Northwest Atlantic Fisheries Organization

Serial No. 1505

NAFO SCR Doc. 88/62

SCIENTIFIC COUNCIL MEETING - JUNE 1988

Reducing Dimensionality in the Temperature and Salinity Data From Station 27: Same Data, Different Analyses

by

J. C. Rice and S. A. Akenhead

Science Branch, Department of Fisheries and Oceans
P. O. Box 5667, St. John's, Newfoundland A1C 5X1

There is increasing interest in the relationships between oceanographic data and fisheries data. There are a large number of candidate oceanographic variables, eg. temperatures and salinities at a variety of depths at each site. An early step in quantifying oceanography-fisheries interrelationships is reduction of the dimensionality of the oceanographic data set. Other talks in this series address the problem of representing the spatial complexity of oceanographic data more simply. In this paper, we consider some aspects of representing the oceanographic data set with fewer variables. We do not consider time series methods directly; although multivariate time series analysis methods exist, their implementation and interpretation are not straightforward. Rather, we focus on similarities and differences in the results from applying various ordination methods to a single data set.

The data:

From the long time series of oceanographic measurements taken at Station 27, off St. John's, Newfoundland (see Myers et al. 1987 for details of this series), we extracted 16 variables: surface and bottom temperatures and salinities for each quarter. The quarterly values were determined by averaging all records within the interval so different values may be a mean of different numbers of observations. In a total of 38 years from 1950 to 1987, a single variable was missing. In most analyses, the missing value was replaced by the mean for that variable across all other years. One set of analyses investigated the effect of this mean replacement procedure, which is necessary in order to obtain values of the new (reduced dimensionality) variables for years with a single missing value in the raw data set.

The analyses:

Two ordination methods were applied to the data - factor analysis (Morrison 1976) and detrended correspondence analysis (DECORANA) (Gauch 1982). In the factor analysis, axes were extracted using principal components analysis, and subjected to either varimax or quartimax rotations. Both rotations are rigid; functionally, varimax rotations are intended to maximize the interpretability of cases (years have scores as close or 0 or to ± 1 as possible on each axis); whereas, quartimax rotations are intended to maximize the interpretability of the axes (each original variable has loadings as close to 0 or ± 1 as possible on each axis). A third factor analysis was performed on the data set after years missing a variable were dropped rather than replacing the missing value with the mean.

DECORANA takes a different approach to ordinating data. Cases and variables are arranged simultaneously, to maximize the correlations between cases (years) and variables (oceanographic attributes). Using a scoring procedure, first, cases are arranged according to their variable scores; then, variables are arranged according to their case scores. This alternation among cases and variables is iterated (hence, the common name of "reciprocal averaging") until patterns stabilize. Although both factor analysis and correspondence analysis are eigenvector techniques and hence share many similarities, they differ in the types of statistical artifacts each is sensitive to and the ways that violations of assumptions affect results.

Results and Discussion

Factor analysis gave five axes with eigenvalues greater than 1.0 (i.e., axes representing at least as much variation as a single variable). Functional interpretations of some axes are readily apparent (Table 1). For example, factor 1 represents a gradient of bottom temperatures, which show similar variation in all quarters; and factor 3 represents variation in bottom salinities which are also consistent across seasons. Other factors may be initially less obvious, but with some thought are reasonable; such as factor 4 the modest inverse association of spring and summer temperatures with winter salinities. Note that influences of surface temperature attributes usually are spread among more than one axes, and no single axes captures variation in surface temperatures or salinities in more than two quarters. This is also reasonable, suggesting important variation in surface attributes occurs on shorter time scales than in bottom attributes.

The 3 different variations of factor analyses extracted almost identical patterns from the data. Correlations of variable loadings between quartimax and varimax rotated axes were consistently near 1.0 for corresponding axes (axis 1, r = 0.998; axis 2, r = 0.996; axis 3, r = 0.999; axis 4, r = 0.979; axis 5, r = 0.844). When years with a missing observation were deleted, the factor analysis reported one fewer axes, but the variable loadings of the axes still matched those of the full data set well (for quartimax rotations: axis 1, r = 0.967; axis 2, r = 0.978; axis 3, r = 0.959; axis 4, r = 0.965). Other than losing an axis representing variation in spring and summer surface temperature, the analyses of the reduced data set clearly presents the same pattern as analysis of the data set where missing data were replaced with the mean for that variable.

Correspondence analysis extracted very different patterns from the same data set. Individually, the axes are as interpretable as those extracted by the factor analyses (Table 2). For example, axis 1 captures variation in surface temperatures and to a lesser extent bottom salinities for spring, summer, and fall. All the salinity measures are reflected in axis 2. The variable loadings of DECORANA axes 1, 2, and 4 are correlated with corresponding loadings from at least two axes from the factor analysis (Table 3). For example, DECORANA axis 1 has similarities with factors 1 and 4; whereas, DECORANA axis 2 relates to factors 1 through 4. DECORANA axis 3 is not closely associated with any factor, and factor 5, the potentially interesting spring/summer surface temperature factor, is not closely associated with any DECORANA axis.

Additional detailed analyses might bring out further subtle relationships between DECORANA and factor axes but would be incidental to the thrust of this paper. We do call attention to one pattern which will be relevant to our discussion. Scores of each year were plotted together for DECORANA and quartimax axes that were correlated in Table 3. For some portions of the interval from 1950 to 1987, matches are good (Fig. 1-4). For example, after 1966, DECORANA axis 1 and quartimax axis 1 show very similar patterns (The signs are arbitrary), but prior to 1966, no association is noticeable. On the other hand, scores on DECORANA axis 4 and quartimax axis 3 match closely until 1972 but not after that time.

The patterns pointed out above could be examined more quantitatively with appropriate partitioning of the full series and correlations analyses of the subsets. Such analyses are both inappropriate and unnecessary, because they would be examples of flagrant a posteriori data snooping and because only a general point needs to be drawn from the graphs. Undoubtedly, there are real gradients and trends in the data sets, but no small number of trends seem to dominate over the full time period. Different analysis methods are likely to pick up major signals in the data set, but the details of how things are patched together is dependent on details of the analysis method chosen.

What alternatives are available to researchers, given this sort of difference between the results of two ordination techniques applied to a single data set? One alternative is simply to avoid data reduction approaches altogether and use the data as collected. Investigations of relationships between oceanographic attributes and fisheries will then have to consider a very large number of independent variables, with corresponding problems of very large Type I and/or Type II error rates, or else researchers will subjectively choose a smaller set of independent oceanographic variables, keeping error rates lower at a cost of throwing away possibly important information.

Another alternative is to replace the statistical methods for condensing large data sets with process oriented studies. Such studies may simplify the data set by elucidating the mechanisms underlying the patterns present in the raw oceanographic data. The potentially smaller numbered parameters of the process oriented models may then be substituted for the original oceanographic variables. There are many reasons to endorse good studies of oceanographic processes, but it is unlikely that such studies will eliminate the need for empirical tools of data reduction in the near future. Good process-oriented studies often require years to complete, while management needs can require investigation of fisheries-oceanography relationships with much shorter completion times. Furthermore, most believe it is unlikely that only a few processes are at work in the Northwest Atlantic. For example, the extensive and thorough work on ice and wind effects on the ocean climate off Newfoundland (Myers et al. 1988) is still basically one process. The variability most directly related to the ir process seems reflected in axes 1 and 3 of the factor analyses. Other factors also could be argued to be consistent with that (or other) specific process but in a post hoc manner. The key point is the occurence of major and apparently systematic variation (factors 1 to 4) which is not clearly attributable to a well documented oceanographic process. It is unrealistic to expect process-oriented oceanographic studies to resolve oceanographic data into a few well-understood parameters in the near future. Such studies should be encouraged, but the problem of data reduction also needs to be addressed.

Another alternative is to select a single data reduction method to apply, on the basis of some best match to statistical properties of the data set, and after considering the assumptions and robustness of the alternative methods. It is always worthwhile to examine the statistical attributes of data sets and the assumptions of one's statistical tools. Realistically, it is unlikely that any single method will emerge as clearly most appropriate. There is a large literature comparing alternative ordination techniques, but recent reviews indicate no consensus is emerging (Gaugh 1982, Greenacre 1984, Morrison 1976, Orloci 1978, Pielou 1984, Wartenberg et al. 1987).

The best alternative is to make the best of what we have. When different data reduction methods give identical answers, enjoy the luxury of a simple answer. When different methods give different answers, or if only a single method is used (including the method of subjectively selecting a few oceanographic attributes based on "insight" or prejudice), use the results which appear most meaningful with no pretention that one has captured the entire and the only truth. There will be many meaningful ways to look at the variation in oceanographic data. More opportunities may be lost than advantages are gained by attaching too much interpretative importance to any single representation. If the procedures followed are presented clearly, and ordinations and relationships are presented phenomenologically, useful advances can be made with a variety of tools. This may be the best we can do with systems influenced by multiple and complex processes, and by events which may unpredictably (or infrequently enough to be functionally unpredictable) introduce major disturbances in the state of the system.

References

Gauch, G. G. Jr. 1982. Multivariate Analysis in Community Ecology. Cambridge University Press.

Greenacre, M. P. 1984. Theory and Application of Correspondence Analysis. Academic Press.

Morrison, D. F. 1976. Multivariate Statistical Methods. 2nd Ed. McGraw-Hill, New York.

Myers, R. A., S. A. Akenhead, and K. Drinkwater. 1988. The North Atlantic Oscillation and the Ocean Climate of the Newfoundland Shelf. NAFO SCR Doc. 88/6/(this session)

Orloci, L. 1978. Multivariate Analysis in Vegetation Research. W. Junk, The Hague.

Pielou, E. C. 1984. The Interpretation of Ecological Data. John Wiley and Sons.

Wartenberg, D., S. Ferson, and F. J. Rohlf. 1987. Putting things in order: A critique of detrended correspondence analysis. American Naturalist 129: 434-448.

Table 1. Factor pattern of quartimax rotation of principal components of 16 oceanographic variables from 1950 to 1987. To simplify reading the table, small factor loadings (between 0.40 and -0.40) are not presented.

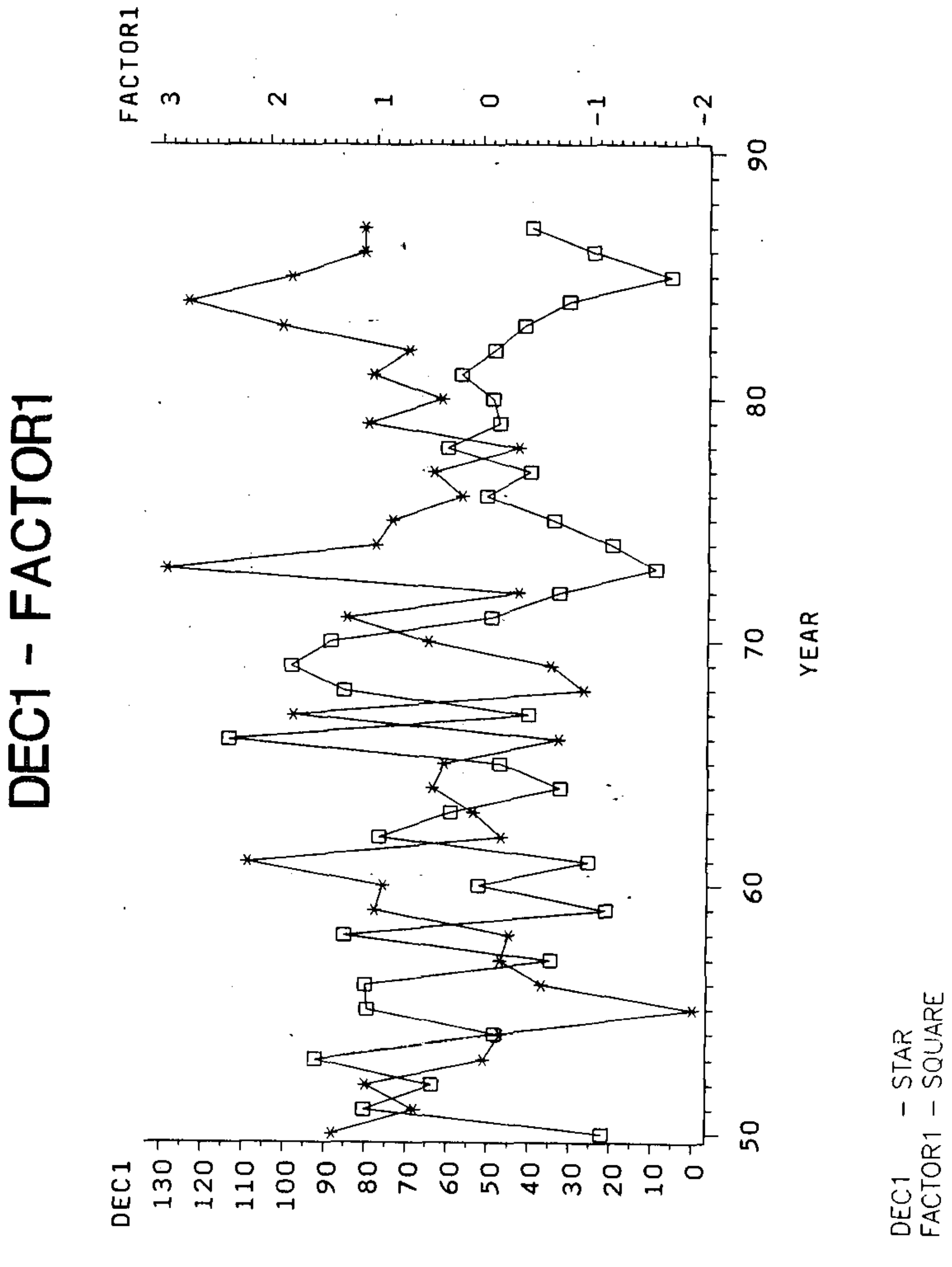
Variable	Factor: 1	2	3	4	5
Surface Temperature:					
Winter	0.892				
Spring	0.426				0.639
Summer				0.648	0.465
Fall				0.811	
Bottom Temperature:					
Winter	0.823				
Spring	0.926				
Summer	0.838				
Fall	0.562	0.571			
Surface Salinity:					
Winter			0.507	-0.460	
Spring				-0.708	
Summer		0.835			
Fall		0.866			
Bottom Salinity:					
Winter					0.623
Spring			0.850		
Summer			0.777		
Fall		0.476	0.601		
% Variance explained by axes	28.4	20.6	9.9	9.3	6.9

Table 2. Scores of the 16 oceanographic attributes on the first four axes of detrended correspondence analysis of the data from 1950 to 1987.

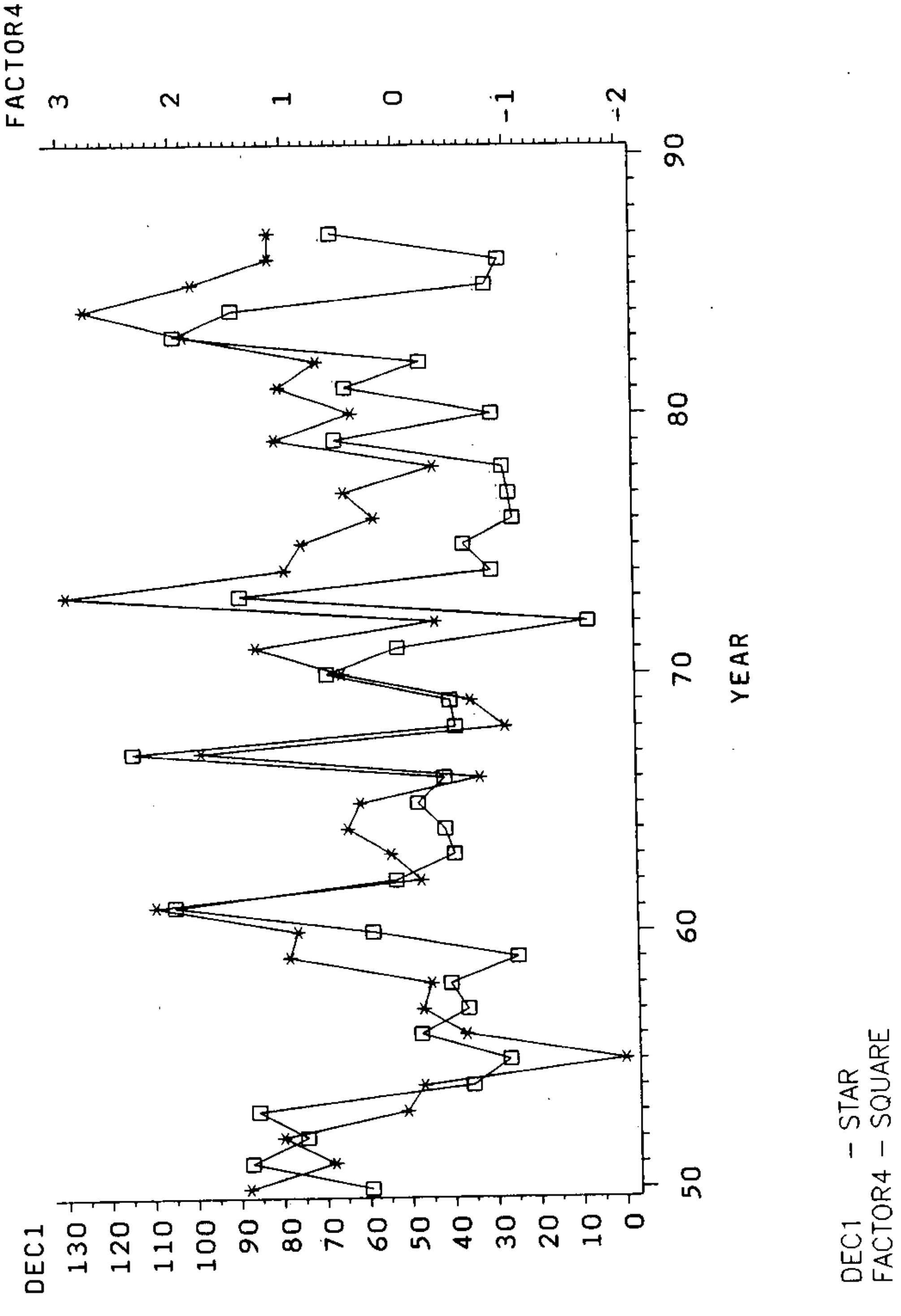
Variable	Axis: 1	2	3	4
Surface Temperatures:				
Winter	- 60	- 72	7	96
Spring	108	- 52	201	80
Summer	228	18	36	-105
Fall	190	50	-123	186
Bottom Temperatures:				
Winter	- 49	45	- 17	- 42
Spring	-125	-113	61	13
Summer	- 50	- 22	-130	- 32
Fall	24	59	- 17	- 27
Surface Salinity:				
Winter	47	255	105	136
Spring	24	198	168	33
Summer	31	135	15	- 32
Fall	75	130	49	- 89
Bottom Salinity:				
Winter	20	122	50	256
Spring	97	151	76	187
Summer	92	216	100	73
Fall	82	193	87	54
% Variance	34.2	26.2	8.8	6.2

Table 3. Pearson product-moment correlation coefficients of quartimax factor loadings (Factors 1-5) and decorana scores of variables (axes 1-4), extracted from the oceanographic data set of 16 variables, collected from 1950 to 1987.

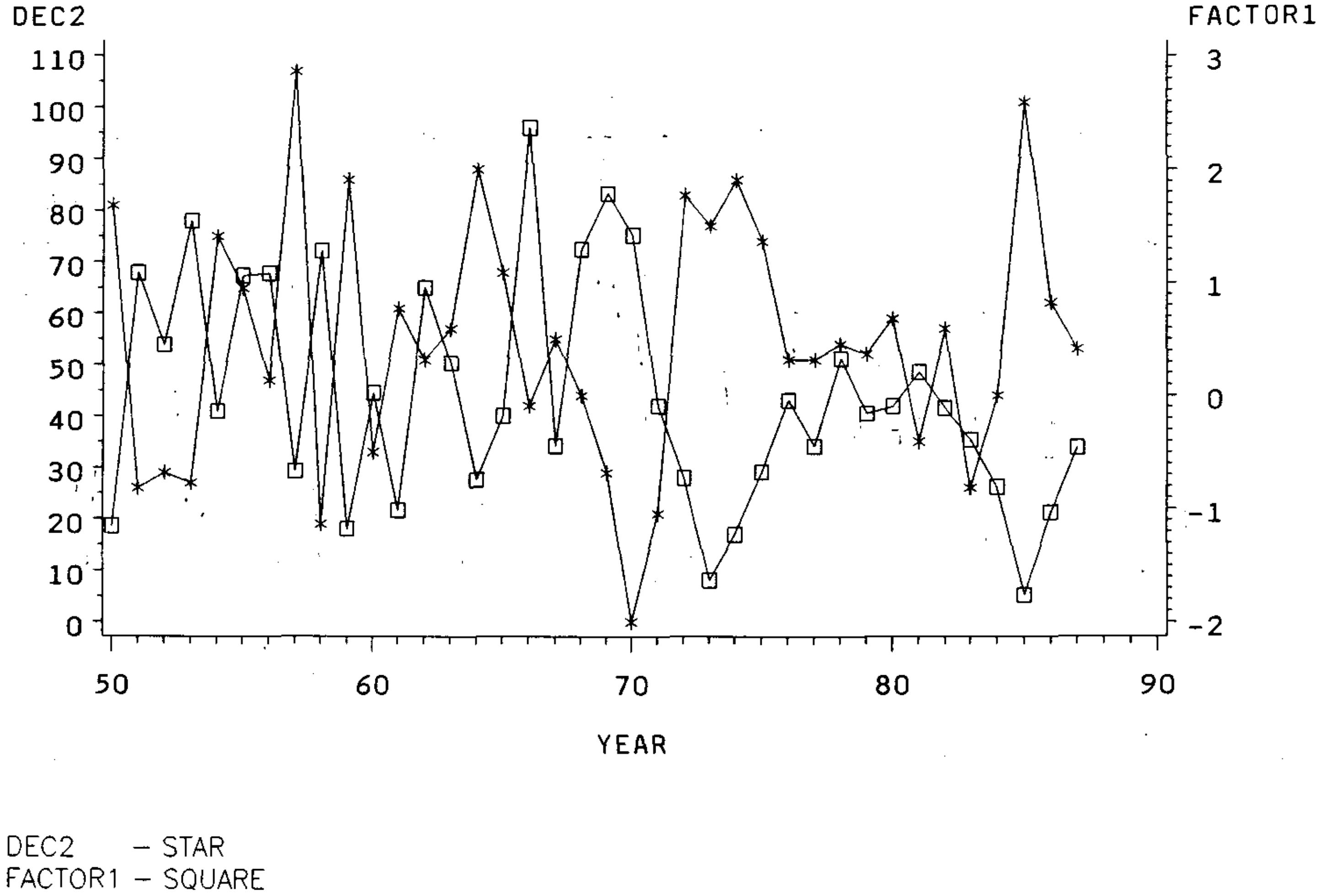
Factor Analysis Axis	Decorana Axis 1	2	3	4
1	-0.748	-0.793	-0.351	-0.309
2	-0.011	0.438	-0.257	-0.551
3	0.089	0.792	0.261	0.735
4	0.507	-0.497	-0.557	0.003
5	0.215	-0.296	0.311	-0.085



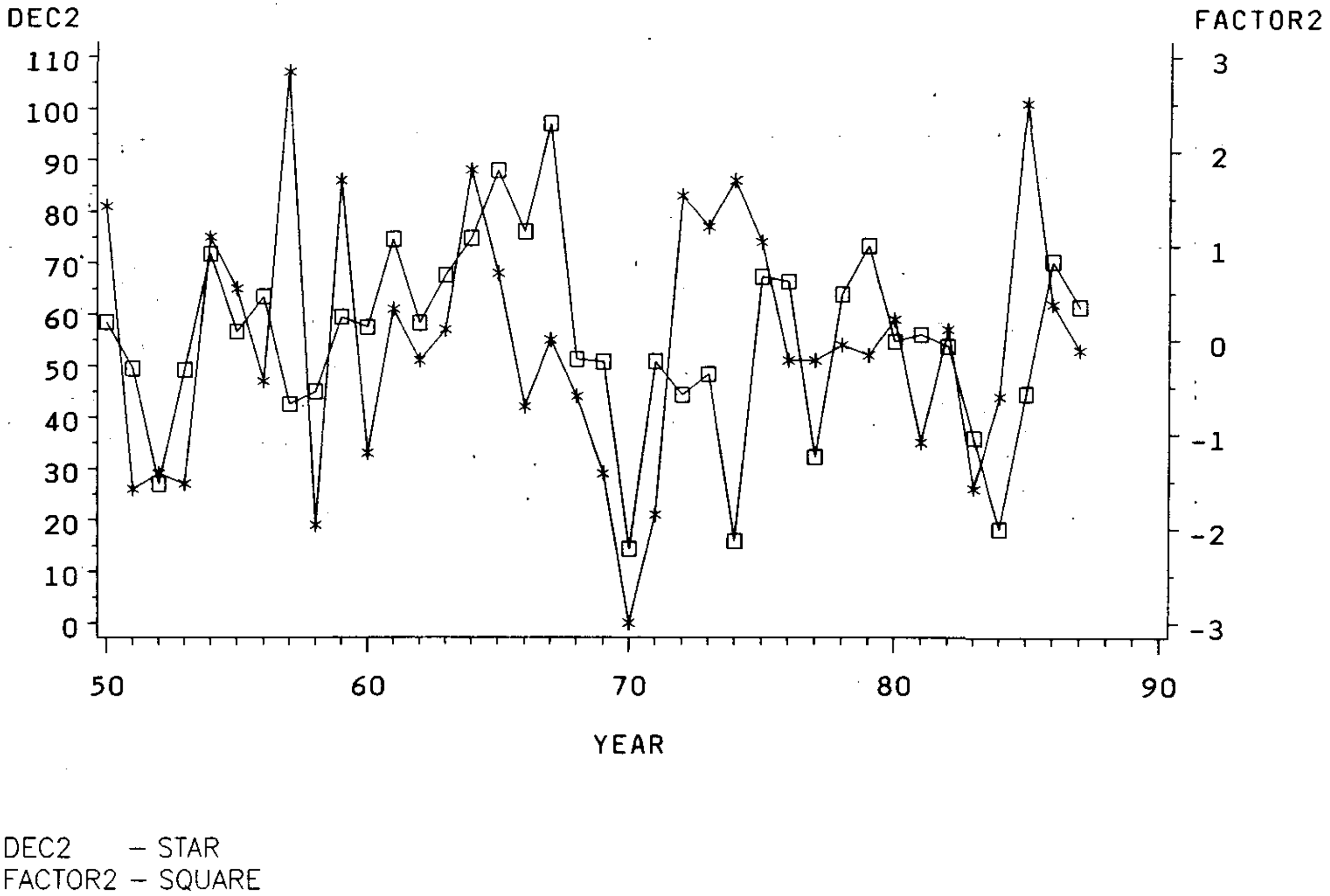
DEC1 - FACTOR4



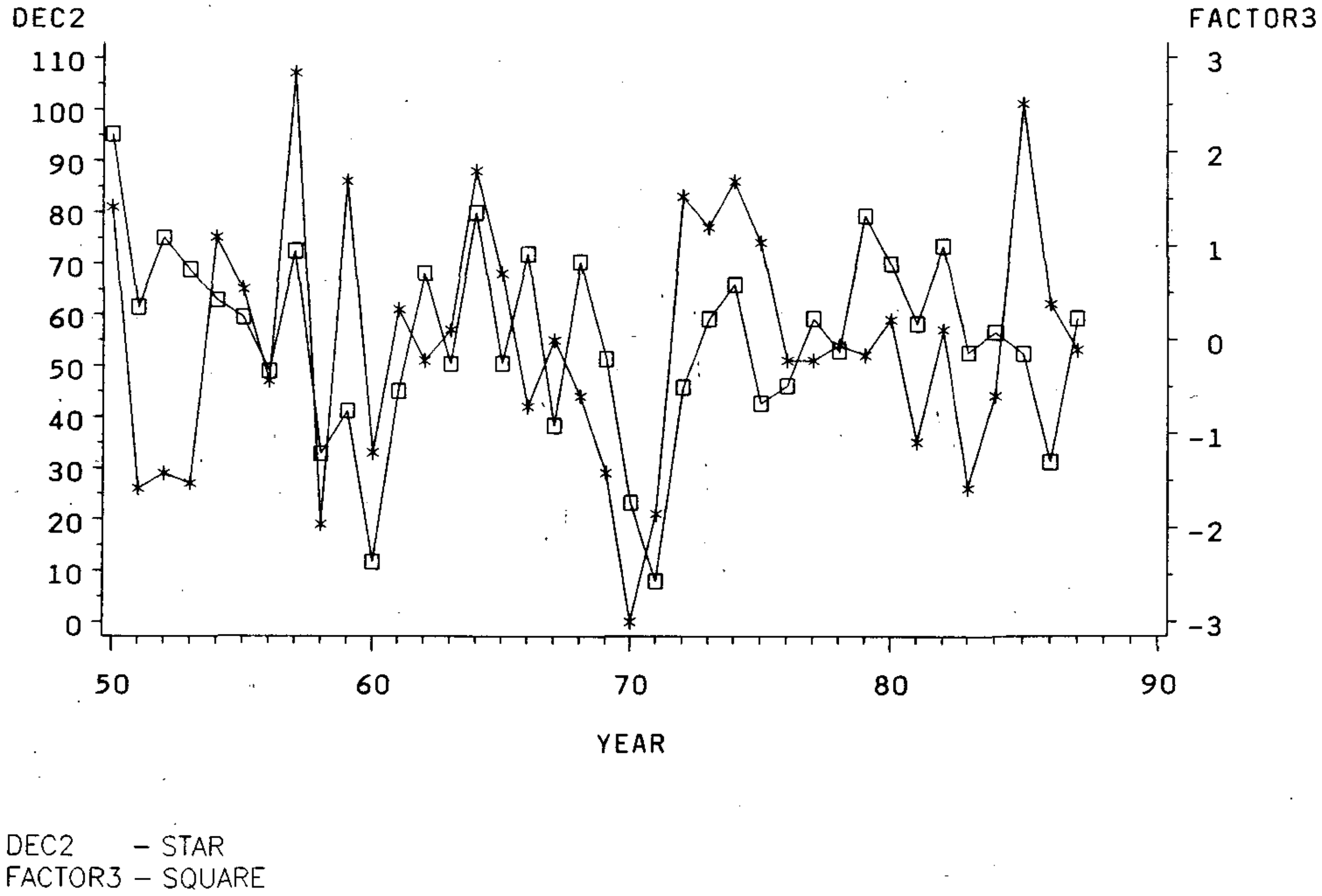
DEC2 - FACTOR1



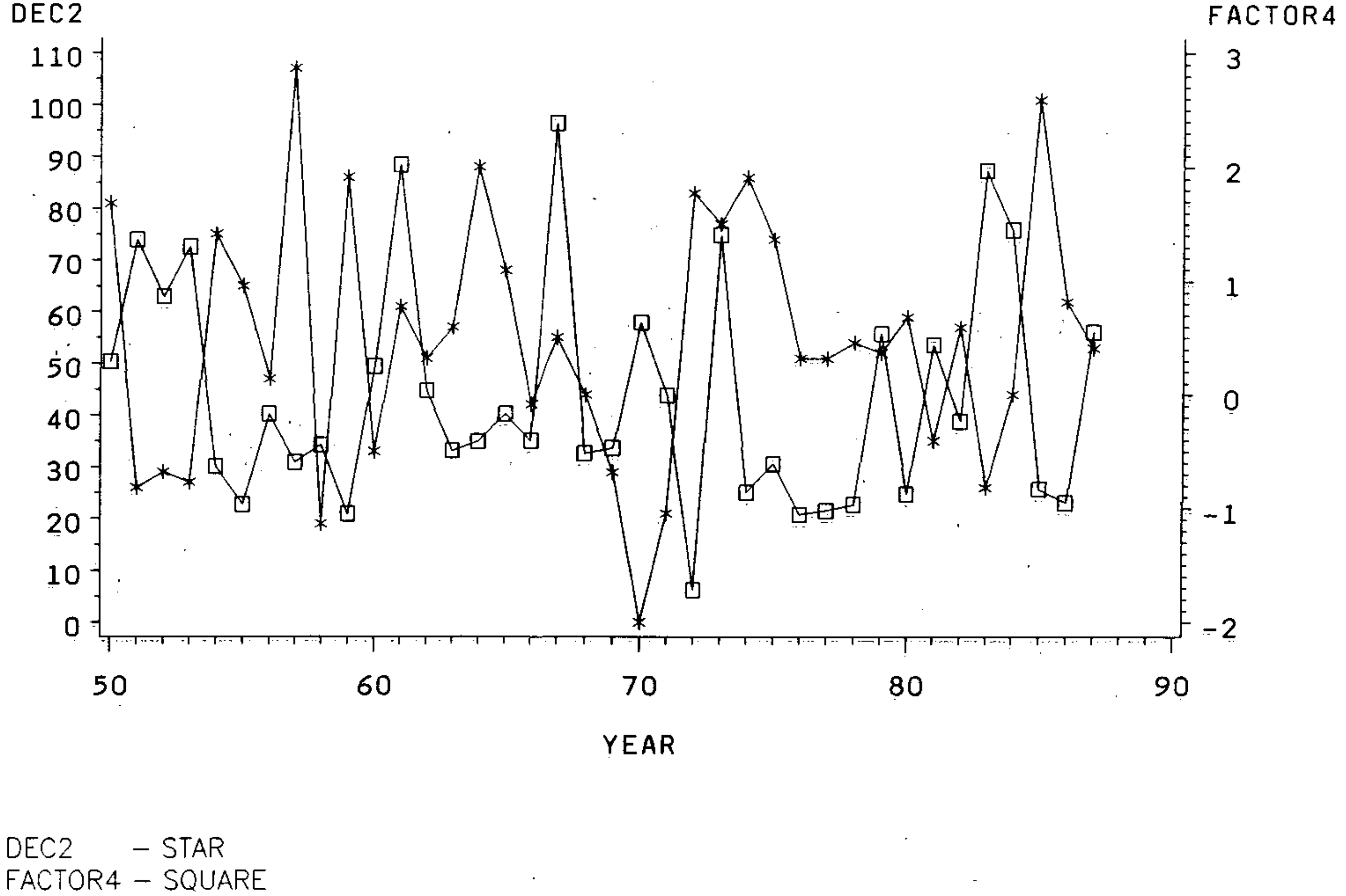
DEC2 - FACTOR2



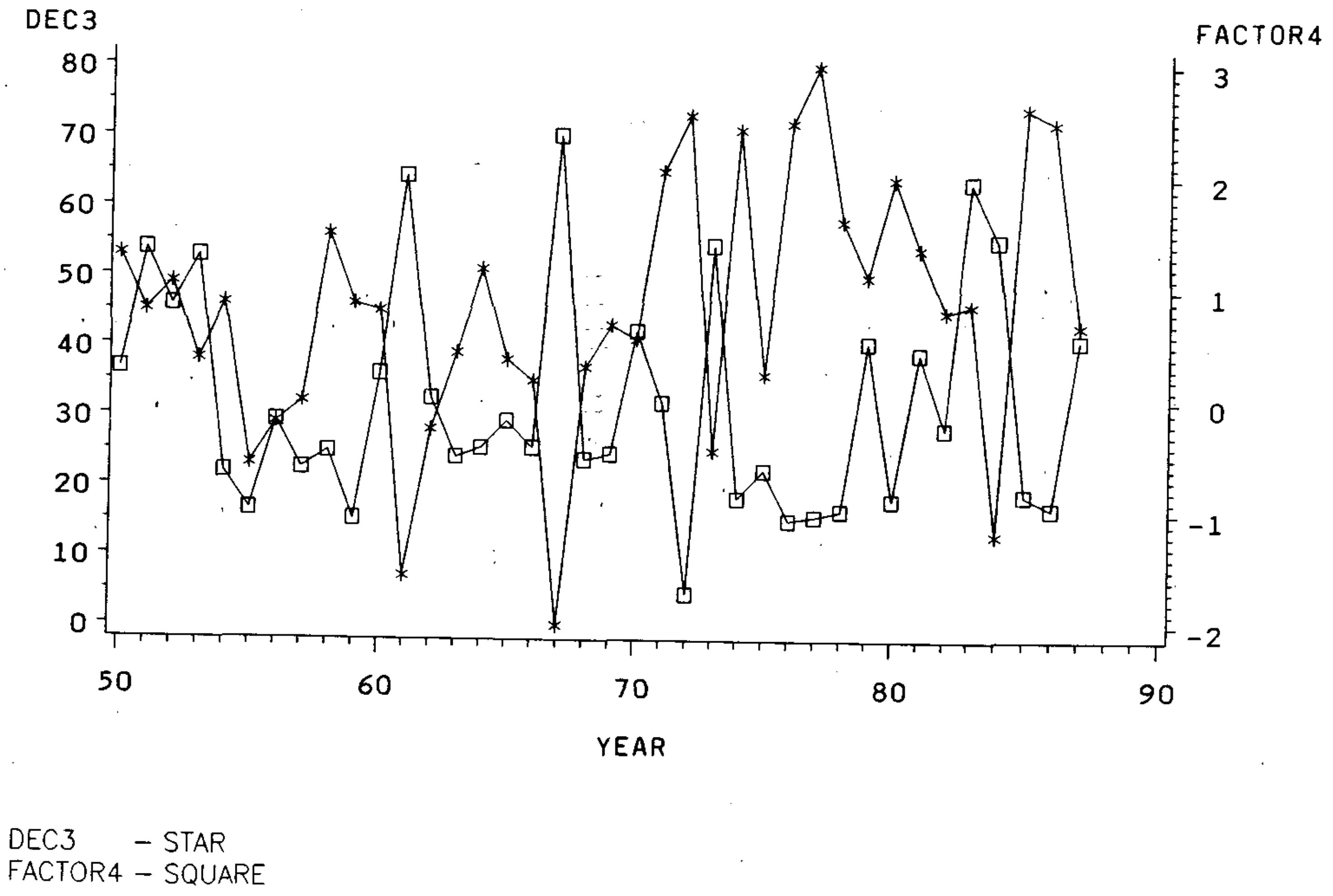
DEC2 - FACTOR3



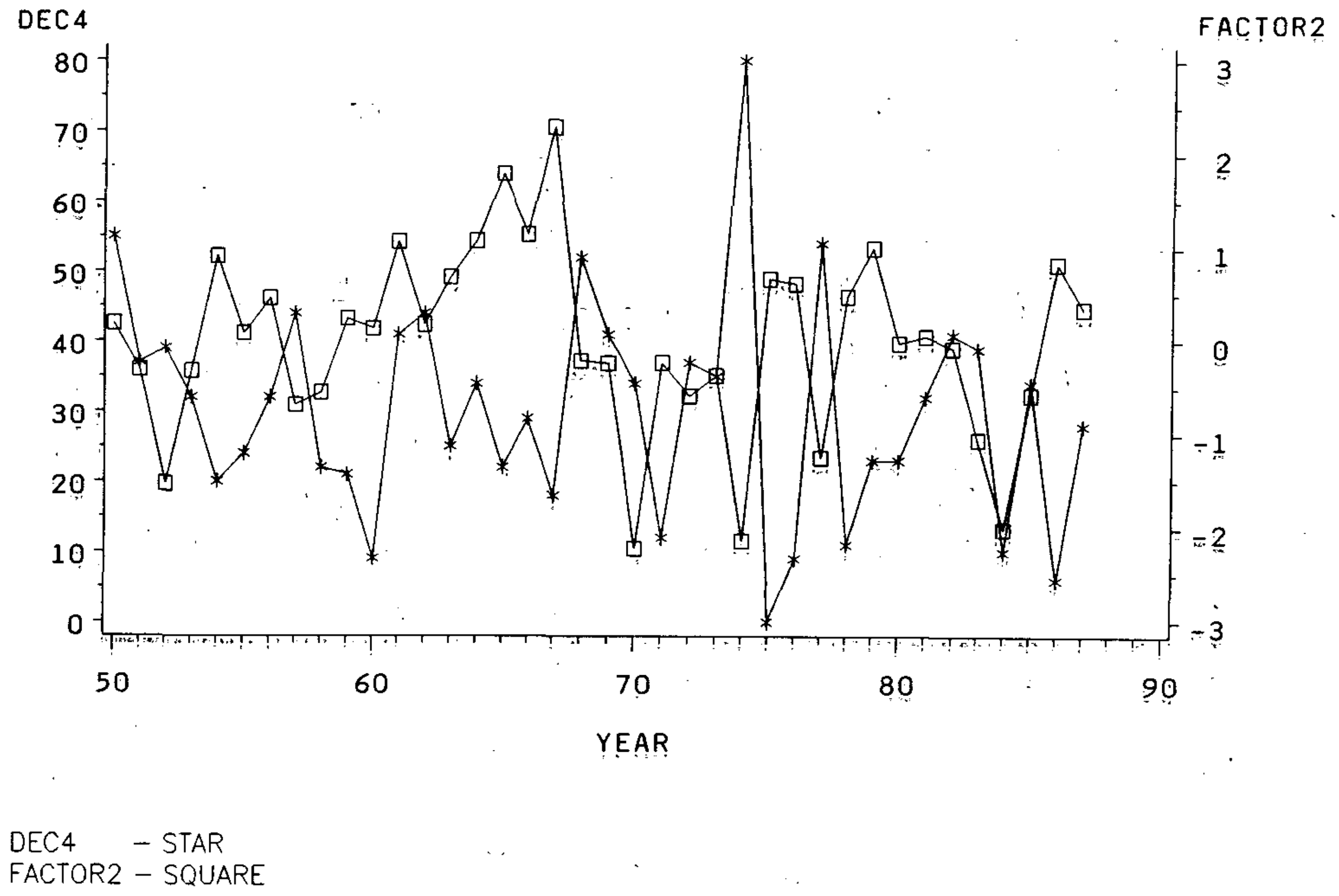
DEC2 - FACTOR4



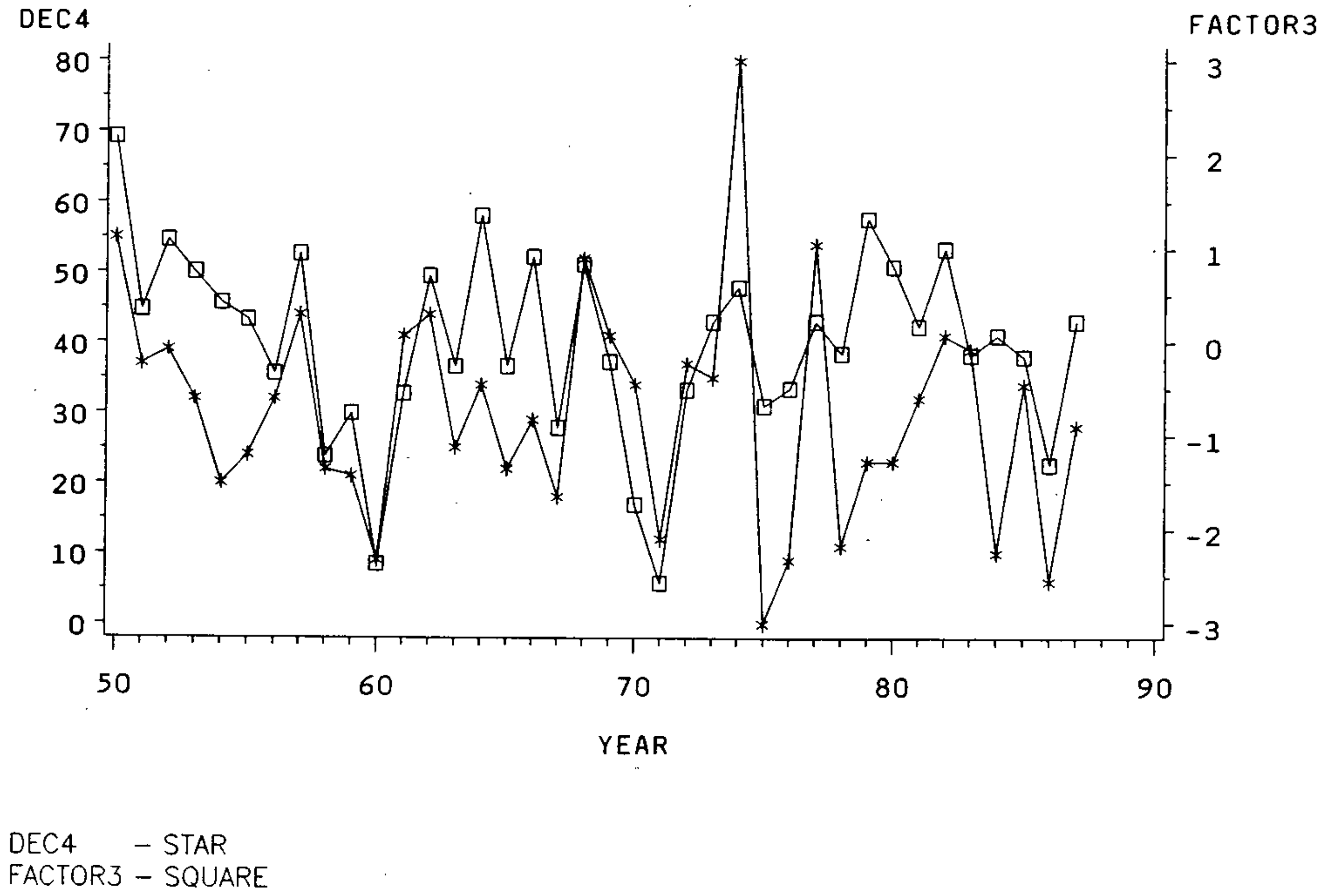
DEC3 - FACTOR4



DEC4 - FACTOR2



DEC4 - FACTOR3


DEC4
FACTOR3
YEAR
DEC4 - STAR
FACTOR3 - SQUARE