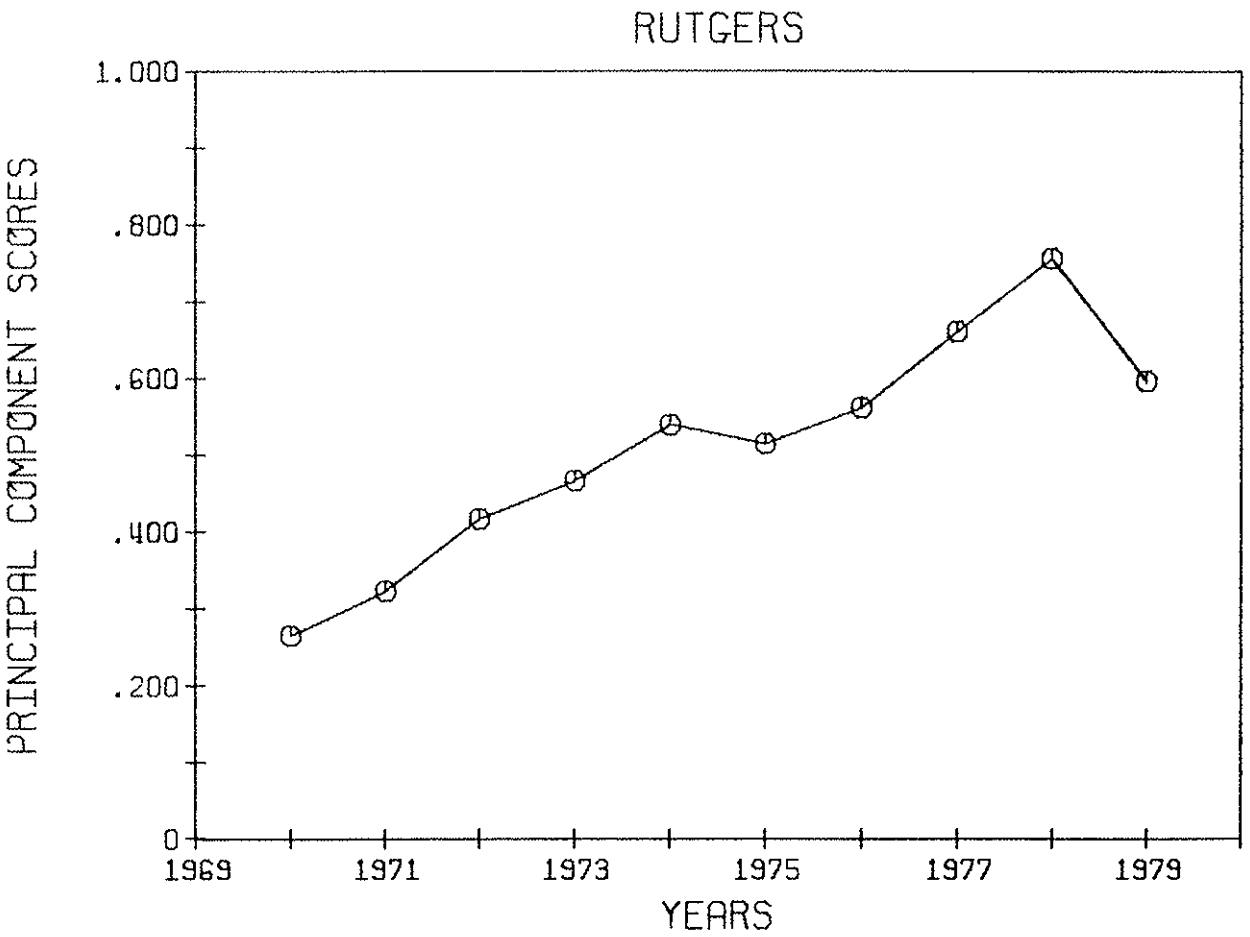

PENNSYLVANIA STATE
PRINCIPAL COMPONENT SCORES
1.200
1.000
.800
.600
.400
.200
1969
1971
1973
1975
1977
1979
YEARS

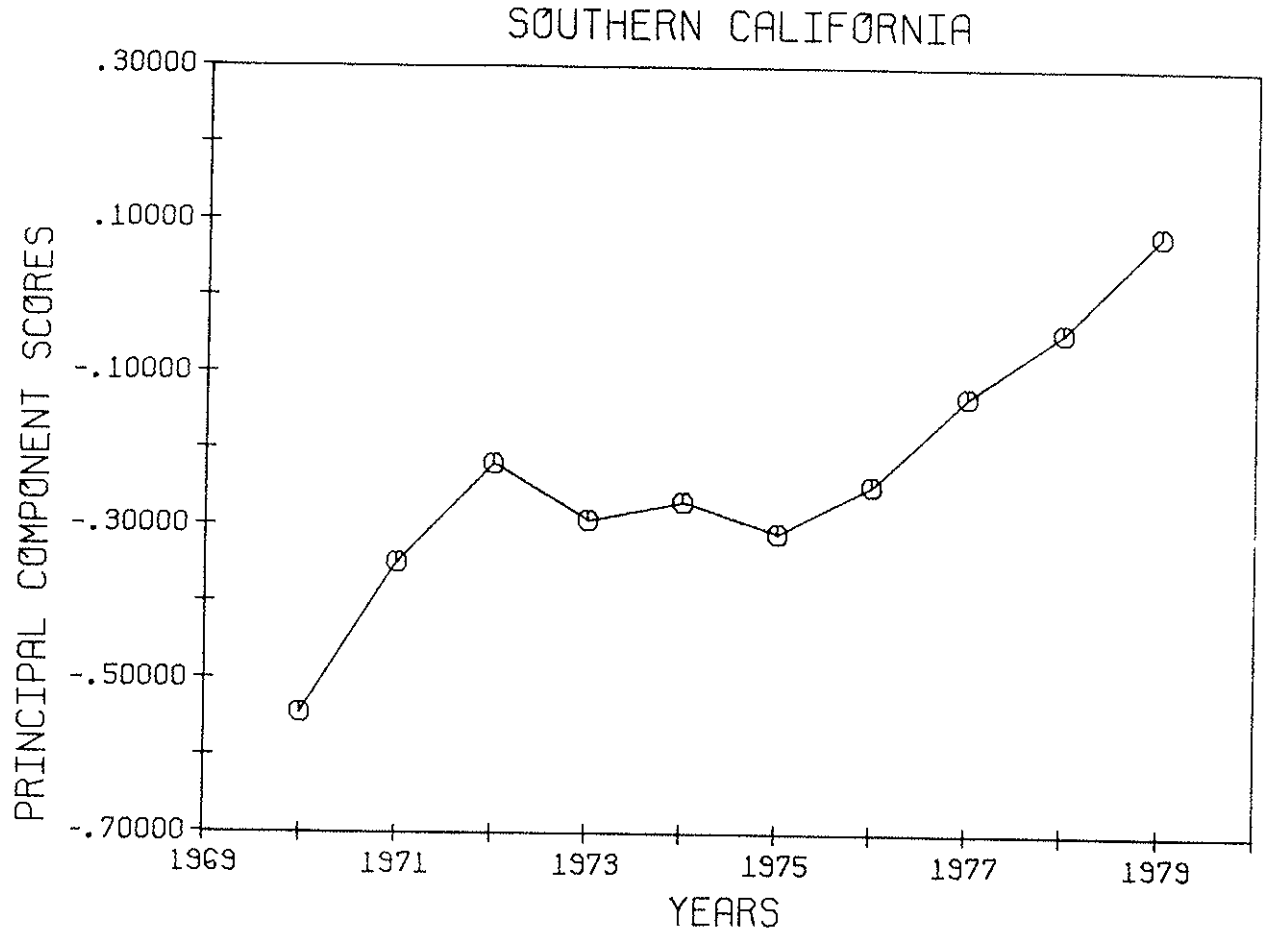

PITTSBURGH
PRINCIPAL COMPONENT SCORES
.80000
.60000
.40000
.20000
-.00000
-.20000
1969
1971
1973
1975
1977
1979
YEARS

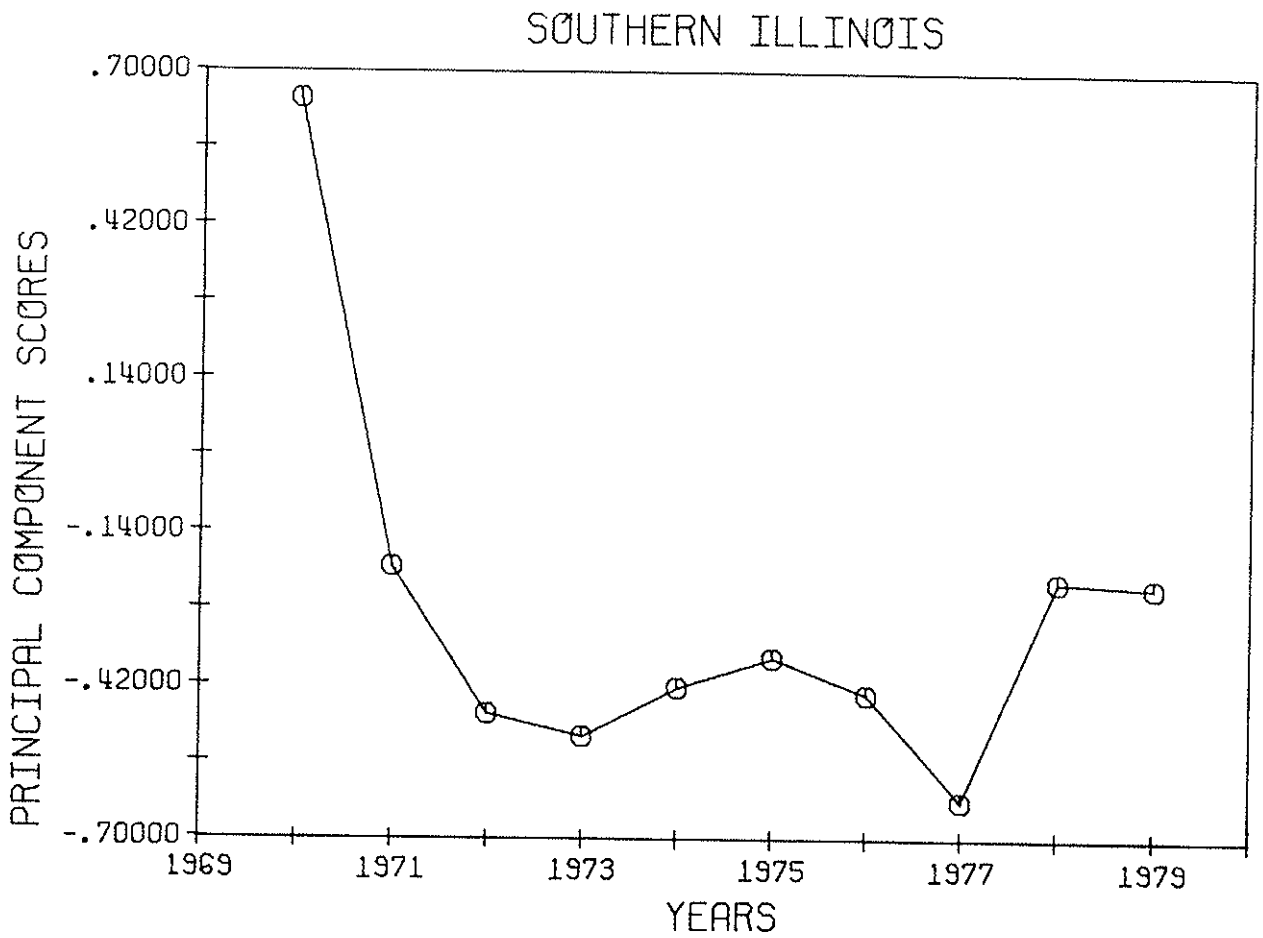

PRINCETON
.90000
.70000
.50000
.30000
.10000
-.10000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

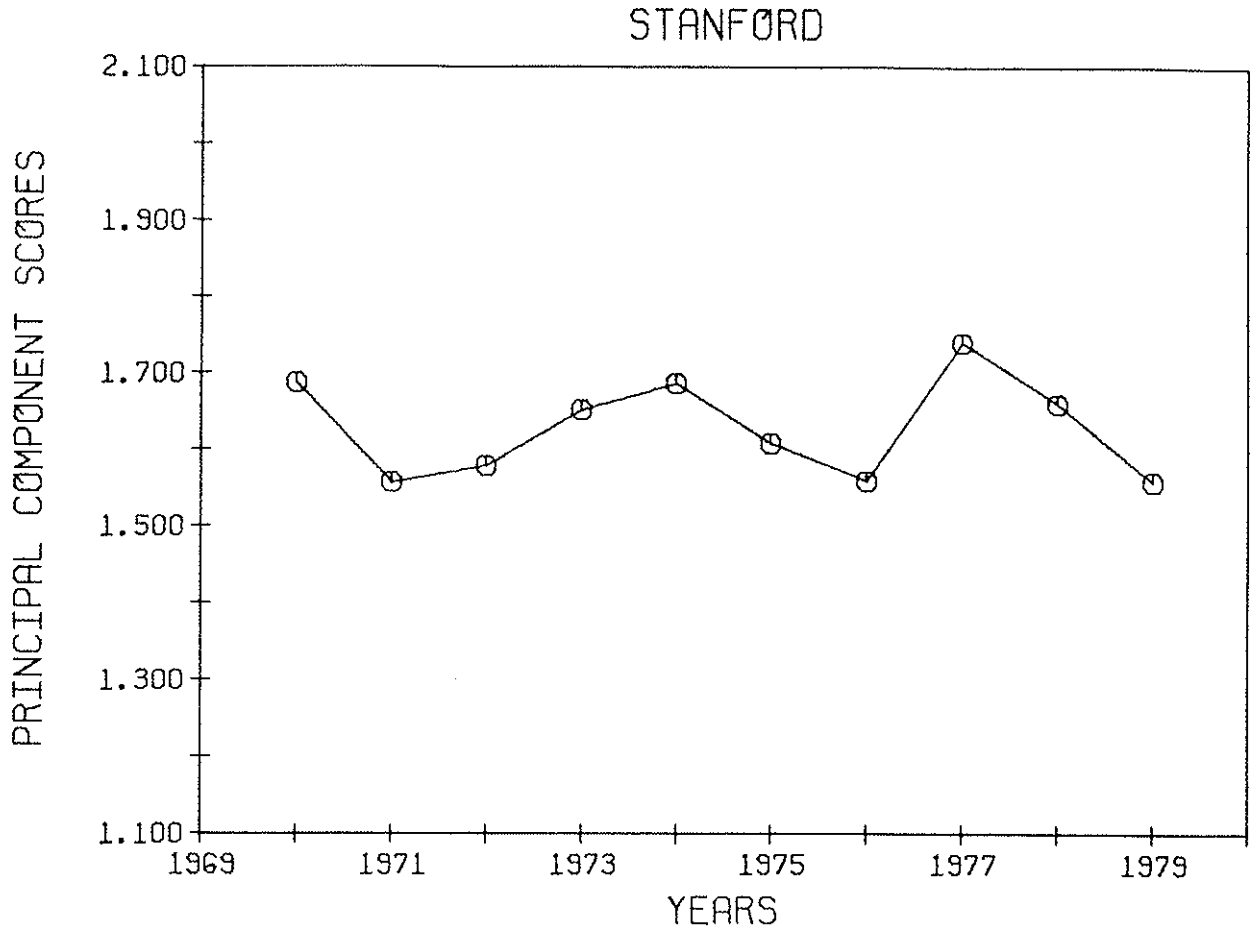

PURDUE
0
-.20000
-.40000
-.60000
-.80000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

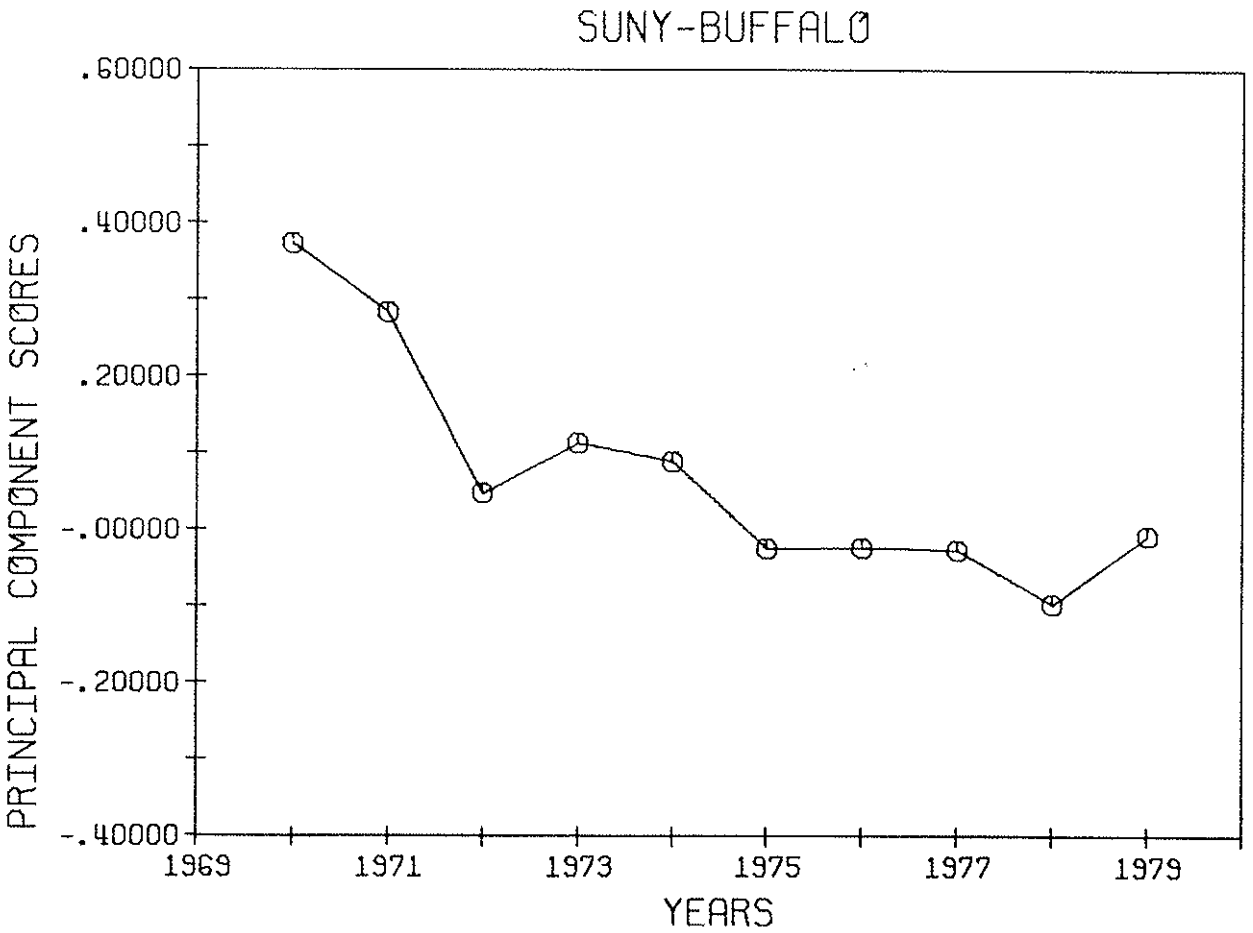

ROCHESTER
PRINCIPAL COMPONENT SCORES
-.100
-.300
-.500
-.700
-.900
-1.100
1969
1971
1973
1975
1977
1979
YEARS

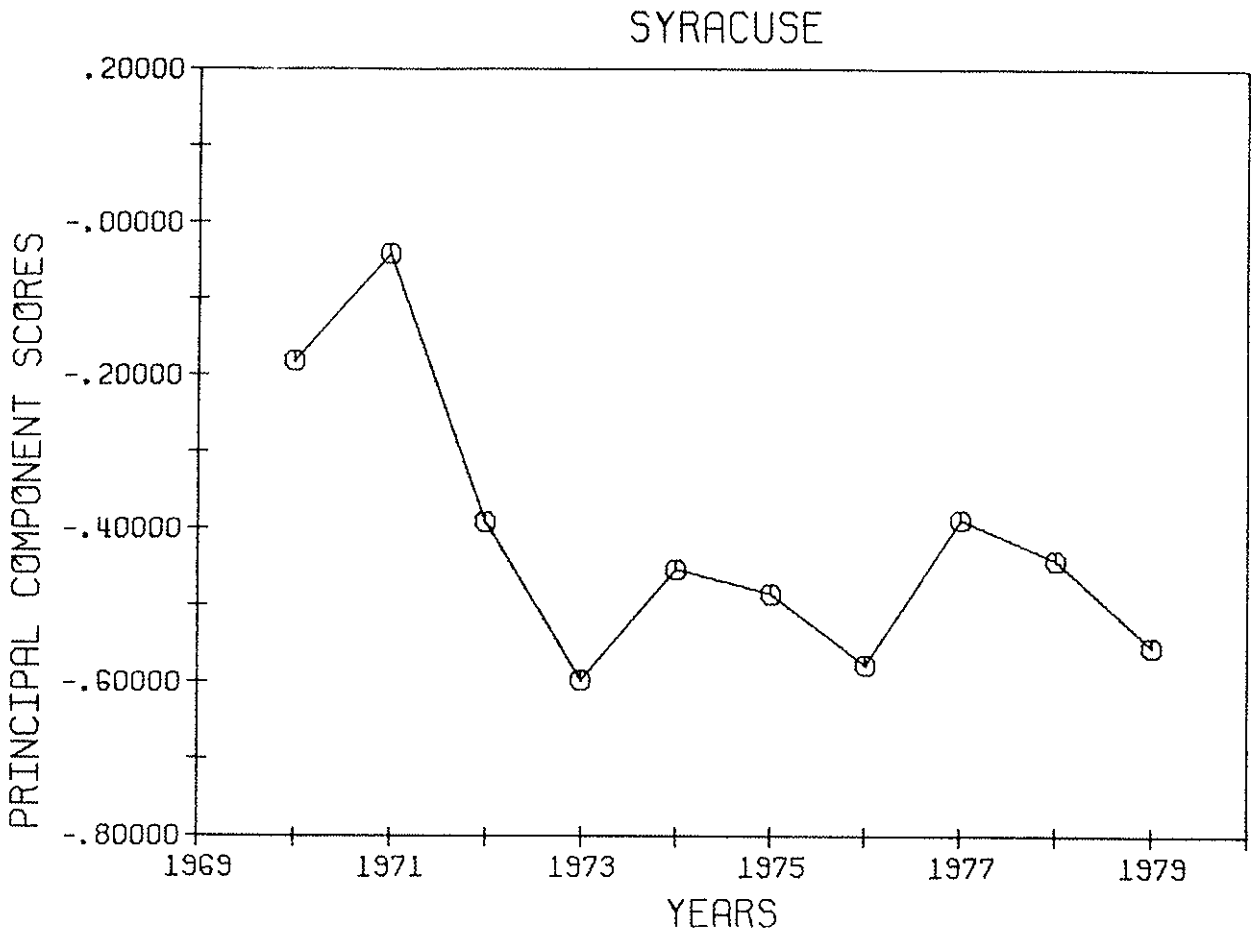

RUTGERS
PRINCIPAL COMPONENT SCORES
1.000
.800
.600
.400
.200
0
1969
1971
1973
1975
1977
1979
YEARS

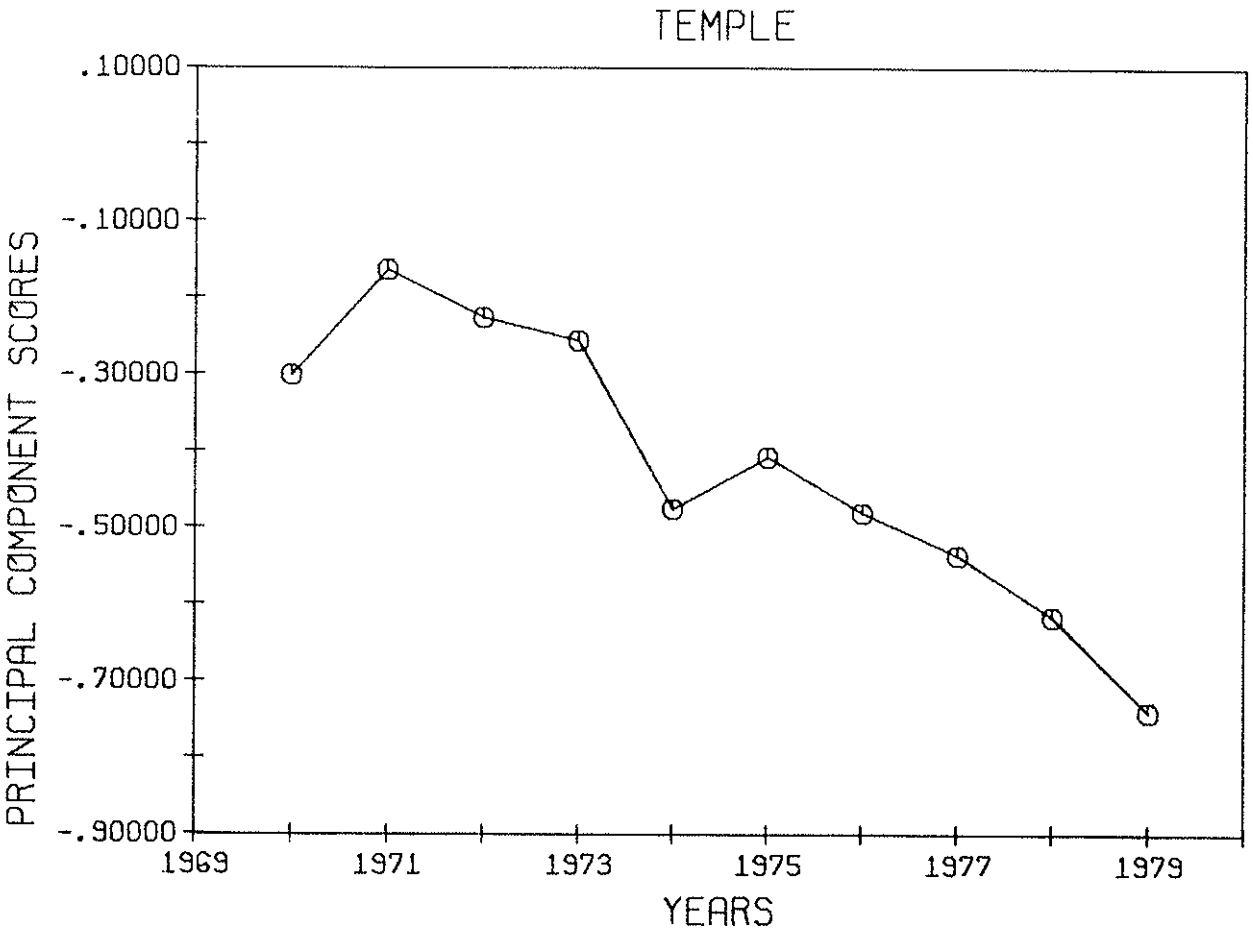

SOUTHERN CALIFORNIA
.30000
.10000
-.10000
-.30000
-.50000
-.70000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

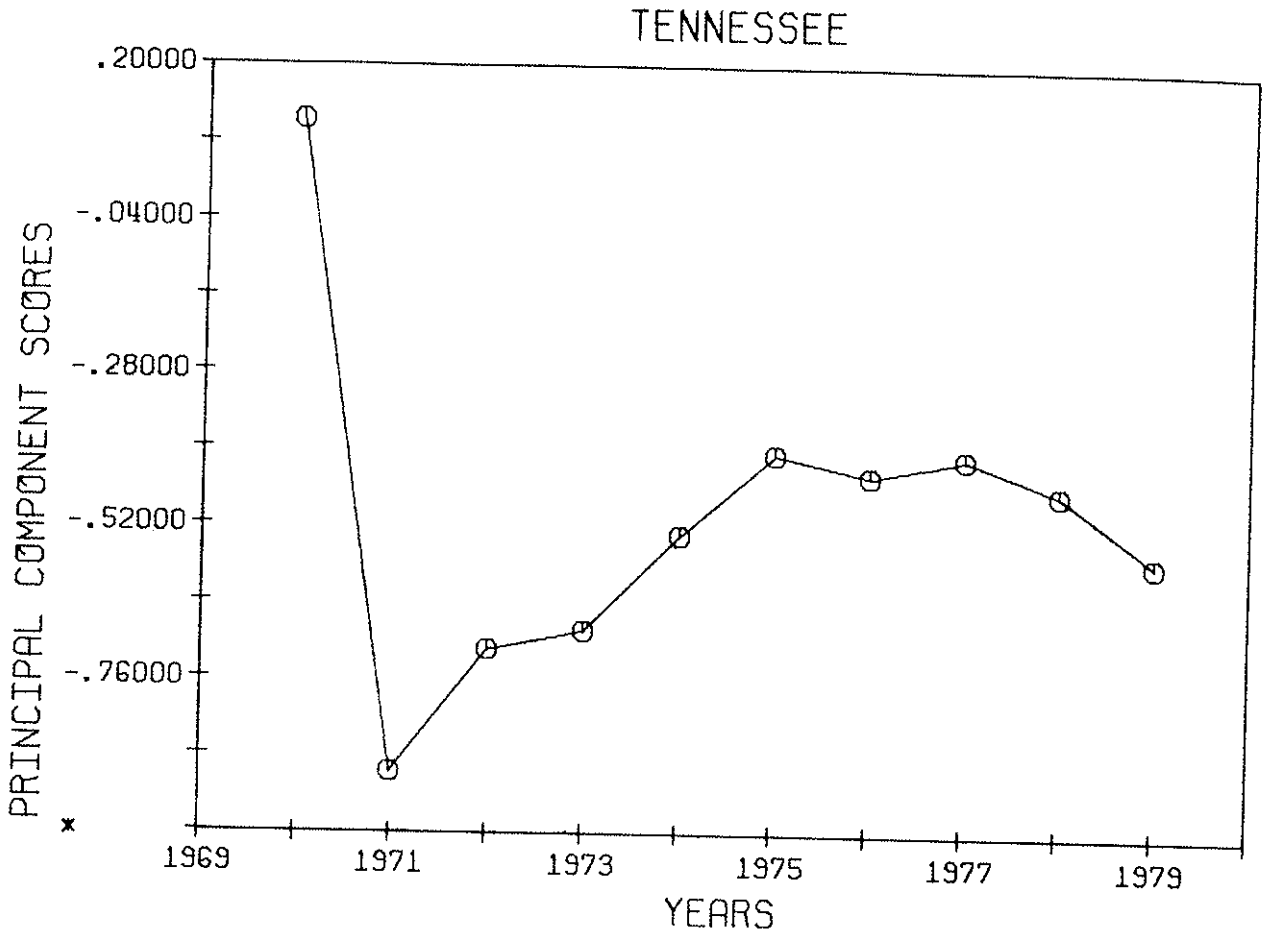

SOUTHERN ILLINOIS
.70000
.42000
.14000
-.14000
-.42000
-.70000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

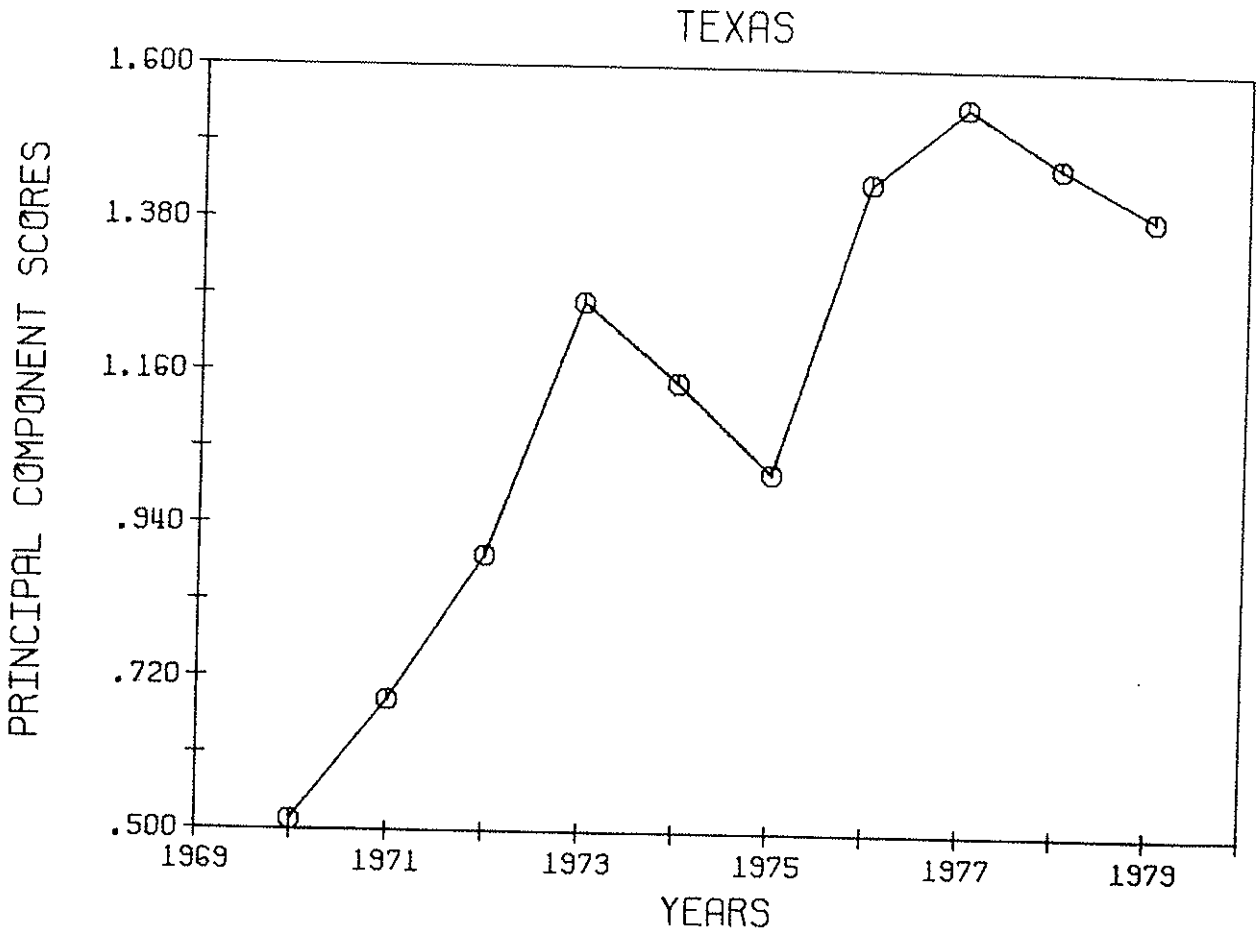

STANFØRD
2.100
1.900
1.700
1.500
1.300
1.100
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

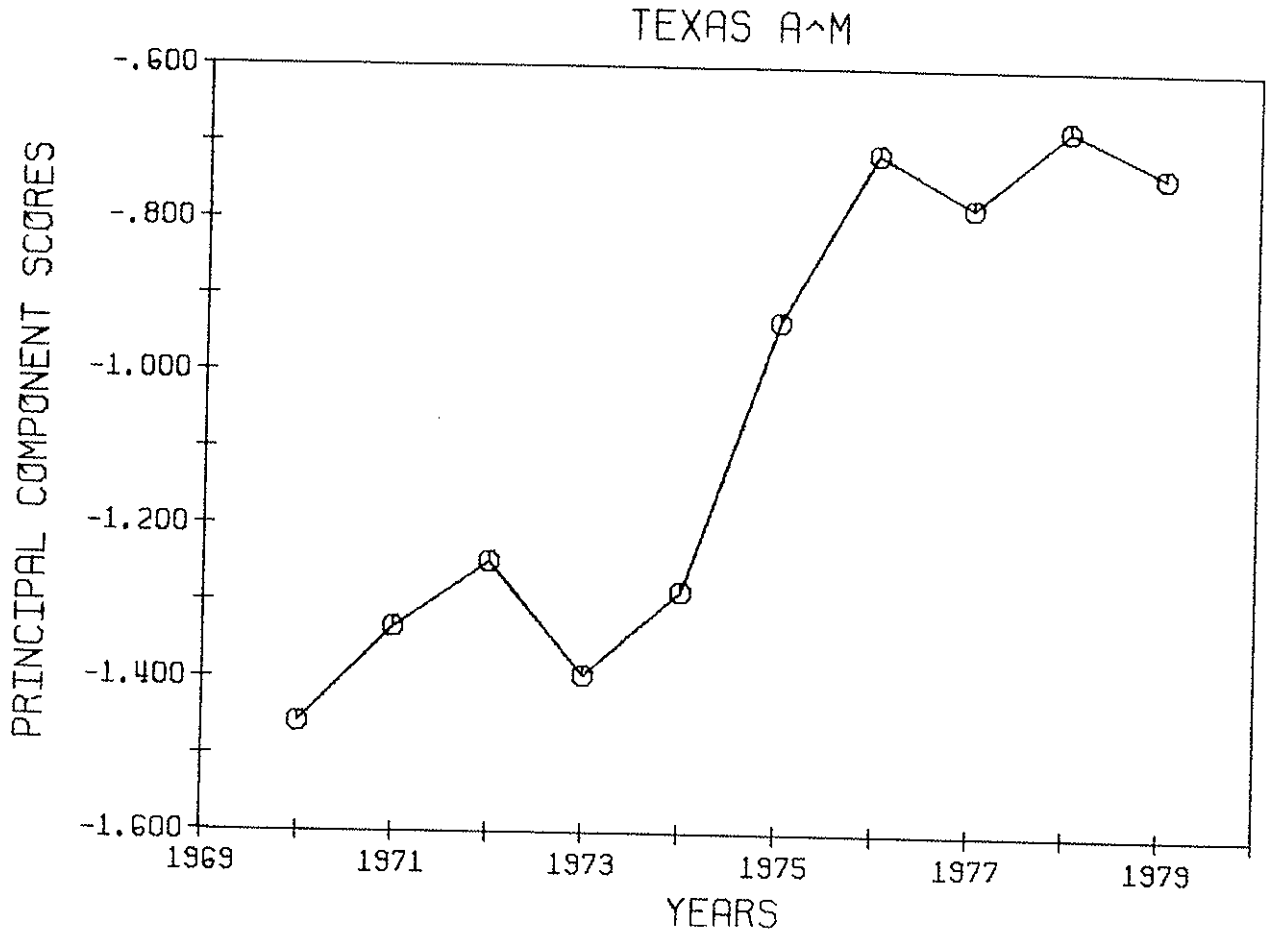

SUNY-BUFFALO
.60000
.40000
.20000
-.00000
-.20000
-.40000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

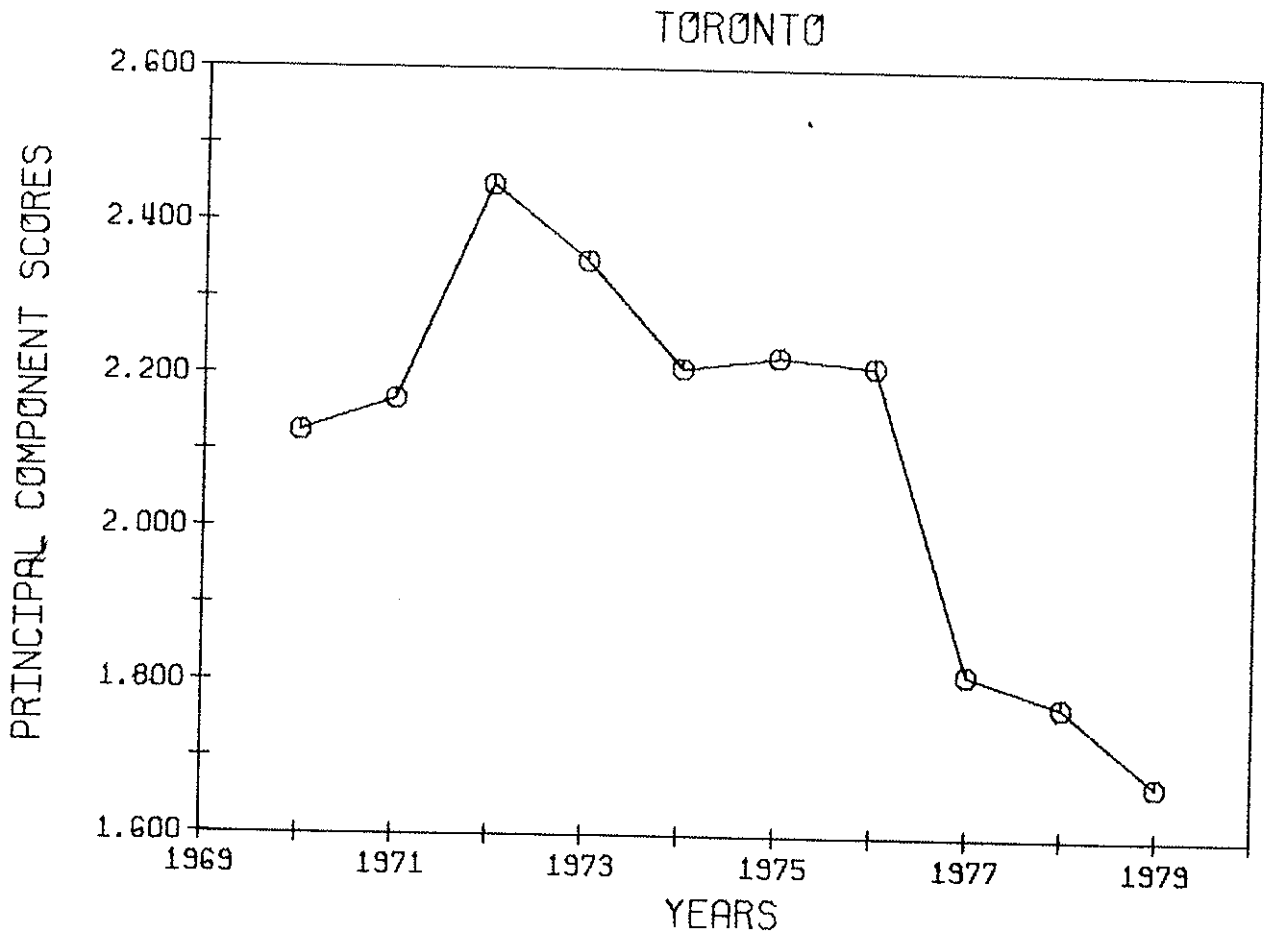

SYRACUSE
.20000
-.00000
-.20000
-.40000
-.60000
-.80000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

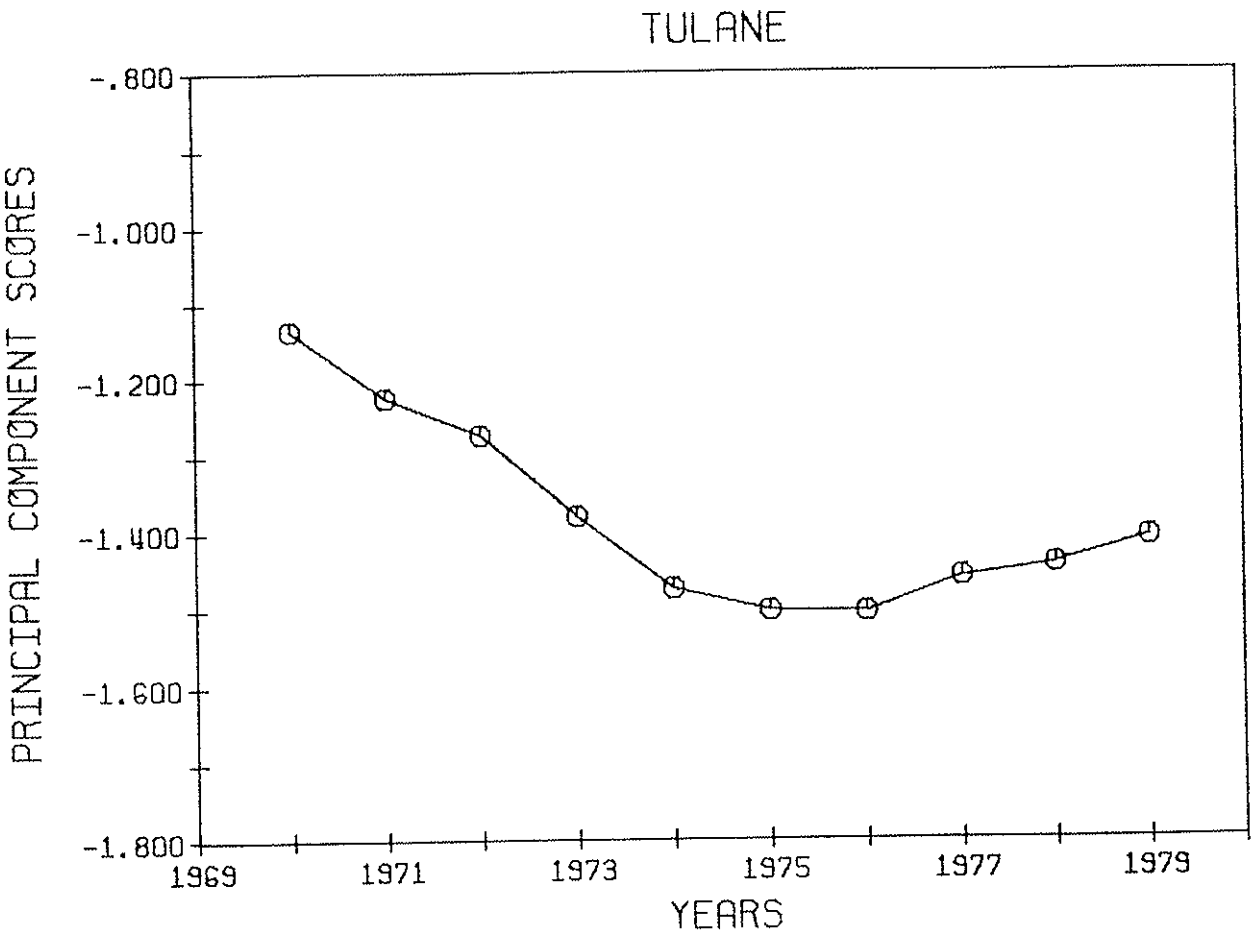

TEMPLE
.10000
-.10000
-.30000
-.50000
-.70000
-.90000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

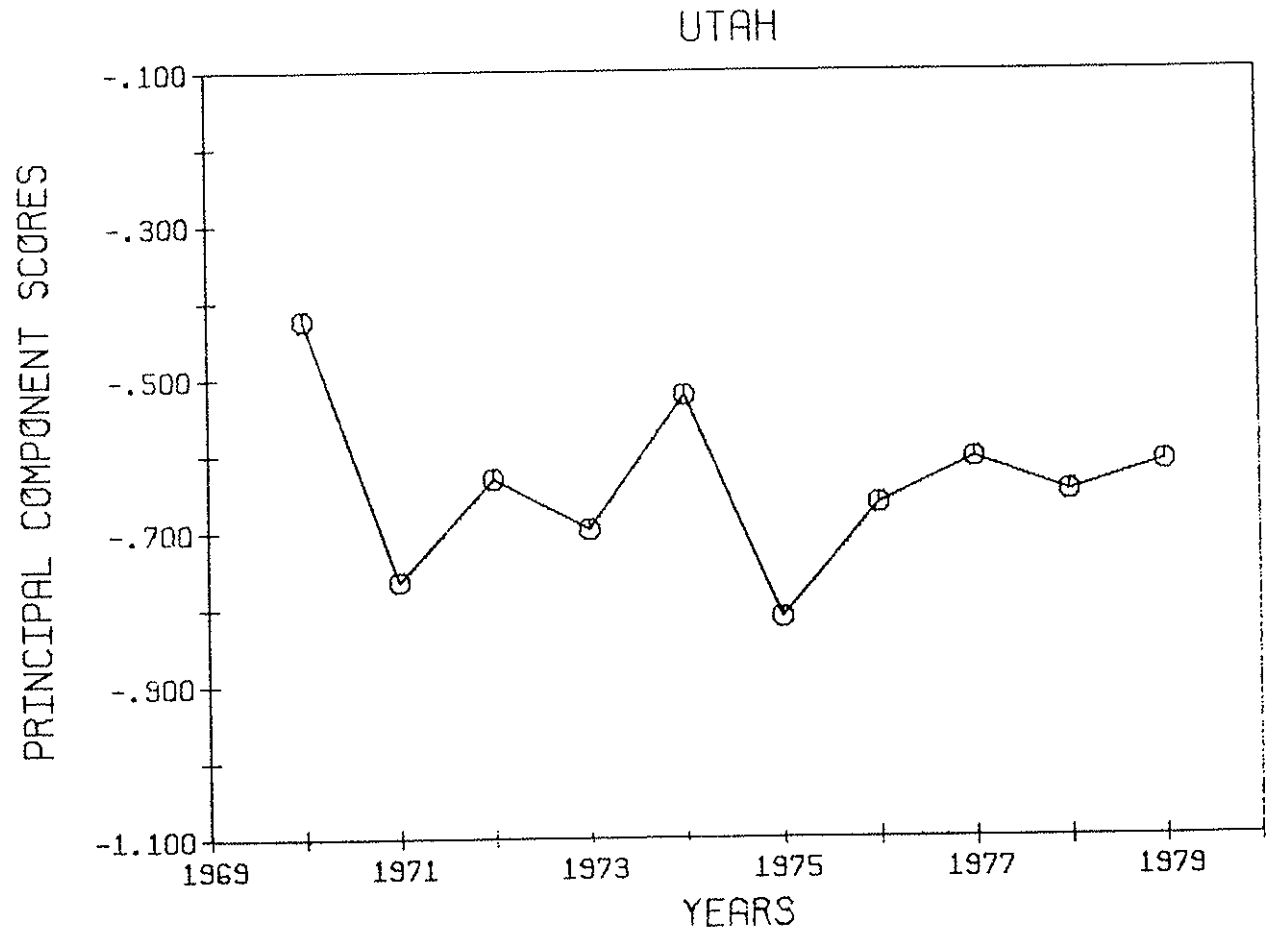

TENNESSEE
.20000
-.04000
-.28000
-.52000
-.76000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

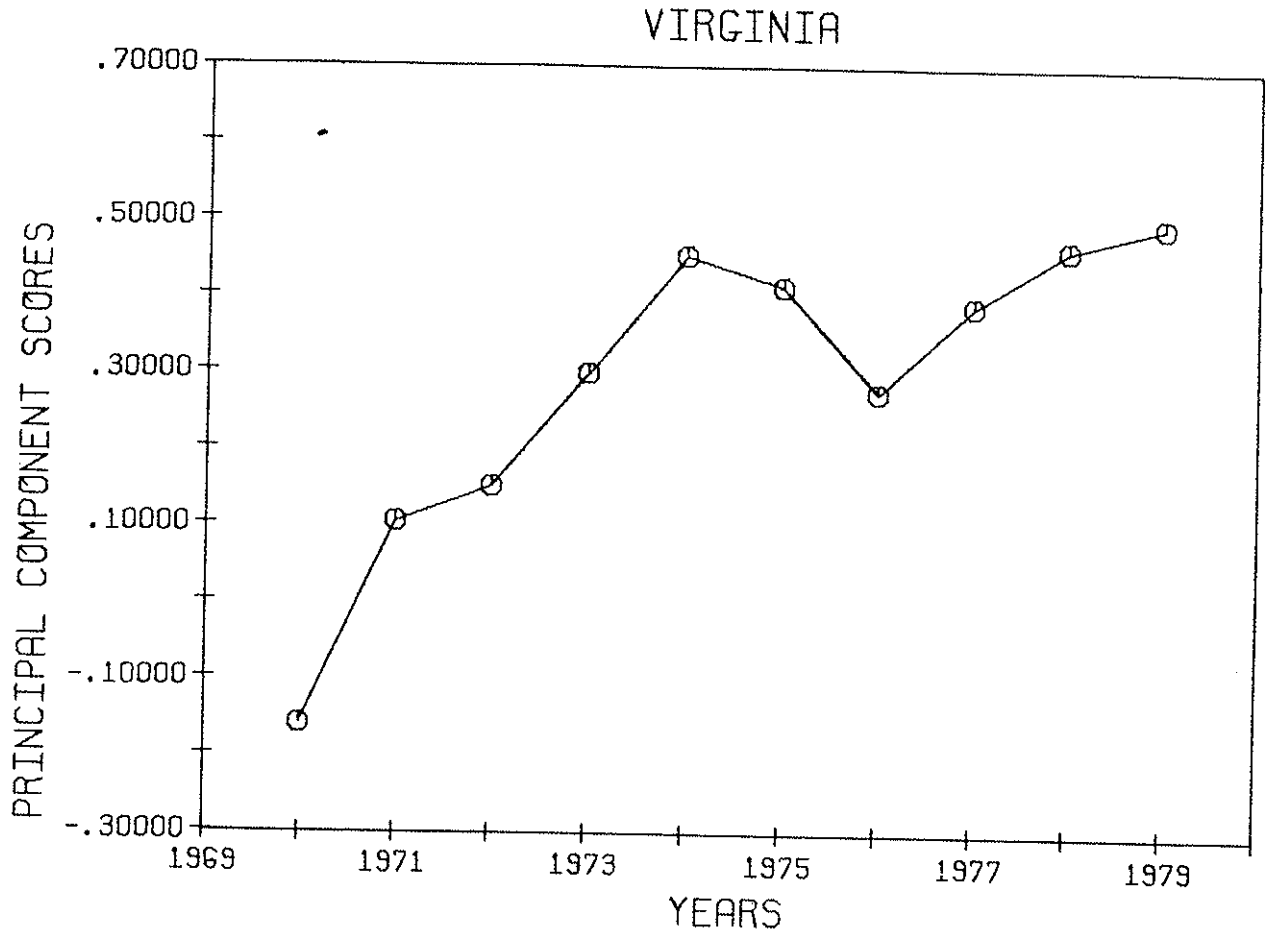

TEXAS
1.600
1.380
1.160
.940
.720
.500
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

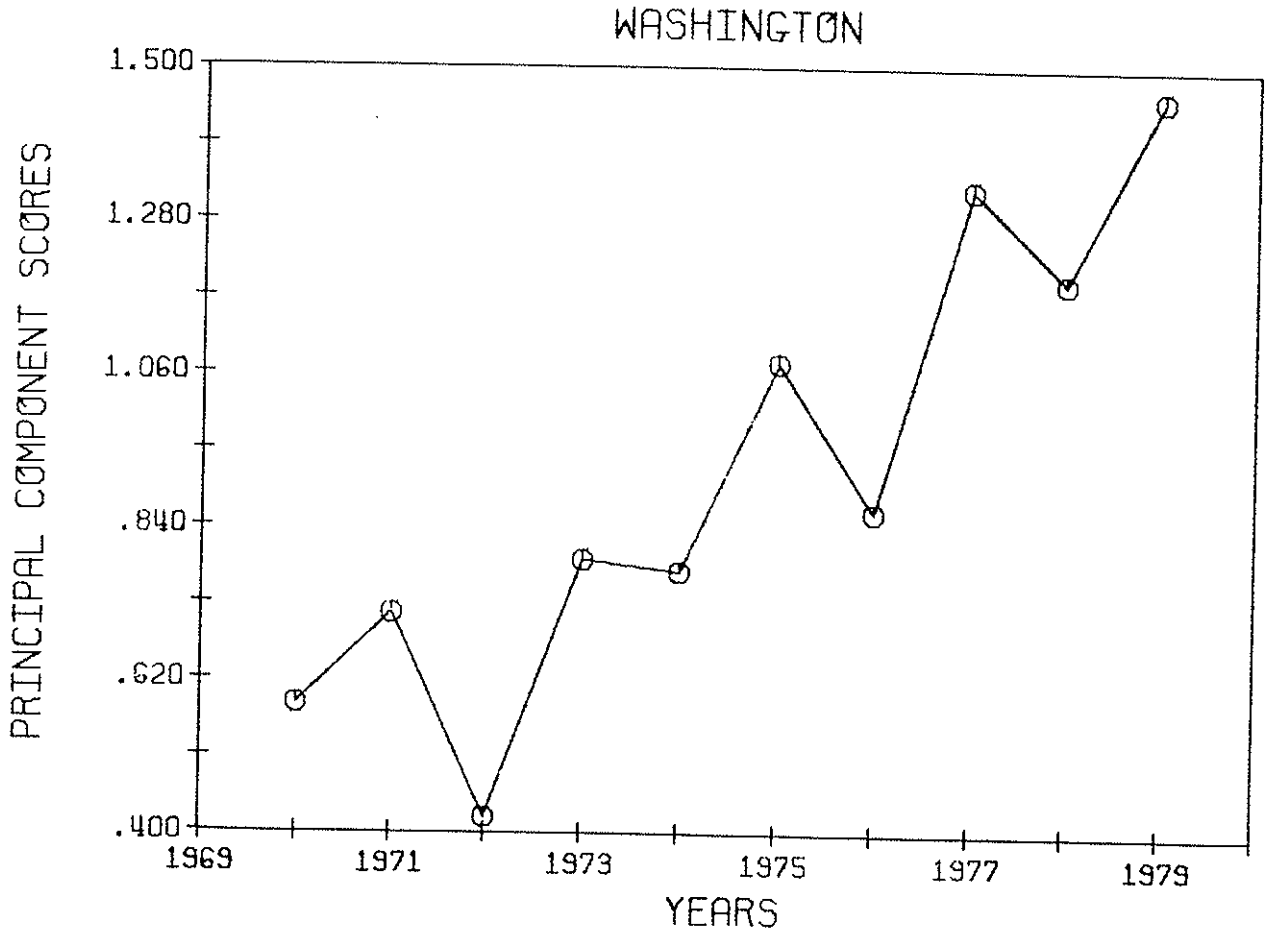

TEXAS A^M
PRINCIPAL COMPONENT SCORES
-.600
-.800
-1.000
-1.200
-1.400
-1.600
1969
1971
1973
1975
1977
1979
YEARS

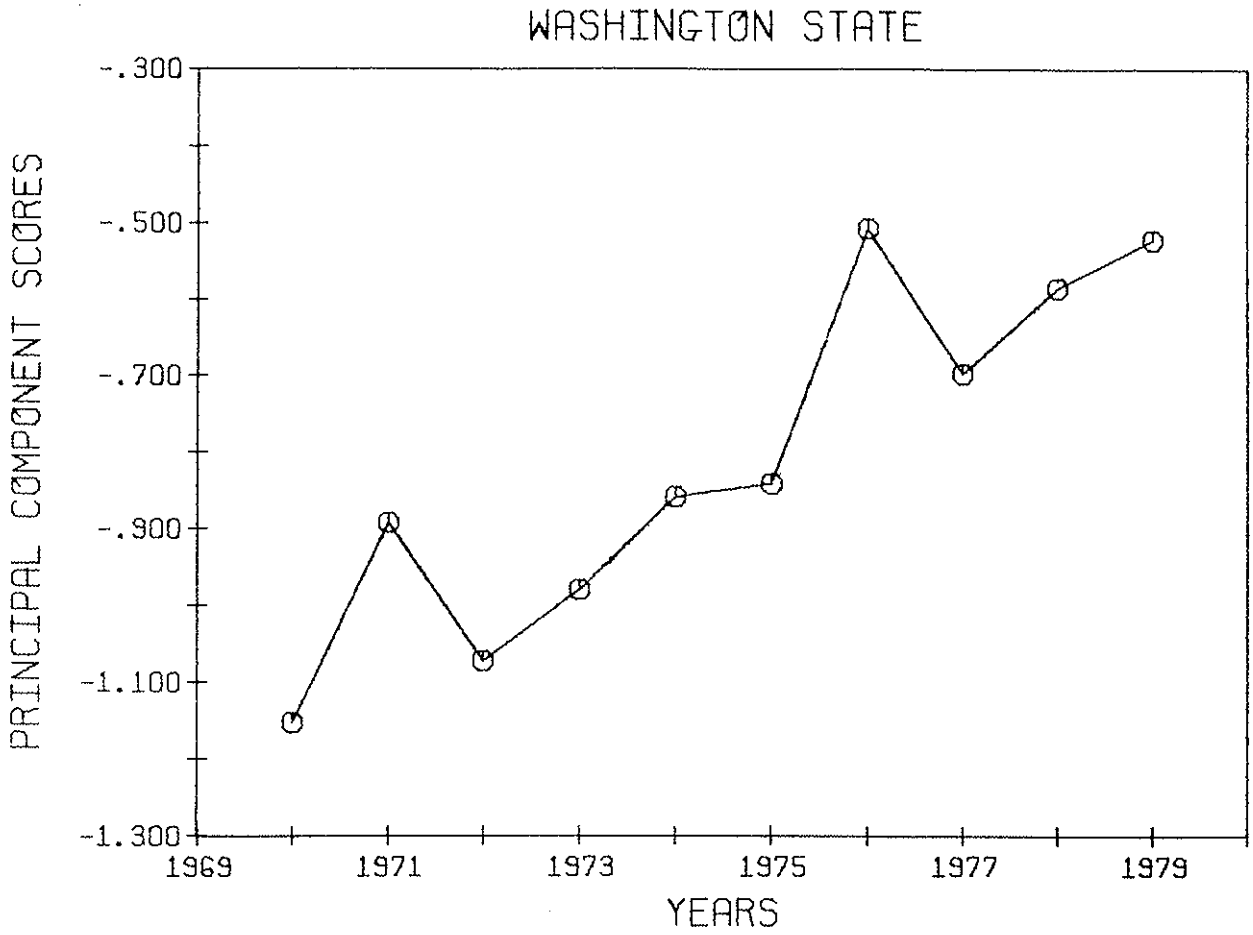

TORONTO
PRINCIPAL COMPONENT SCORES
2.600
2.400
2.200
2.000
1.800
1.600
1969
1971
1973
1975
1977
1979
YEARS

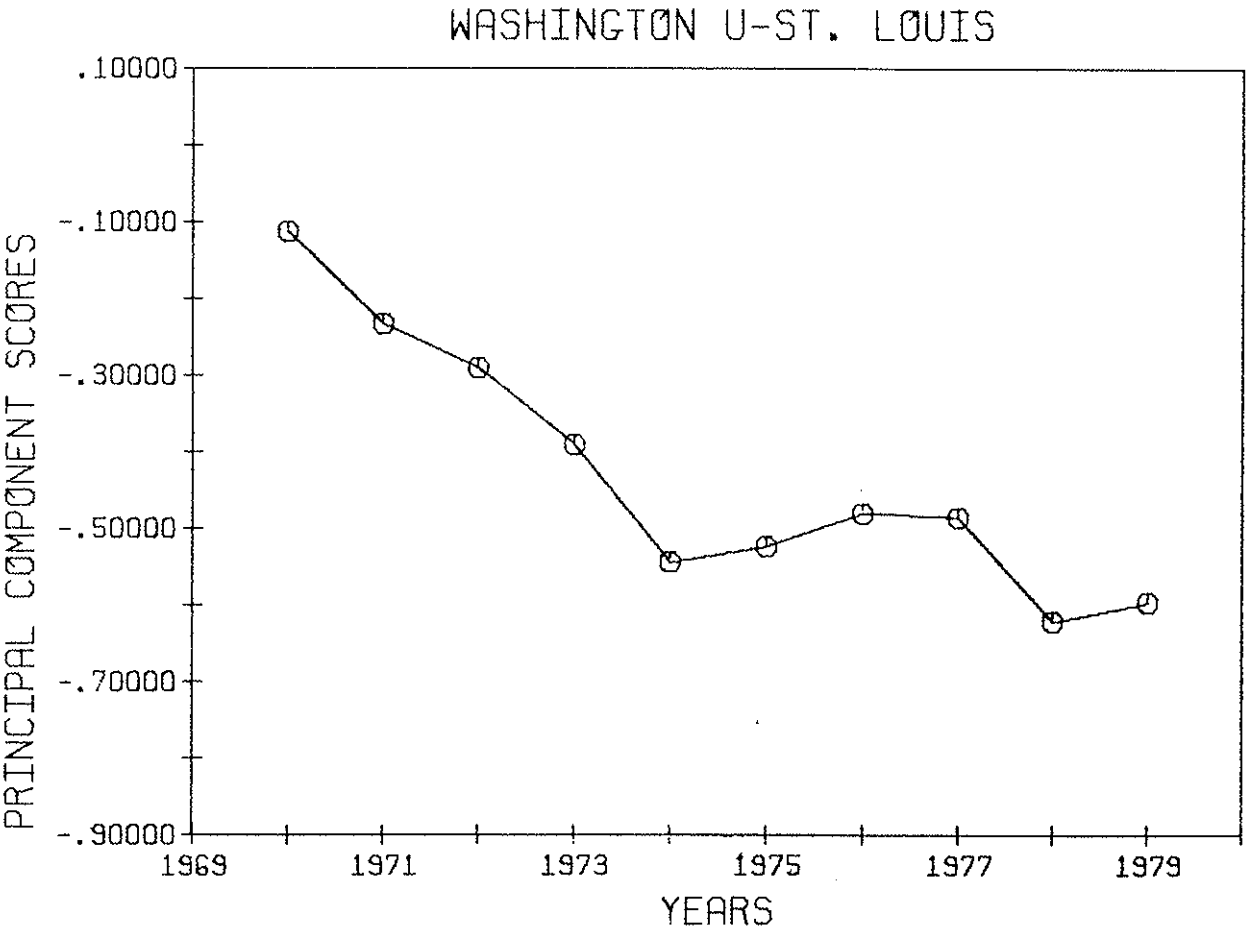

TULANE
PRINCIPAL COMPONENT SCORES
-.800
-1.000
-1.200
-1.400
-1.600
-1.800
1969
1971
1973
1975
1977
1979
YEARS

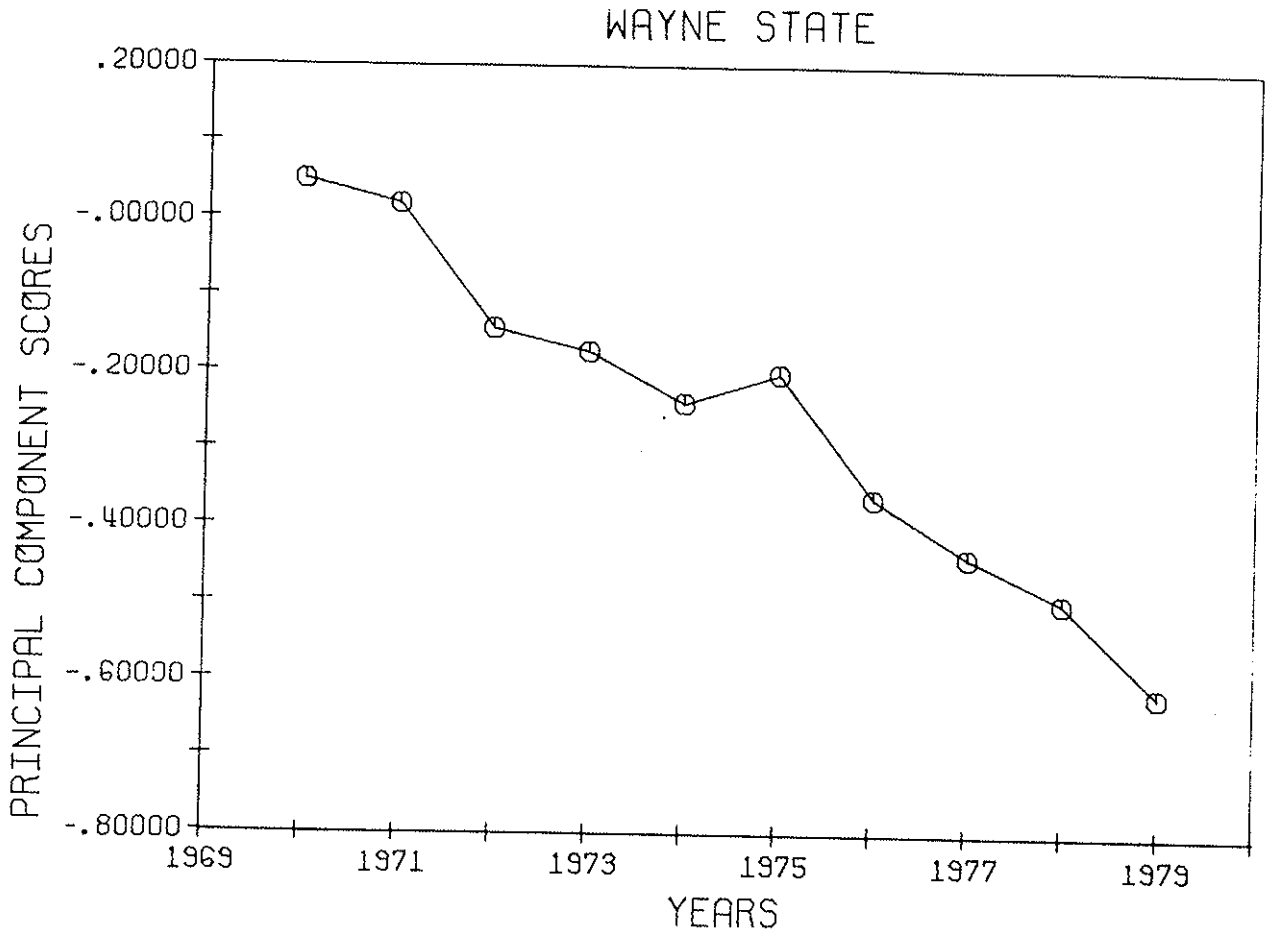

UTAH
PRINCIPAL COMPONENT SCORES
-.100
-.300
-.500
-.700
-.900
-1.100
1969
1971
1973
1975
1977
1979
YEARS

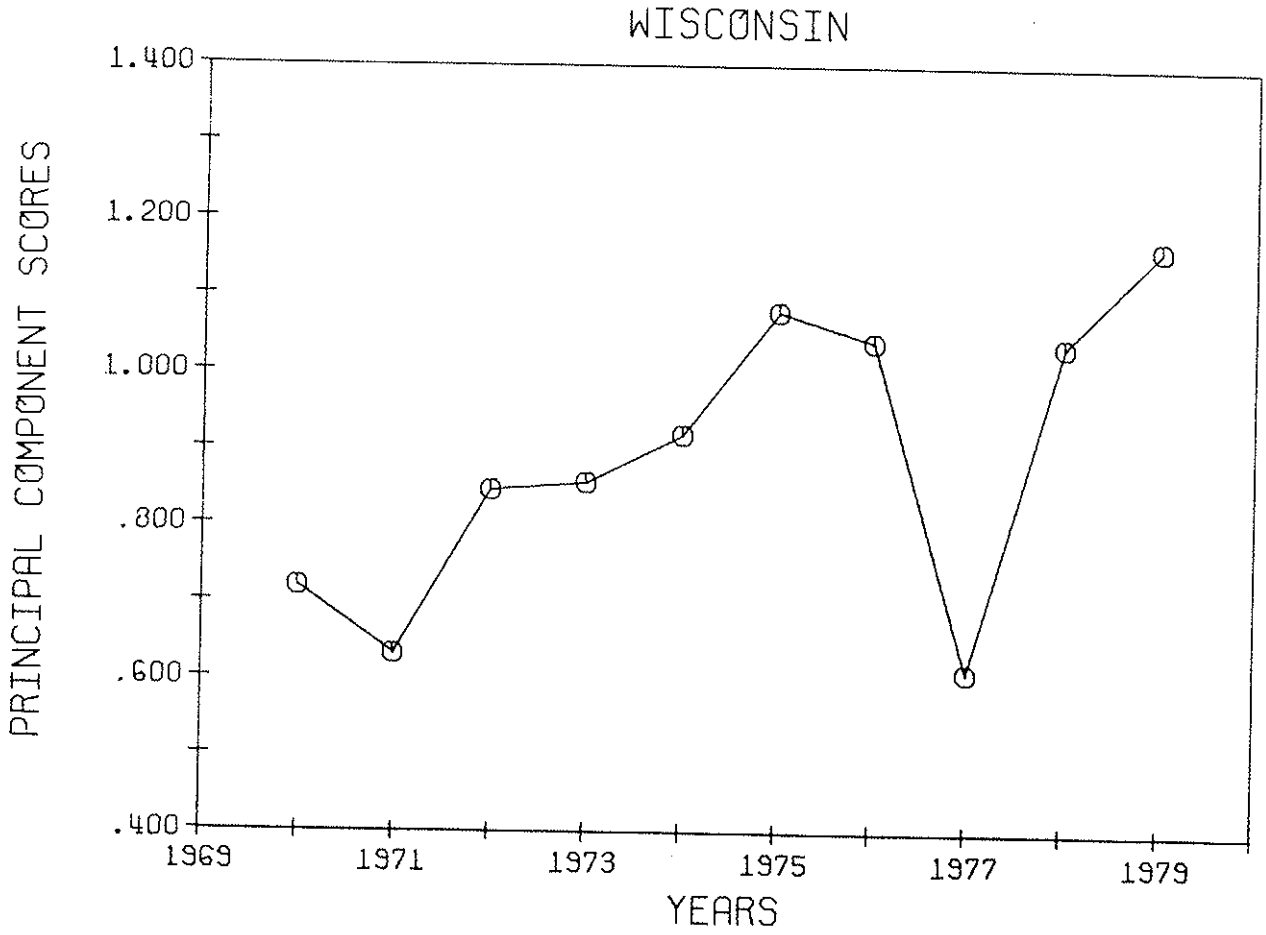

VIRGINIA
.70000
.50000
.30000
.10000
-.10000
-.30000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS

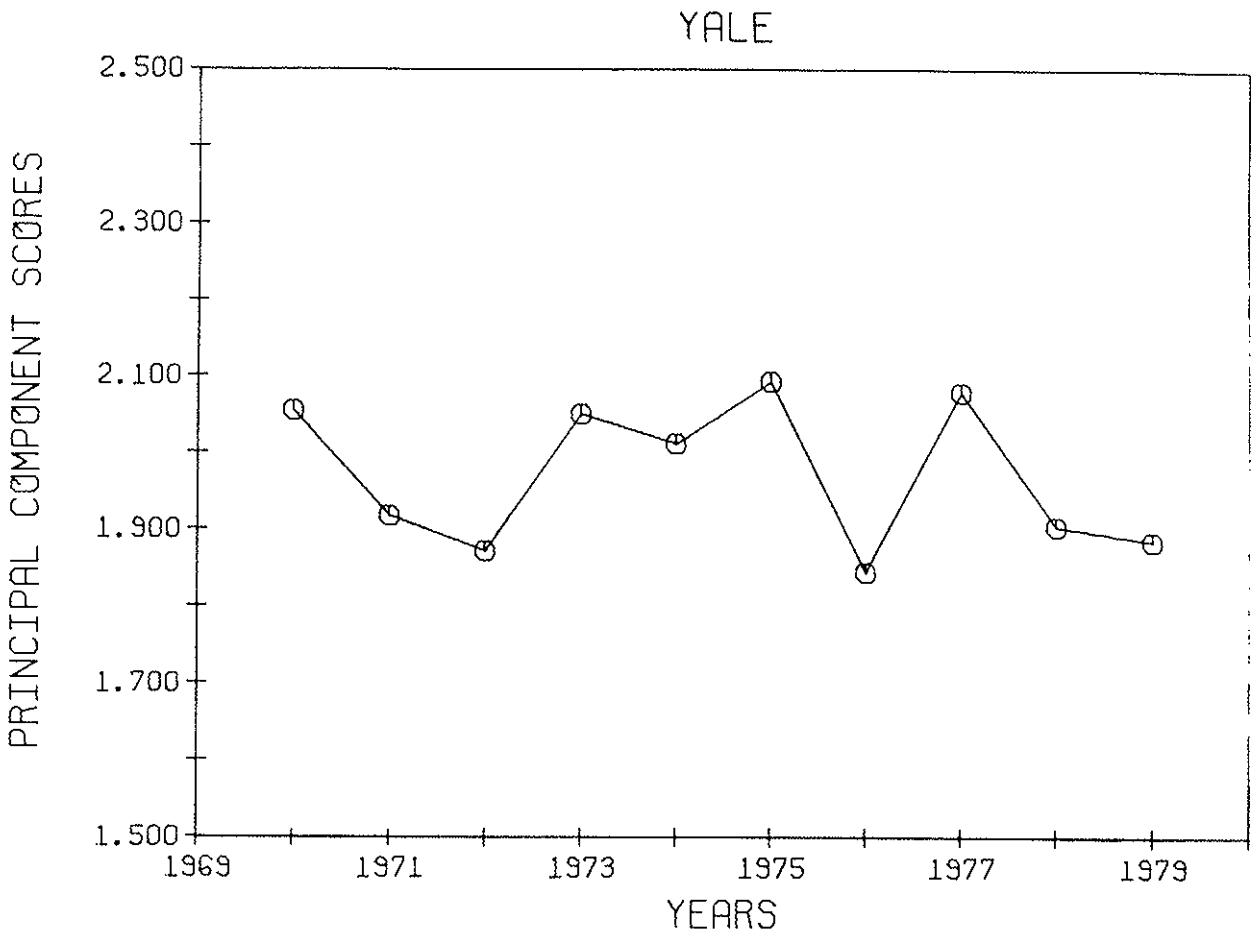

WASHINGTON
1.500
1.280
1.060
.840
.620
.400
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS


WASHINGTON STATE
PRINCIPAL COMPONENT SCORES
-.300
-.500
-.700
-.900
-1.100
-1.300
1969
1971
1973
1975
1977
1979
YEARS


WASHINGTON U-ST. LOUIS
PRINCIPAL COMPONENT SCORES
.10000
-.10000
-.30000
-.50000
-.70000
-.90000
1969
1971
1973
1975
1977
1979
YEARS


WAYNE STATE
.20000
-.00000
-.20000
-.40000
-.60000
-.80000
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS


WISCONSIN
1.400
1.200
1.000
.800
.600
.400
PRINCIPAL COMPONENT SCORES
1969
1971
1973
1975
1977
1979
YEARS


YALE
PRINCIPAL COMPONENT SCORES
2.500
2.300
2.100
1.900
1.700
1.500
1969
1971
1973
1975
1977
1979
YEARS

Appendix: The Calculation of Principal Component Scores

The calculation of scores requires that we multiply the component coefficients or weights by each library's data and sum the ten products. Unfortunately, the data must first be transformed in two ways; and these transformations undoubtedly make a simple process seem very mysterious indeed.

In the first place, various studies have shown that the ARL data (and other bibliometric data) are lognormal rather than normal.[8] The use of lognormal data in principal component analysis (and in other statistical procedures) has various debilitating effects. It is therefore necessary to use logarithms of the ARL data rather than the raw data in the principal component calculations. In the second place, we want the scores in standard normal form. Any data can be standardized by subtracting the mean and dividing by the standard deviation. For example, in 1978-1979 the mean of the logarithms of volumes in the ARL is 6.2839, and the standard deviation is .2191. The standard value of volumes for any ARL library is therefore the log of that library's volumes minus 6.2839 divided by .2191.

These two transformations do not change the rank of libraries in any category of data. Even after logarithms are taken and the data are standardized, each library's rank in volumes, volumes added, etc., is precisely the same as when the raw data rankings are displayed in the ARL Statistics. The transformations, in other words, do not cause any loss of information from the raw data.

As a result of the transformations the formula for component scores is still component weight times data, but the data are now in the form: logs of raw data minus means divided by standard deviations, for each of the ten library size variables.[9] The 1978-1979 formula is thus

1978-1979 ARL principal component scores =

.12431 (log of volumes held - 6.2839)/.2191

+ .11905 (log of volumes added gross - 4.8686)/.2113

+ .07064 (log of microforms - 6.0572)/.1839

+ .12297 (log of current serials - 4.3312)/.2381

+ .12284 (log of expenditures for materials - 6.2000)/.1752

+ .11364 (log of expenditures for binding - 5.0255)/.2414

+ .12743 (log of total salaries - 6.4432)/.2123

+ .10920 (log of operating expenditures - 5.6388)/.2565

+ .12674 (log of professional staff - 1.8318)/.1979

+ .11923 (log of nonprofessional staff - 2.1523)/.2119

This formula can be simplified mathematically to

.56737 x log of volumes held

+ .56342 x log of volumes added gross

+ .38412 x log of microforms

+ .51646 x log of current serials

+ .70114 x log of expenditures for materials

+ .47075 x log of expenditures for binding

+ .60024 x log of total salaries

+ .42573 x log of operating expenditures

+ .64042 x log of professional staff

+ .56267 x log of nonprofessional staff

-26.23700

Notes

1. It is not possible within the scope of this paper to describe the derivation of r and r^2. Many introductions to statistics include chapters on r and r^2 and other topics of correlation and regression. For example, there is a lucid discussion in Frederick Herzon, Introduction to Statistics for the Social Sciences (New York: Crowell, 1976), pp. 325-373.

2. William Baumol and Matityahu Marcus, Economics of Academic Libraries (Washington: American Council on Education, 1973), pp. 85-86.

3. For further examples of regression and correlation among ARL data see Donald Koepp et al., Regression Analysis of the ARL Data (Washington: Association of Research Libraries, 1978).

4. The standard monograph on factor analysis is Harry Harman, Modern Factor Analysis (Chicago: University of Chicago Press, 1976; 3rd edition rev.). More accessible (and less mathematical) treatments are Jae-On Kim and Charles Mueller, Introduction to Factor Analysis (Beverly Hills: Sage, 1978); Jae-On Kim and Charles Mueller, Factor Analysis: Statistical Methods and Practical Issues (Beverly Hills: Sage, 1978); R.J. Rummel, Applied Factor Analysis (Evanston: Northwestern University Press, 1970); and various texts on multivariate analysis, such as Spencer Bennett and David Bowers, An Introduction to Multivariate Techniques for Social and Behavioural Sciences (New York: Wiley, 1976), pp. 8-71. Factor analysis has even entered popular literature. In his novel The Terminal Man (New York: Knopf, 1972), pp. 46-50, Michael Crichton describes in some detail the use of factor analysis to measure the growing psychotic tendencies of the "terminal man".

5. Principal component analysis is discussed in most works on factor analysis. For the rationale of using principal components in calculating scores, see Wayne Velicer, "The Relation Between Factor Score Estimates, Image Scores, and Principal Component Scores," Educational and Psychological Measurement, 36 (Spring, 1976), 149-159.

6. For some recent arguments that library size and quality of use are not strongly related see Richard DeGennaro, "Library Statistics & User Satisfaction: No Significant Correlation," Journal of Academic Librarianship, 6 (May, 1980), 95; and Stella Bentley, "Academic Library Statistics: A Search for a Meaningful Evaluative Tool," Library Research, 1 (Summer, 1979), 143-152.

7. The libraries with ranges greater than 1.0 in these plots are Colorado (1.1), Florida (1.2), Maryland (1.1), Southern Illinois (1.4), Tennessee (1.2), Texas (1.1), Washington (1.1).

8. The lognormal nature of library data is discussed in Allan Pratt, "The Analysis of Library Statistics", Library Quarterly, 45 (July, 1975), 275-286; George Piternick, "ARL Statistics -- Handle With Care," College and Research Libraries, 38 (September, 1977), 419-423; and elsewhere.

9. In the calculations of scores and in the other calculations described in this paper expenditures by Canadian libraries are converted to U.S. dollar equivalents. The conversion factors are the average for each fiscal year of the average monthly noon exchange rates reported in the Bank of Canada Review and in Canada Year Book. The exchange rates in Canadian dollars per U.S. dollar are

1969-1970:	1.0724
1970-1971:	1.0159
1971-1972:	1.0023
1972-1973:	.9922
1973-1974:	.9872
1974-1975:	.9968
1975-1976:	1.0056
1976-1977:	1.0130
1977-1978:	1.1030
1978-1979:	1.1666