On a Differentiability of Four Subregions of the New York Bight Based on 1980 Summer Monitoring Data

by

Geung-Ho Kim, John E. O'Reilly and Robert N. Reid

U. S. Department of Commerce
National Oceanic and Atmospheric Administration
National Marine Fisheries Service
Northeast Fisheries Center
Sandy Hook Laboratory
Highlands, New Jersey 07732

Report No. SHL-83-01

(January 1983)

On a Differentiability of Four Subregions of the New York Bight Based on 1980 Summer Monitoring Data

A major problem associated with monitoring pollution-related alterations in the sediment composition and benthic communities in the apex of the New York Bight is to separate the pollution effects resulting from contaminant inputs via sewage sludge, dredge spoils, the Hudson-Raritan estuarine plume and other sources. Proper understanding of the respective characteristics and distribution of these pollutant inputs, as well as of possible relationships among them should be instrumental to sound waste management decisions. As a preliminary step toward this sort of management support, a portion of the sediment monitoring data collected in the New York Bight during July-August, 1980 has been analyzed using various multivariate statistical techniques. See the NEMP Report edited by Robert Reid et al. (1982, Ref. 1) for details of the data base.

The main objective of this exploratory analysis is to clarify the nature of the explanatory power evidenced in some sediment constituents, individually or collectively, toward discrimination of sediments from some suitably chosen subregions of the New York Bight. This is done in two stages: first partition into four regions including sewage sludge-affected region, dredge spoil-affected region, region peripheral to dump sites, and the rest of the Bight, then evaluate the explanatory power of the variables involved. The sediment variables subjected to analysis include six metals (Cd, Cr, Cu, Ni, Pb, and Zn), two sediment size indicators (% silt, and % clay), and four organic compounds (Total Organic Carbon, Total Kjeldahl Nitrogen, Polychlorinated Biphenyls and Coprostanol).

After elimination of those stations (or observations) containing missing entries, the total sediment data from 38 stations was available for analyses. In

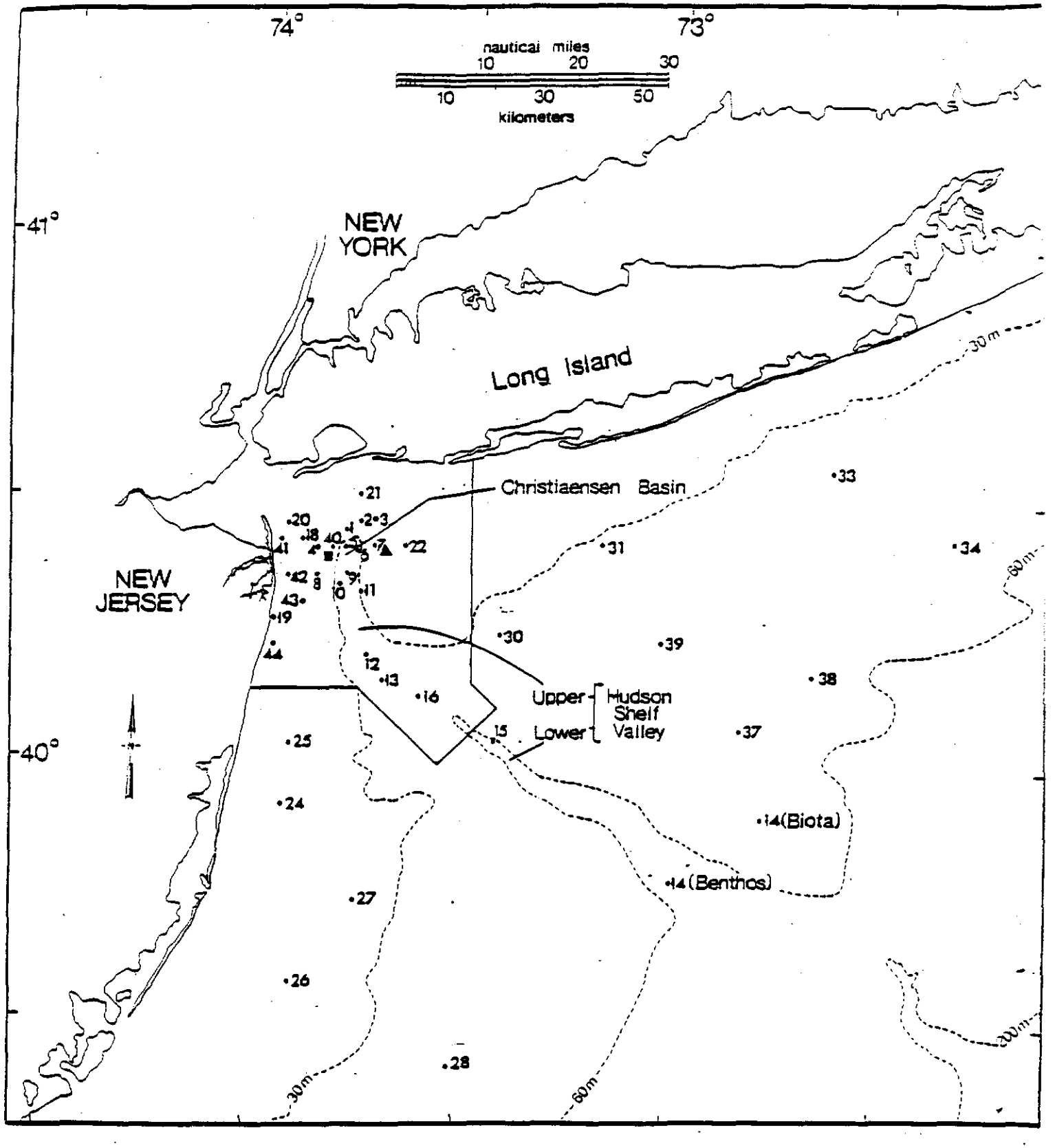
74°
73°
nautical miles
10 20 30
10 30 50
kilometers
41°
NEW YORK
Long Island
30m
Christiaensen Basin
NEW JERSEY
Upper
Lower
Hudson Shelf Valley
40°
14(Biota)
14(Benthos)
60m
200m

Figure 1

view of this rather small total number of stations, each subset of stations (see Figure 1) allocated to a subregion is typically too few to check the general conformity of the data with respect to underlying statistical assumptions, on which the development and interpretations of the analytic results are based. This is one of the reasons why we use the term "exploratory" in the preceeding paragraph, and the main emphasis here is to make a preliminary presentation of some main features of the data, without dwelling unduly on the theoretical aspect of assessing the reliability of various significance statements to be made.

The multivariate techniques routinely employed for this purpose consist of principal components analysis (PCA), canonical correlation analysis (CCA), and discriminant analysis (DA). See Green (1979), and Orloci et al. (1979) for details of these techniques. PCA is a tool for dimensional reduction of multivariate data sets, and corresponding graphic representation of all the data points in a properly reduced number of dimensions usually reveals how to produce a reasonable partition. Checking the validity of any partition formulated in this step is done using CCA.

The form of CCA used here is mathematically equivalent to the multivariate analysis of variance test which is designed to assess the combined strength of the conjectured grouping obtained from PCA. Typically the ratio

$$\frac{\text{Variability due to the grouping structure}}{\text{Variability due to the random error}}$$

is computed using the data values and, if this ratio is unreasonably large, then the null hypothesis of no grouping is rejected and, at the same time, the alternative hypothesis of significant group difference is entertained. CCA actually tells us more than this. When the null hypothesis is rejected, the analysis can provide somewhat detailed information on what are the most significant contrasts among subregions (for example, in the present case of 4

subregions six different pairwise comparisons can be made), and what are the subsets of variables most responsible for the regional differences.

Finally, DA is a tool employed to understand the effectiveness of each variable toward discrimination. Our approach here is to graphically assess the incremental contribution of each individual variable toward separation of the subregions sorted out previously. It begins with an initial two-dimensional plot of all the stations in terms of two variables, % silt and coprostanol, which are declared most discriminatory by a certain stepwise DA procedure. The pattern of separation shown in this initial plot is expected to be improved upon addition of other variables not in the analysis one by one stepwise. Although this rather qualitative graphical procedure could be backed up by some numerical estimates called F-statistics of pairwise comparisons, the probabilistic interpretation of these numbers is not easy to justify (in view of the small samples and other untested statistical properties of the data).

Recapitulating the main points, we summarize the operational steps to be followed:

I) Formulation: Initital simplification of the data to formulate a reasonable hypothesis on grouping of stations. (using PCA).

II) Confirmation: Test of the strength of the hypothesized grouping, and identification of variables responsible for the regional differences in sediment composition. (CCA)

III) Stepwise examination: Graphical demonstration of an individual variable's incremental discriminatory power. This is to be shown in a sequential manner. (DA)

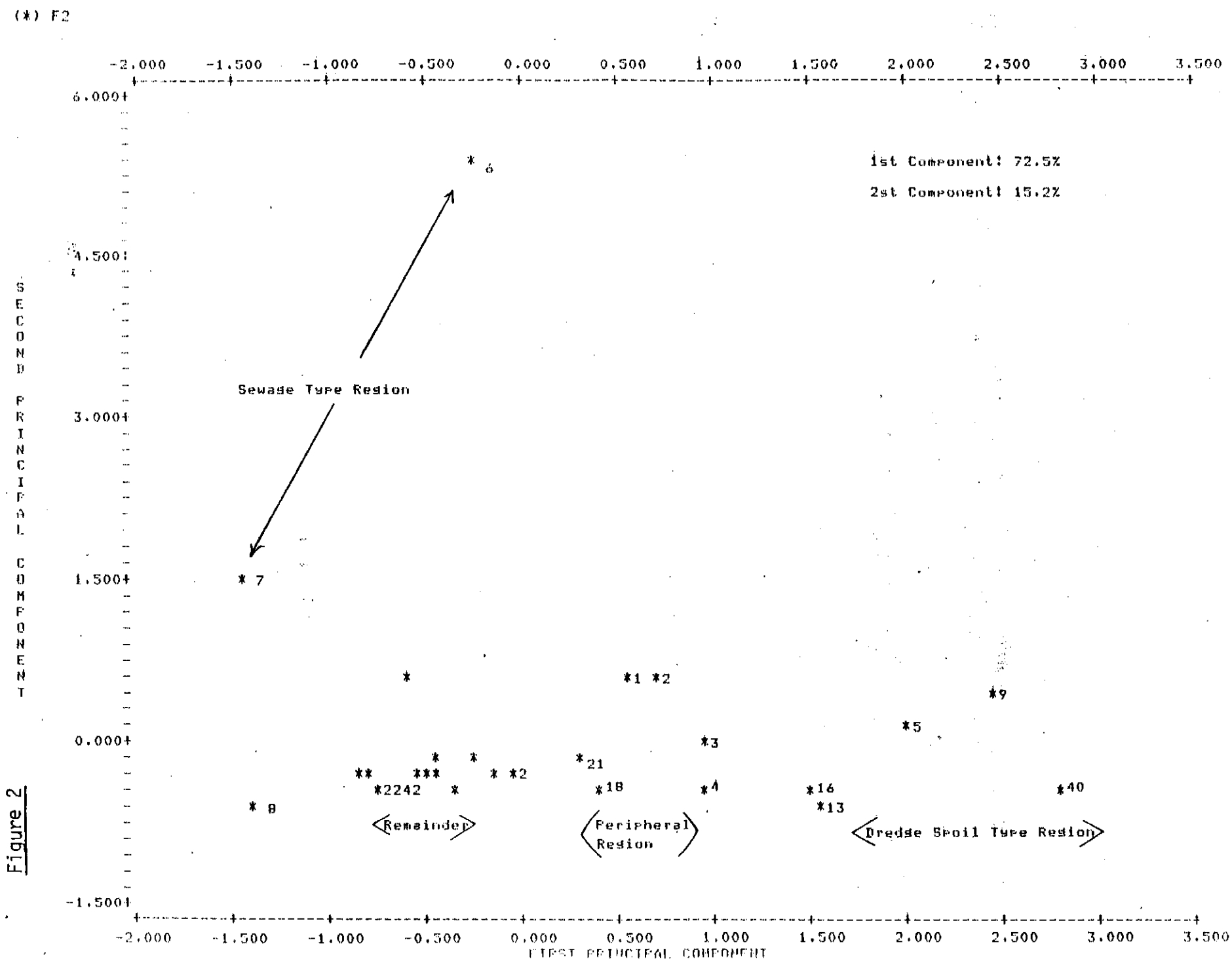
RESULT OF PRINCIPAL COMPONENT ANALYSIS
(*) F2
-2.000 -1.500 -1.000 -0.500 0.000 0.500 1.000 1.500 2.000 2.500 3.000 3.500
6.000
4.500
3.000
1.500
0.000
-1.500
SECOND PRINCIPAL COMPONENT
1st Component: 72.5%
2st Component: 15.2%
Sewage Type Region
Remainder
Peripheral Region
Dredge Spoil Type Region
FIRST PRINCIPAL COMPONENT

Figure 2

Let us look at these steps one by one.

(I) Formulation

The PCA representation of 38 New York Bight stations based on the 12 variables mentioned above is given by Figure 2. Based on this graph as well as consideration of geographical proximity of stations (in Figure 1), it is plausible to partition the set of stations into four subregions:

D. Dredge spoil type region (5 stations): 5, 9, 40, 13, 16

S. Sewage sludge type region (2 stations): 6, 7

P. Region peripheral to D and S (6 stations): 1, 2, 3, 4, 18, 21

R. Remainder of the bight (24 stations).

Station 8 is somewhat isolated from the nearby major clusters in Figure 1. It could be classified into sewage region or remainder according to the graph, although it is closer to dredge region geographically. It seems reasonable to treat this station as an outlier to R and to exclude this from the subsequent analyses, in view of its exploratory nature.

(II) Confirmation of grouping

This step is to test the strength or adequacy of the grouping conjectured above, and, if these groups are judged to be significantly different, proceed to identify those variables that can explain the group differences. CCA was carried out and the corresponding result is summarized in the following factor matrix table (Table 1). Some essential features of this table are itemized below.

a) Each one of the three canonical variates R_1, R_2, and R_3 represents a suitably chosen linear contrast of the 4 monitoring subregions. Similarly, each one of the three canonical variates V_1, V_2, and V_3 represents a suitably weighted linear combination of the 12 monitoring variables.

b) The first two canonical pairs (R_1, V_1) and(R_2, V_2) have highly significant correlation coefficients of 0.991 and 0.972 respectively. This

Table 1

Factor Matrix

Canonical variates / Regions	R_1	R_2	R_3	communality
D) Dredge dump	0.812	-0.282	0.510	1.0
S) Sewage dump	-0.401	-0.915	-0.052	1.0
P) Peripherals	0.267	0.566	-0.834	1.0
R) Remainders	-0.678	0.631	0.376	1.0
% Variance	42.7%	43.8%	13.5%	100.0%
Redundancy	41.9%	41.4%	7.7%	91.1%
Canonical corr.	0.991	0.972	0.756	
Canonical variates / Monitoring variables	V_1	V_2	V_3	communality
1.Cd	0.319	-0.692	-0.148	0.603
2.Cr	0.602	-0.535	-0.202	0.689
3.Cu	0.391	-0.672	-0.205	0.646
4.Ni	0.735	-0.446	-0.032	0.741
5.Pb	0.496	-0.669	-0.201	0.734
6.Zn	0.442	-0.672	-0.170	0.675
7.%Silt	0.929	-0.285	0.034	0.945
8.%Clay	0.460	-0.146	-0.240	0.290
9.TOC	0.757	-0.450	-0.246	0.836
10.TKN	0.790	-0.390	-0.087	0.784
11.PCB	0.506	-0.512	-0.110	0.530
12.COP	-0.294	-0.885	-0.130	0.887
% Variance	35.1%	31.8%	2.8%	69.7%
Redundancy	34.5%	30.1%	1.6%	66.2%

** Each number indicates the correlation between the original variable and the corresponding canonical variate.

implies that the hypothesis of no difference among regions should be rejected, i.e., the four regions are significantly different. The third canonical pair (R_3, V_3), with r = 0.756, is moderately significant, too. However, since the proportion of variability carried by the third pair is less than one tenth of either of the first two pairs (i.e., 2.8% vs. 35.1% and 31.8%), we can conclude that the associated explaining power is insignificant.

c) The first pair (R_1, V_1) tells us that the major contrast is between D region and R region and that this contrast is accounted for mainlyby four variables; % silt (r = .929), TKN (r = .790), TOC (r = .757), and Ni (r = .735). Cr, PCB, and Pb appear to be variables of secondary importance.

d) The second pair (R_2,V_2) explains the difference between S Region and the combined region of R and P in terms of seven variables; Coprostanol (r = .885), PCB (r = .512), and five metal variables ($-.692 \leq r \leq -.535$). Coprostanol turned out to be the strongest indication of the sewage sludge. This is reasonable, since coprostanol is a sterol indicative of mammalian feces, and has been used to identify sludge in other studies (Hatcher and McGillivary 1979; Boehm 1980). The fact that Cr, Pb and PCB are represented in both pairs with considerable correlations indicates that these variables are commonly high in both D and S regions.

E) The third pair (R_3, V_3) tells us that P region is distinguished by % clay primarily and by TOC secondarily. As alluded earlier, this third relationship pair does not carry a significant portion of the overall variability of the monitoring variable set.

As a summary, relative positioning of the above four groups of stations in a two-dimensional space is depicted in Figure 3. The essential feature of the region-wise comparison specified by the first canonical pair is illustrated along the horizontal axis, while the same is true for the second canonical pair along

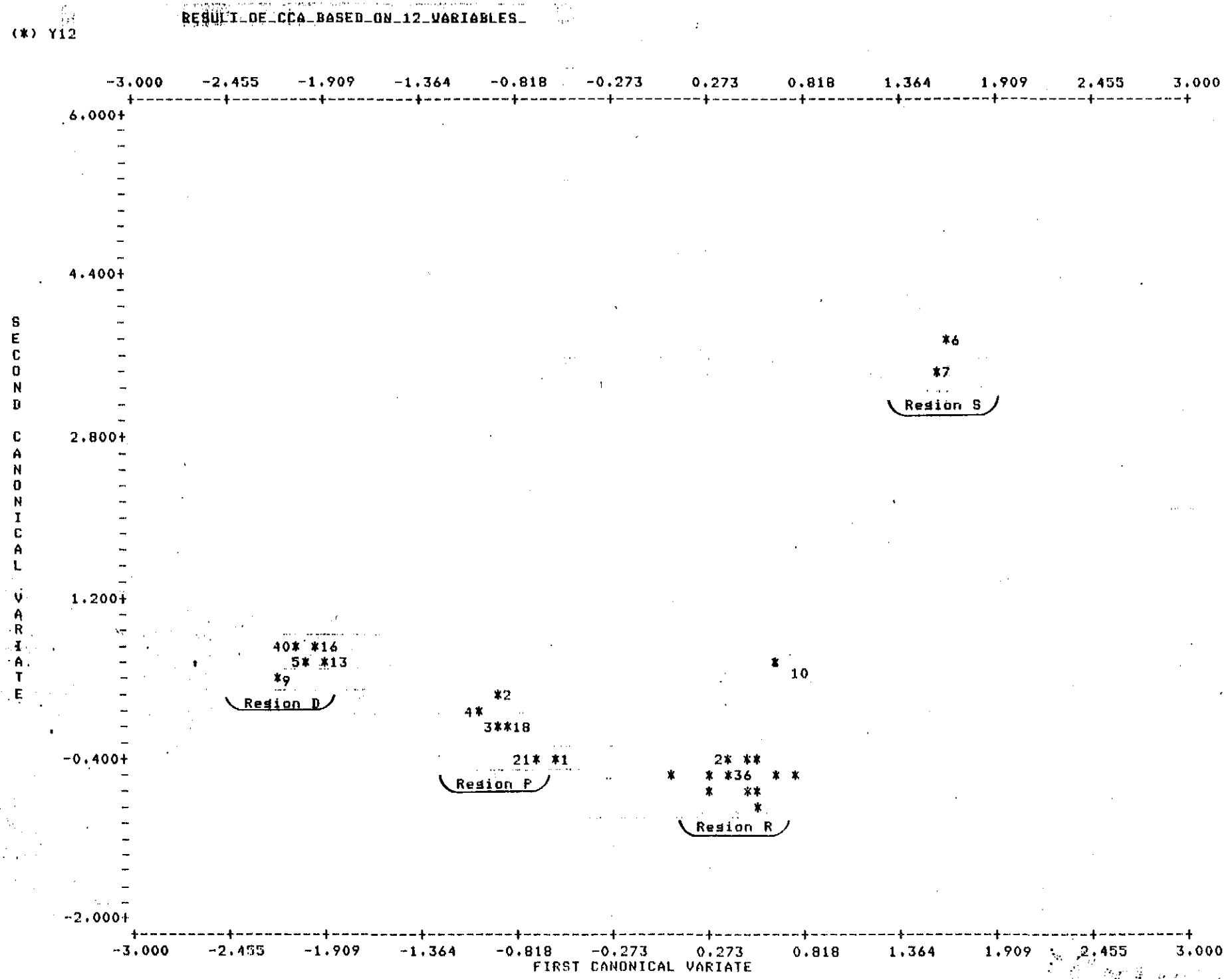
RESULT OF CCA BASED ON 12 VARIABLES
(*) Y12
-3.000 -2.455 -1.909 -1.364 -0.818 -0.273 0.273 0.818 1.364 1.909 2.455 3.000
6.000
4.400
2.800
1.200
-0.400
-2.000
SECOND CANONICAL VARIATE
*6
*7
Region S
40* *16
5* *13
*9
Region D
*2
4*
3**18
21* *1
Region P
10
2* **
* * *36 * *
* **
*
Region R
FIRST CANONICAL VARIATE

Figure 3

the vertical axis. Addition to the third axis explaining the third canonical pair appears to be needless in view of the poor accountability mentioned above.

(III) Stepwise examination

The question relevant to this step may be phrased as: How can one assess the incremental discriminatory power carried by each individual variable? It implicitly assumes a sequential introduction of the variables of interest instead of a simultaneous introduction of all variables, upon which the discussion of the preceding steps is based. Since a comprehensive and rigorous treatment of the question requires some detailed verification of statistical properties of the data, such a treatment is not intended in this brief report. Rather, we proceed to generate some qualitative solution based on simple graphical comparisons. Namely, two successive graphs, one plotted before addition of a specified variable and the other plotted after addition of the same variable, are to be compared and on the basis of this comparison of two possibly different partitioning patterns, we hope to say something about the effectiveness of the variable.

In this connection, it is customary to specify a sequence according to which variables are to be introduced to the analysis one at a time. After application of a forward stepwise discriminant analysis based on the Wilks' distance statistic, the resultant sequence is; % silt, coprostanol, % clay, TKN, TOC, Cu, PCB, Cr, Zn, Pb, Ni, and Cd. (Note that ordering of variables via DA can be affected by various factors. These are the choice of the distance statistics, the specification of the threshold values for inclusion and exclusion of a new target variable, and whether forward or backward procedure is used.) Then, we proceed with the two-dimensional plot (Figure 4) depicting the result of CCA done with the first two variables. Here X-axis represents the silt level, while the

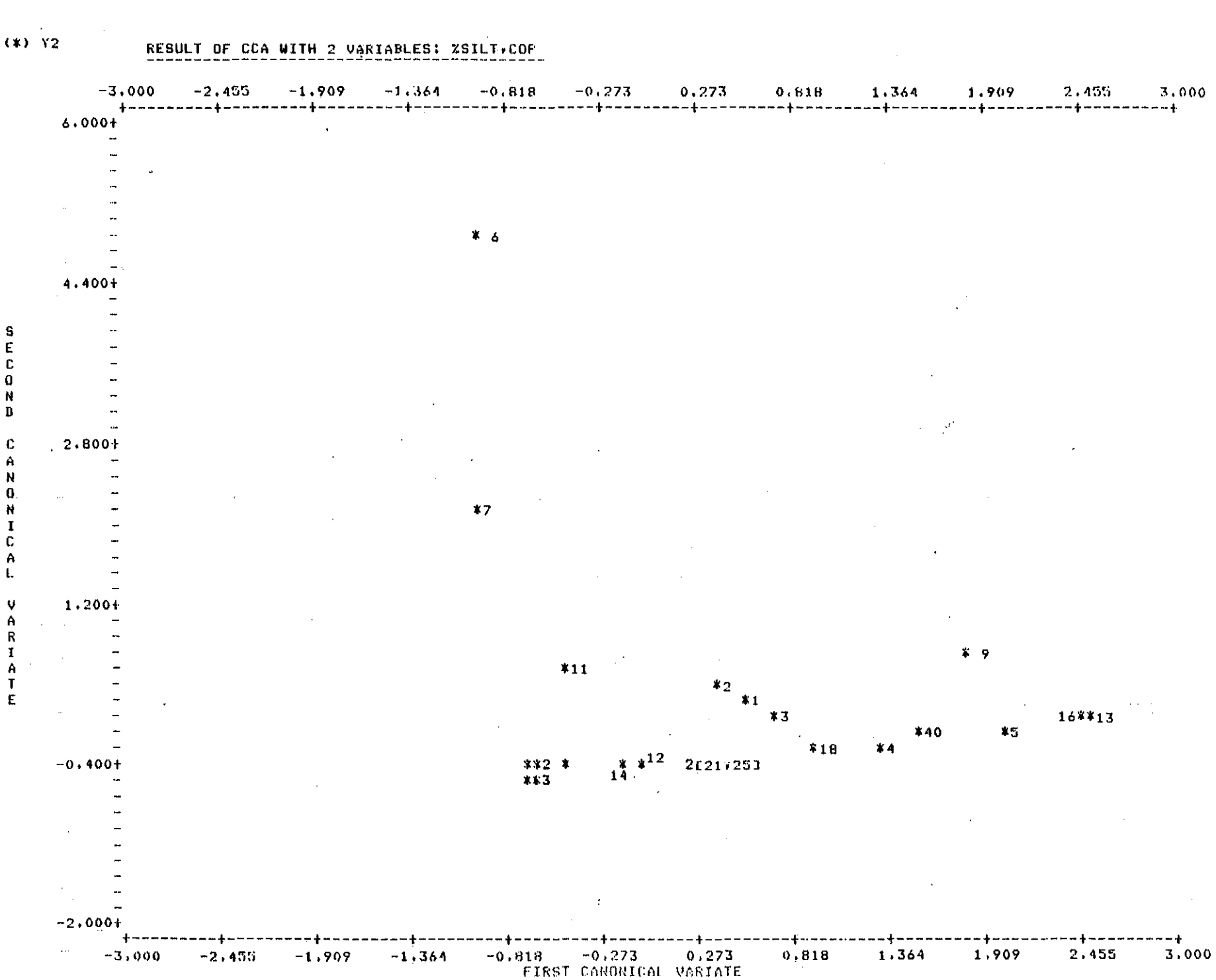
(*) Y2
RESULT OF CCA WITH 2 VARIABLES: %SILT,COF
-3.000 -2.455 -1.909 -1.364 -0.818 -0.273 0.273 0.818 1.364 1.909 2.455 3.000
6.000
4.400
2.800
1.200
-0.400
-2.000
SECOND CANONICAL VARIATE
FIRST CANONICAL VARIATE
6
7
11
2
1
3
9
16
13
40
5
18
4
12
14
2[21,25]

Figure 4

(*) Y5

RESULT OF CCA WITH 5 VARIABLES:%SILT,COP,%CLAY,TKN,TOC

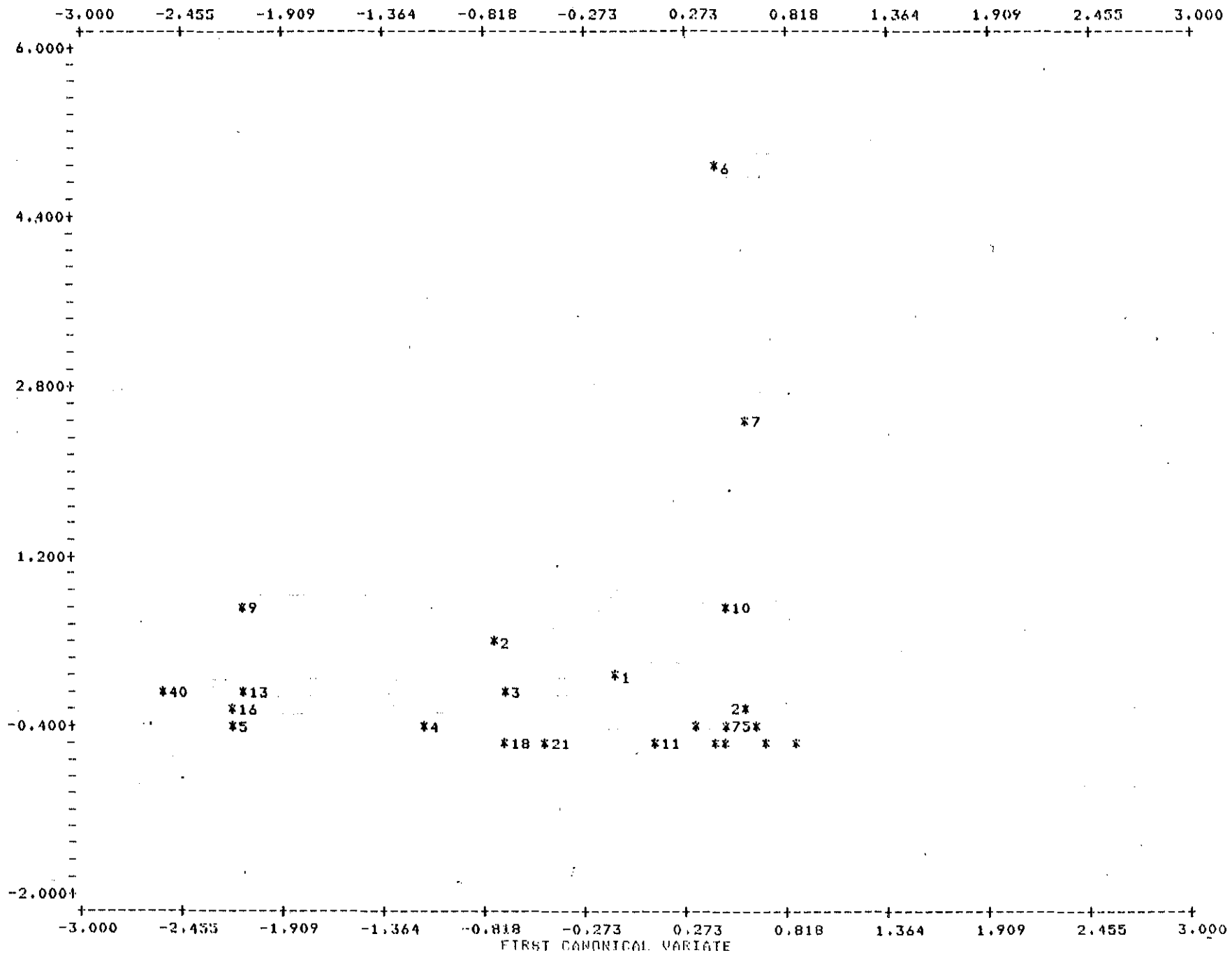

Figure 5

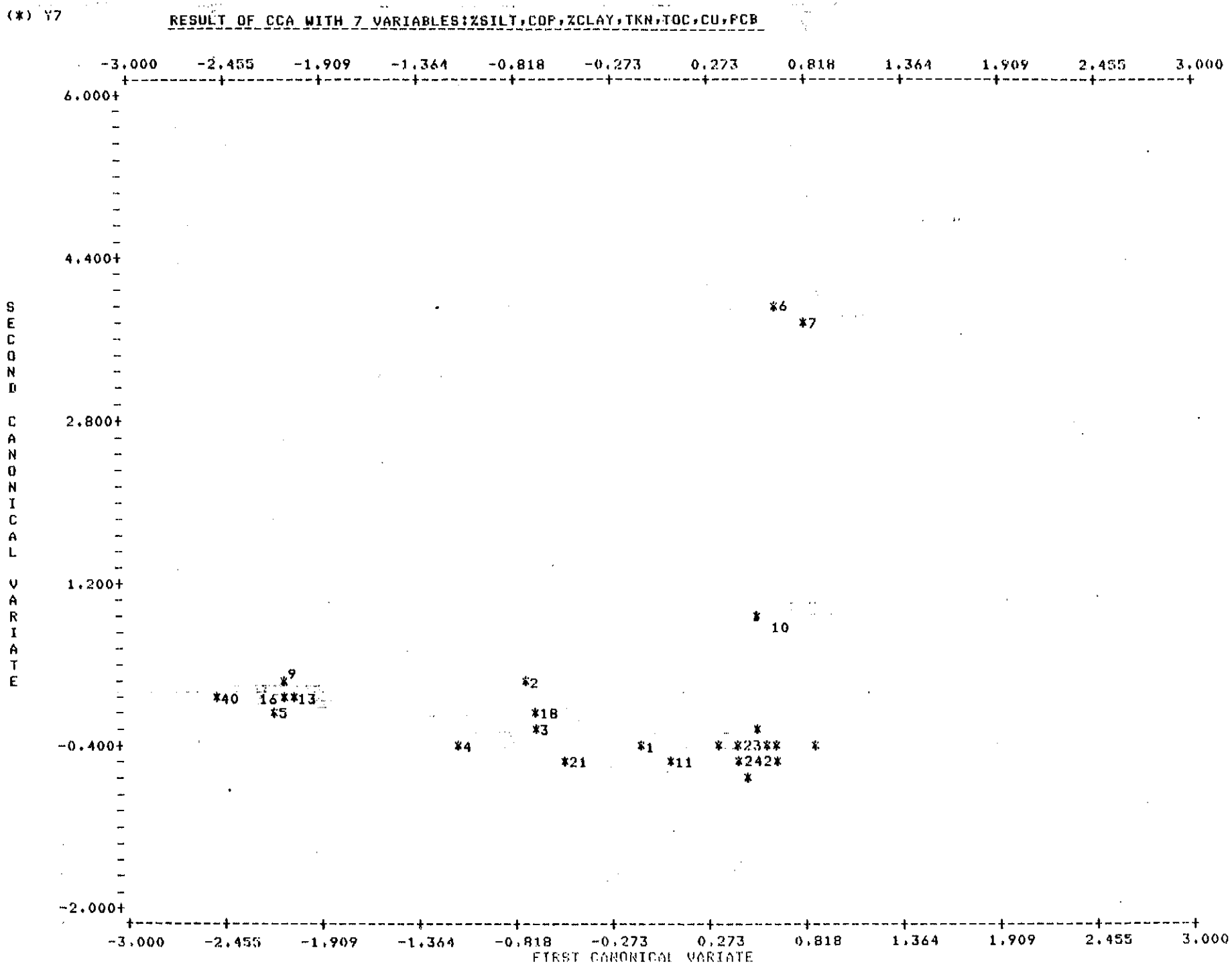
(*) Y7
RESULT OF CCA WITH 7 VARIABLES:%SILT,COP,%CLAY,TKN,TOC,CU,PCB
-3.000 -2.455 -1.909 -1.364 -0.818 -0.273 0.273 0.818 1.364 1.909 2.455 3.000
6.000
4.400
2.800
1.200
-0.400
-2.000
SECOND CANONICAL VARIATE
FIRST CANONICAL VARIATE
*6
*7
10
*2
*9
*40
16
*13
*5
*18
*3
*4
*1
*21
*11
*23
*242

Figure 6

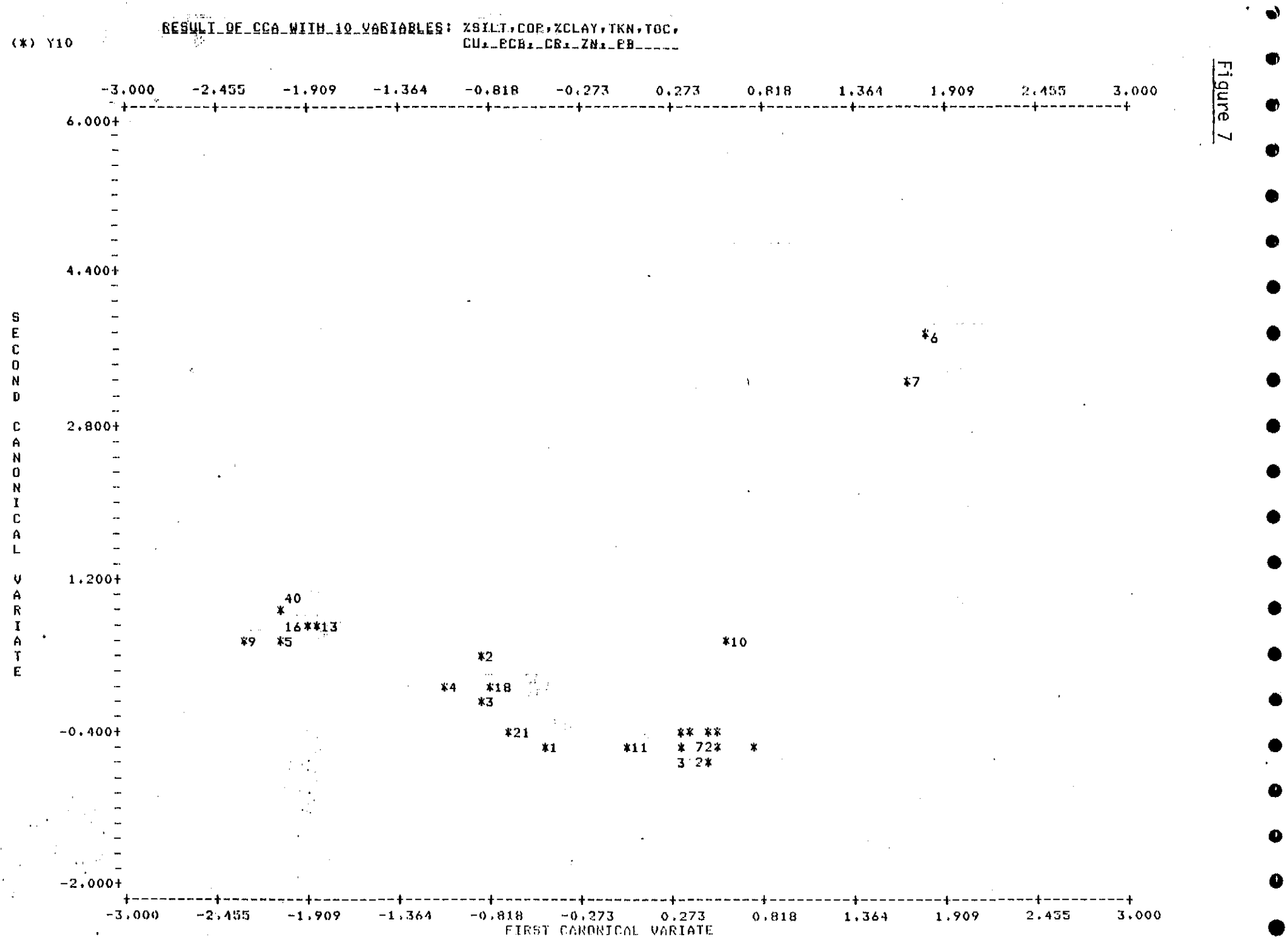
RESULT OF CCA WITH 10 VARIABLES: %SILT,COP,%CLAY,TKN,TOC,
CU,_ECB,_CB,_ZN,_EB_____
(*) Y10
-3.000 -2.455 -1.909 -1.364 -0.818 -0.273 0.273 0.818 1.364 1.909 2.455 3.000
6.000
4.400
2.800
1.200
-0.400
-2.000
SECOND CANONICAL VARIATE
FIRST CANONICAL VARIATE
40
16
13
9
5
2
4
18
3
21
1
11
10
6
7

Figure 7

Y-axis represents the coprostanol level. It does not appear to present a clear pattern of separation, except the two sewage dump stations.

1) Addition of % Clay

Change in separation pattern is not noticeable except the slight reduction of within-variability of the region R (i.e., resulting in a higher concentration within R). This fact agrees with the CCA result stating that the variable clay is mainly responsible for the third canonical pair which is not included in the graphic analysis.

2) Addition of TKN and TOC (Figure 5)

Separation in terms of the X-axis is improved, in particular, between regions D and R, and between regions D and P. This confirms the CCA results that the variables are contributing to the X-axis (or the first canonical axis). Note that the overall variability is still great along the Y-axis.

3) Addition of Cu

This is a Y-axis variable according to CCA. The variability along the Y-axis of respective regions is generally reduced except S region. This reduction results in better separation between regions S and R, and between S and P.

4) Addition of PCB (Figure 6)

Respective within-region variabilities of the two dump-related regions D and S are greatly reduced, and separation of these from the R becomes more vivid. This accords with the previous result that PCB is a sediment variable prevalent in both dump-related regions.

5)Addition of Cr

Like PCB, Cr is indicative of both types of dump regions. Interesting features are: a) difference between regions S and R in terms of X-axis starts appearing. b) Difference between regions D and R in terms of Y-axis start appearing. Therefore, regions D and S are getting closer in terms of Y-axis. In

other words, regions D and S become less distinguishable by addition of this metal. The next few graphs show us more of this general tendency by adding more metals.

6) Addition of Zn, Pb (Figure 7)

In terms of X-axis, separation of S region from R is improved, i.e., contribution of these metals to the X-axis helps to differentiate the sewage region. However, contribution of these metals to the Y-axis helps reduce the distance between the two types of dump regions.

7) Addition of Ni

This contributes to the X-axis primarily, and Y-axis secondarily, and results in improved separation along the X-axis.

8) Addition of Cd (Figure 3)

Shows more reduction in within-region variability for all groups. Incremental contribution is not outstanding mainly because Cd is highly correlated with all the metals previously included in the analysis.

As a tentative conclusion, we may say that the combined strength of all 12 monitoring variables can sufficiently differentiate the four regions in question. As far as the difference between the two types of dump related regions is concerned, Silt and Coprostanol appear to be most effective. TKN, TOC, and Ni may be of secondary importance. Heavy metals like Cu, Cr, Zn, and Pb seems ineffective toward discrimination of the two because of their equally high prevalence in both regions.

In view of our initial remarks on the small sample size per group and other unverifiable statistical properties of the data, it should be mentioned that some of the significance statements and the corresponding conclusions above are less

reliable than they should be. Another conservative consequence to the remarks is that the conclusions drawn above should be construed as peculiarities of the given data set, and therefore extrapolation of the result to other similar data sets should be treated carefully.

References

1. Northeast Monitoring Program (NEMP). 1982.
   Contaminants in New York Bight and Long Island Sound Sediments and Demersal Species, and Contaminant Effects on Benthos, Summer 1980. NOAA Tech. Memo. NMFS-F/NEC-16.

2. Green, R. H. 1979.
   Sampling Design and Statistical Methods for Environmental Biologists. John Wiley and Sons, New York, N.Y.

3. Orloci, L., C. R.Rao, and W. M.Stiteler (eds.). 1979.
   Multivariate Methods in Ecological Work. International Co-operative Publishing House, Fairland, Maryland.

4. Hatcher, P. G., and P. A. McGillivary. 1979.
   Sewage contamination in the New York Bight. Coprostanol as an indicator. Envir. Sci. Technol. 13: 1225-1229.

5. Boehm, P. 1980.
   New York Bight benthic sampling survey: Coprostanol, polychlorinated biphenyl and polynuclear aromatic hydrocarbons measurements in sediments. NOAA, Northeast Monitoring Program Report No. NEMP-III-80-B-0046.