Environmental Applications of Chemometrics

EDITED BY
Joseph J. Breen
Philip E. Robinson

ACS Symposium Series
292

ACS SYMPOSIUM SERIES 292

Environmental Applications of Chemometrics

Joseph J. Breen, EDITOR
Office of Toxic Substances
U.S. Environmental Protection Agency

Philip E. Robinson, EDITOR
Office of Toxic Substances
U.S. Environmental Protection Agency

Developed from a symposium sponsored by
the Division of Environmental Chemistry
at the 188th Meeting
of the American Chemical Society,
Philadelphia, Pennsylvania,
August 26–31, 1984

American Chemical Society, Washington, D.C. 1985

Library of Congress Cataloging in Publication Data

Environmental applications of chemometrics.
(ACS symposium series, ISSN 0097-6156; 292)

"Developed from a symposium sponsored by the Division of Environmental Chemistry at the 188th meeting of the American Chemical Society, August 26-31, 1984," in Philadelphia, Pa.

Includes bibliographies and index.

1. Chemistry—Mathematics—Congresses.
2. Chemistry—Statistical methods—Congresses.
3. Environmental chemistry—Congresses.

I. Breen, Joseph J., 1942- . II. Robinson, Philip E., 1948- . III. American Chemical Society. Division of Environmental Chemistry. IV. American Chemical Society. Meeting (188th: 1984: Philadelphia, Pa.) V. Series.

QD39.3.M3E58 1985 628.5'028 85-22878
ISBN 0-8412-0945-6

PRINTED IN THE UNITED STATES OF AMERICA

ACS Symposium Series

M. Joan Comstock, *Series Editor*

Advisory Board

FOREWORD

The ACS SYMPOSIUM SERIES was founded in 1974 to provide a medium for publishing symposia quickly in book form. The format of the Series parallels that of the continuing ADVANCES IN CHEMISTRY SERIES except that, in order to save time, the papers are not typeset but are reproduced as they are submitted by the authors in camera-ready form. Papers are reviewed under the supervision of the Editors with the assistance of the Series Advisory Board and are selected to maintain the integrity of the symposia; however, verbatim reproductions of previously published papers are not accepted. Both reviews and reports of research are acceptable, because symposia may embrace both types of presentation.

PREFACE

ENVIRONMENTAL APPLICATIONS OF CHEMOMETRICS are of interest because of the concern about the effects of chemicals on humans. The symposium upon which this book is based served as an important milestone in a process we, the editors, initiated in 1982. As members of the Environmental Protection Agency's Office of Toxic Substances (OTS), we have responsibilities for the acquisition and analysis of human and environmental exposure data in support of the Toxic Substances Control Act. OTS exposure studies invariably are complex and range from evaluating human body burden data (polychlorinated biphenyls in adipose tissue, for example) to documenting airborne asbestos levels in schools.

The proper conduct of complex exposure studies requires that the quality of the data be well defined and the statistical basis be sufficient to support rule making if necessary. These requirements, from study design through chemical analysis to data reduction and interpretation, focused our attention on the application of chemometric techniques to environmental problems.

In the fall of 1982, OTS and the Agency's Office of Research and Development's (ORD) Environmental Monitoring Systems Laboratory (Research Triangle Park, NC) hosted a 2-day workshop for researchers active in chemometrics. The participants represented various agency program offices and ORD laboratories, as well as researchers from the National Fisheries and Wildlife Service, Columbia, MO; University of Illinois, Chicago; and Infometrix, Seattle, WA. It was evident that isolated attempts were in progress to apply chemometric techniques to complex environmental problems. What was lacking was a coherent chemometrics program with well-defined objectives.

The advent of analytical techniques capable of providing data on a large number of analytes in a given specimen had necessitated that better techniques be employed in the assessment of data quality and for data interpretation. In 1983 and 1984, several volumes were published on the application of pattern recognition, cluster analysis, and factor analysis to analytical chemistry. These treatises provided the theoretical basis by which to analyze these environmentally related data. The coupling of multivariate approaches to environmental problems was yet to be accomplished.

This multivariate data analysis challenge is aggressively being met by a number of researchers. The result is a vibrant and growing literature filled with software acronyms such as ARTHUR, SIMCA, CHEOPS, CLEOPATRA,

EIN*SIGHT, and others. All of these programs are specifically directed toward the multivariate analysis of analytical chemical data both in assessing data quality (quality control and quality assurance) and in interpreting data to provide insight into the complex system under investigation.

The fall of 1983 also saw the North Atlantic Treaty Organization host an Advanced Studies Institute in Cosenza, Italy, entitled "Chemometrics: Mathematics and Statistics in Chemistry." One hundred scientists—a most unusual collection of chemists, engineers, and statisticians from academia, industry, and government—representing a dozen countries assembled to discuss the role of sophisticated multivariate statistics in the daily routine of an analytical chemistry laboratory.

With this backdrop, we approached the ACS Division of Environmental Chemistry with the request to sponsor a symposium on the application of chemometrics to environmental problems.

This volume represents a majority of the presentations made at the symposium. The broad range of topics can be seen in the table of contents. Thought-provoking discussions at the symposium revealed that significant progress has been made in the application of chemometrics to environmental problems.

DISCLAIMER

This book was edited by Joseph J. Breen and Philip E. Robinson in their private capacity. No official support or endorsement by the U.S. Environmental Protection Agency is intended or should be inferred.

JOSEPH J. BREEN
PHILIP E. ROBINSON
Office of Toxic Substances
U.S. Environmental Protection Agency
Washington, DC 20460

CONTENTS

Preface ix

1. **Soft Independent Method of Class Analogy: Use in Characterizing Complex Mixtures and Environmental Residues of Polychlorinated Biphenyls** 1
   D. L. Stalling, T. R. Schwartz, W. J. Dunn III, and J. D. Petty

2. **Evaluating Data Quality in Large Data Bases Using Pattern-Recognition Techniques** 16
   Robert R. Meglen and Robert J. Sistko

3. **Exploratory Data Analysis of Rainwater Composition** 34
   Richard J. Vong, Ildiko E. Frank, Robert J. Charlson, and Bruce R. Kowalski

4. **Multivariate Analysis of Electron Microprobe–Energy Dispersive X-ray Chemical Element Spectra for Quantitative Mineralogical Analysis of Oil Shales** 53
   Lawrence E. Wangen, Eugene J. Peterson, William B. Hutchinson, and Leonard S. Levinson

5. **Application of Pattern Recognition to High-Resolution Gas Chromatographic Data Obtained from an Environmental Survey** 69
   John M. Hosenfeld and Karin M. Bauer

6. **Quality Assurance Applications of Pattern Recognition to Human Monitoring Data** 83
   Philip E. Robinson, Joseph J. Breen, and Janet C. Remmers

7. **Description of Air Pollution by Means of Pattern Recognition Employing the ARTHUR Program** 93
   F. W. Pijpers

8. **Application of Soft Independent Modeling of Class Analogy Pattern Recognition to Air Pollutant Analytical Data** 106
   Donald R. Scott

9. **Cluster Analysis of Chemical Compositions of Individual Atmospheric Particles Data** 118
   T. W. Shattuck, M. S. Germani, and P. R. Buseck

10. **Monitoring Polycyclic Aromatic Hydrocarbons: An Environmental Application of Fuzzy C-Varieties Pattern Recognition** 130
    R. W. Gunderson and K. Thrane

11. **Applications of Molecular Connectivity Indexes and Multivariate Analysis in Environmental Chemistry** 148
    Gerald J. Niemi, Ronald R. Regal, and Gilman D. Veith

12. **Pattern Recognition of Fourier Transform IR Spectra of Organic Compounds** 160
    Donald S. Frankel

13. **Use of Composited Samples To Increase the Precision and Probability of Detection of Toxic Chemicals** 174
Gregory A. Mack and Philip E. Robinson

14. **The Alpha and Beta of Chemometrics** 184
George T. Flatman and James W. Mullins

15. **Statistical Issues in Measuring Airborne Asbestos Levels Following an Abatement Program** 191
Jean Chesson, Bertram P. Price, Cindy R. Stroup, and Joseph J. Breen

16. **Estimation of Spatial Patterns and Inventories of Environmental Contaminants Using Kriging** 203
Jeanne C. Simpson

17. **Simple Modeling by Chemical Analogy Pattern Recognition** 243
W. J. Dunn III, Svante Wold, and D. L. Stalling

18. **A Quality Control Protocol for the Analytical Laboratory** 250
Robert R. Meglen

19. **Statistical Receptor Models Solved by Partial Least Squares** 271
Ildiko E. Frank and Bruce R. Kowalski

Author Index 280

Subject Index 280

Soft Independent Method of Class Analogy

Use in Characterizing Complex Mixtures and Environmental Residues of Polychlorinated Biphenyls

D. L. Stalling[1], T. R. Schwartz[1], W. J. Dunn III[2], and J. D. Petty[1]

[1]Columbia National Fisheries Research Laboratory, U.S. Fish and Wildlife Service, Columbia, MO 65201

[2]Health Sciences Research Center, Department of Medicinal Chemistry and Pharmacognosy, University of Illinois at Chicago, Chicago, IL 60612

Pattern recognition studies on complex data from capillary gas chromatographic analyses were conducted with a series of microcomputer programs based on principal components (SIMCA-3B). Principal components sample score plots provide a means to assess sample similarity. The behavior of analytes in samples can be evaluated from variable loading plots derived from principal components calculations. A complex data set was derived from isomer specific polychlorinated biphenyl (PCBS) analyses of samples from laboratory and field studies. The application of chemometrics to these problems includes three segments: analytical quality control; method and data base development; and modeling Aroclor composition and PCB residues in bird eggs.

Chemometrics, as defined by Kowalski (1), includes the application of multivariate statistical methods to the study of chemical problems. SIMCA (Soft Independent Method of Class Analogy) and other multivariate statistical methods have been used as tools in chemometric investigations. SIMCA, based on principal components, is a multivariate chemometric method that has been applied to a variety of chemical problems of varying complexity. The SIMCA-3B program is suitable for use with 8- and 16-bit microcomputers.

Four levels of pattern recognition have been defined by Albano (2). Levels I and II are most frequently used to determine the similarity of objects, or to characterize clusters of samples and to classify unknown objects. Level III takes advantage of the reduction of data dimensions resulting from SIMCA and seeks to establish a correlation of sample scores with independent variables

0097-6156/85/0292-0001$06.00/0

such as chemical functions or variables, spectroscopic data or chemical toxicity. This approach is often used in quantitative structure-activity relationships (3-5). Level IV is most frequently applied to complex spectroscopic calibration problems and in situations where composition prediction or estimation is to be made from spectroscopic data.

The SIMCA approach can be applied in all of the four levels of pattern recognition. We focus on its use to describe complex mixtures graphically, and on its utility in quality control. This approach was selected for the tasks of developing a quality control program and evaluating similarities in samples of various types. Principal components analysis has proven to be well suited for evaluating data from capillary gas chromatographic (GC) analyses (6-8).

Analytical quality control (QC) efforts usually are at level I or II. Statistical evaluation of multivariate laboratory data is often complicated because the number of dependent variables is greater than the number of samples. In evaluating quality control, the analyst seeks to establish that replicate analyses made on reference material of known composition do not contain excessive systematic or random errors of measurement. In addition, when such problems are detected, it is helpful if remedial measures can be inferred from the QC data.

Our progress in the application of chemometrics to capillary GC data was advanced by the development of a laboratory chromatography data base (9). This development followed from our decision to use capillary GC in most of our laboratory analyses for environmental contaminants. A data base was considered necessary because large amounts of data were being generated from the analysis of laboratory and field studies on complex mixtures of organochlorine contaminants. A data base is an important, but not essential, factor in using pattern recognition for quality control.

The most advanced application of pattern recognition (Level IV) offers the possibility of predicting independent variables by using latent variables derived from examining training sets of dependent and independent variables (10). The application of partial least squares in the prediction of the composition of mixtures of Aroclors was previously explored (6) by using the program, PLS-2 provided by the SIMCA-3B programs (11-12).

The first results from the use of PLS were reported by Dunn *et al* (6) who estimated the composition of PCB contaminated waste oil in terms of Aroclor mixtures. Stalling *et al* (13), who reported on the characterization of PCB mixtures and the use of three-dimensional plots derived from principal components, demonstrated that the fractional composition of TCDD and other PCDD residues were related to their geographical origins. These two reports (6,13) described the application of an advanced chemometric tool in residue studies and illustrated the

use of pattern recognition to extract quantitative information about sample similarity.

In our present investigations, we encountered a pressing need for an objective, statistically based way of evaluating concentrations of as many as 105 individual PCB isomers in each sample analyzed by capillary GC. We summarize here some of the experience obtained in our laboratories from the use of SIMCA to characterize Aroclor mixtures and environmental PCB residues in a series of bird eggs.

METHODS

Sampling. Eggs of Forster's tern (Sterna forsteri) were collected in 1983 from nests in two colonies in Wisconsin--one on Lake Poygan and the other on Oconto Marsh, Green Bay--as part of a study on impaired reproduction. Lake Poygan is a relatively clean lake whereas Green Bay is heavily contaminated from the Fox River with many industrial chemicals--particularly PCBs and chlorophenols, which are known sources of PCDFs and PCDDs. Reproductive success has declined and the incidence of deformed young has increased in the Green Bay colony (14).

Analysis of PCBs. PCB residues in extracts of egg samples were enriched by using a combination of gel permeation chromatography on BioBeads S-X3 and 1:1 (v/v) cyclohexane:methylene chloride. Adsorption column chromatography on silicic acid was used to separate PCBs from other co-extractives and contaminants (15).

The PCB congeners were separated by using a glass capillary chromatographic column (30 M x .25 mm i.d.) coated with C_{87}-hydrocarbon stationary phase (Quadrex Corp., New Haven, CT 06525); a 60-cm uncoated fused silica retention gap connected the injector to the analytical column and a 15 cm uncoated fused silica column connected analytical column to the detector. The data sampling and gas chromatography program was controlled by a Varian Autosampler Model 8000, which also delivered a calibrated amount of sample to the GC injection port. Chromatographic conditions were similar for all of the analyses: initial temperature, 80 ^{o}C, programmed at 3 ^{o}C/min to a final temperature of 265 ^{o}C; detector temperature, 320 ^{o}C; and injector temperature (direct inject) 220 ^{o}C.

An IBM CS9000 microcomputer was interfaced with the GC which acquired data generated by the electron capture detector. In processing the data, we used the CS9000 and a software package designed for laboratory data collection (Capillary Applications Program [CAP], IBM Instrument division, Danbury, CT 06810). We organized the processed peak data, using a basic program, into a series of files on hard disk media and transferred these files off-line to a Digital Equipment Corp. (DEC) PDP-11/34 minicomputer. We

then organized the data into tree-structured disk files, using our specialized laboratory data base management computer programs written in DSM-11 (Digital Standard MUMPS) for the PDP-11 family of computers.

We separated 105 constituents and achieved calibration by using a 1:1:1:1 (w/w/w/w) mixture of Aroclors 1242, 1248, 1254 and 1260. The last two digits of the Aroclor number designates the percentage chlorine in the Aroclor. A chromatogram of this mixed Aroclor standard is shown in Figure 1. The method of peak identification was a retention index system utilizing n-alkyl trichloroacetates (16). Molar response factors were determined from a flame ionization detector by using the computer-based calculation methods described by Schwartz et al. (16).

After we determined the concentrations of individual isomers, we retrieved the data from the MUMPs based laboratory data base, and transferred them to an IBM-XT (IBM Corporation, Boco Raton, FL 33432) by way of a RS-232 link, using the program Cyber (Department of Linguistics, University of Illinois at Champaign-Urbana, Urbana, IL 61820). In performing principal components analyses, we used SIMCA-3B for MS-DOS based microcomputers (Principal Data Components, 2505 Shepard Blvd., Columbia, MO 65201).

A series of Aroclors and known Aroclor mixtures were analyzed by these techniques to provide a training data set for SIMCA-3B. These standards included replicate analyses, a 1:1 (w/w) mixture of each Aroclor in combination with one other Aroclor, and a 1:1:1:1 mixture of each Aroclor (Table I).

Principal Components Analysis

We examined the data by calculating principal components sample scores (Thetas) and variable loading terms (Betas), using the program CPRIN from the SIMCA-3B programs. After calculating two or three principal components for a class model, one can prepare a plot of sample similarity, by using the sample scores (Theta-1 vs Theta-2), as well as variable loadings (Beta-1 vs Beta-2). Sample similarity was determined by calculating sample scores (θ-values, Equation [1]).

$$X_{ik} = \bar{X}_i + \sum_{a=1}^{A} \theta_{ka} \cdot B_{ai} + E_{ik} \qquad [1]$$

The likeness of samples within the class can be assessed by the proximity of samples to each other in plots derived from principal components models. The statistical technique of cross-validation (17) was used to

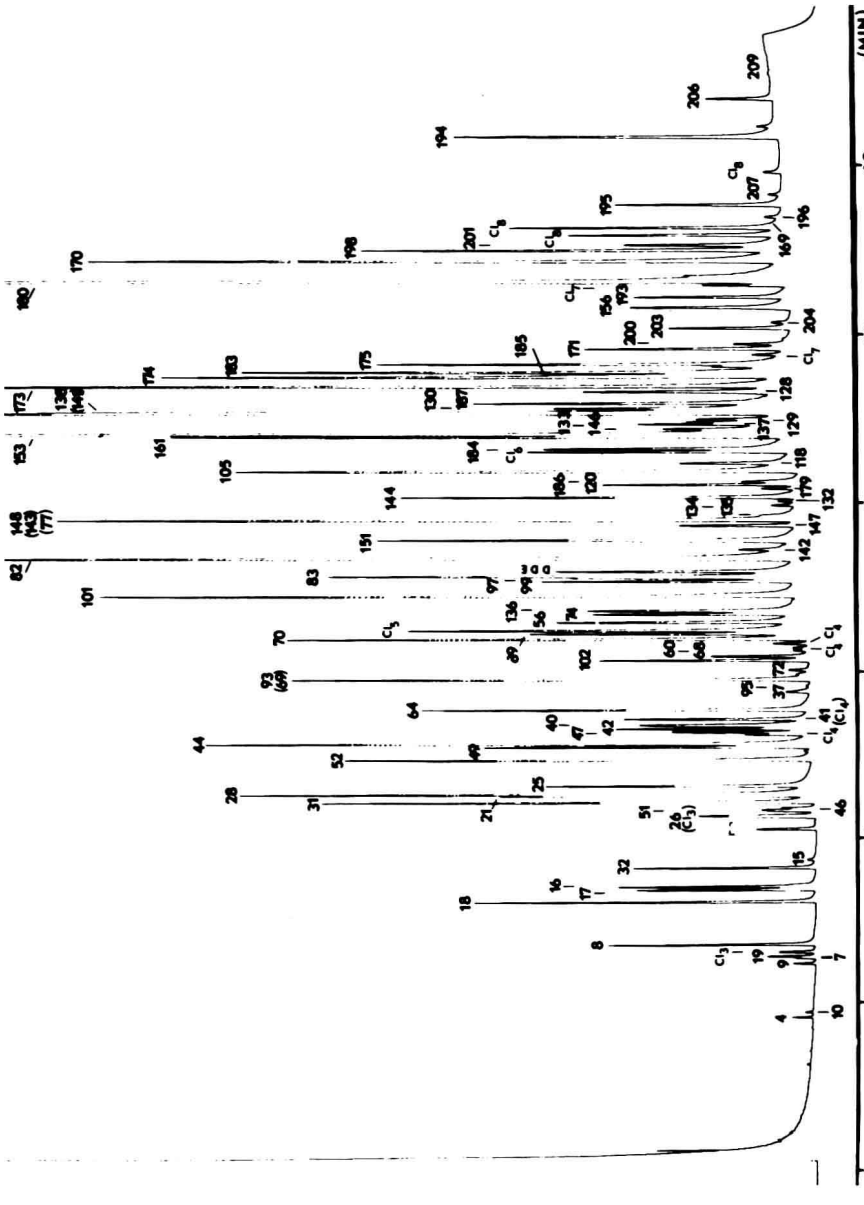

Figure 1. Gas chromatogram of a 1:1:1:1 mixture (w/w/w/w) of aroclors 1242:1248:1254:1260 (See text for chromatographic conditions.)

determine the number of components that were statistically significant.

Table I. Aroclors Samples Composing the Training Data Set.

	Aroclor Composition				
Sample #	1242	1248	1254	1260	Replicate #
1	0	0	1	0	1254-1
2	0	1	0	0	1248-1
3	0	0	0	1	1260-1
4	1	0	0	0	1242-1
5	1	0	0	0	1242-2
6	0	0	0	1	1260-2
7	0	0	1	0	1254-2
8	1	1	1	1	1:1:1:1-1
9	1	1	1	1	1:1:1:1-2
10	1	0	1	0	-
11	1	1	0	0	-
12	0	1	1	0	-
13	0	0	1	1	-
14	0	1	0	1	-
15	1	0	0	1	-

Principal Components Plots

By using SIMCA-3B program, FPLOT.EXE, one can plot numerous variables derived from the principal components calculations. Because a printer in the character mode is used with this program to plot variables, the plots are restricted to two-dimensional presentations.

The program 3DPC.BAS (Principal Data Components) provides a means to plot sample scores in 3-D and color if three principal components are calculated. The 3-D display derived from the sample score values may be transferred to a disk file by using the program, FRIEZE.COM, supplied as part of PC-PAINT BRUSH or 4-Point Graphics (International Microcomputer Software, Inc., [IMSII]), San Rafel, CA 94991). The image is stored on disk and can be edited, enhanced, or labeled with a commercial software package such as PC Paint Brush (IMSII). The screen image can also be printed on a color or black/white printer.

RESULTS and DISCUSSION

Analyses of PCBs can create large data sets that are difficult to interpret, since there are 209 PCB isomers. Isomer compositions may vary widely due to differential partitioning or metabolism of compounds. In addition, wide differences in residue profiles may exist in the biota locally because of variations in effluents, combustion, or other source of residues. Chemometric methods can

greatly improve the analyst's ability to describe and model residues in these diverse samples.

The utility of principal components modeling of multivariate data like those encountered in these complex mixtures, originates from graphical presentations of sample similarity, as well as from statistical results calculated by the SIMCA-3B programs (3). Sample data are treated as points in higher dimensional space, and projections of these data are made in two- or three-dimensional space in a way that preserves most of the existing relations among samples and variables (3). This feature is especially helpful in visualizing data of more than three dimensions.

The calculations involved in principal components are summarized in Equation [1]. The objective was to derive a model of a data set having k samples and i variables in which the concentration or value of any measured value, X_{ik}, could be calculated. The principal component term is the product of Θ_{ka} and B_{ai}, where Θ_{ka} (Theta) is designated the a^{th} component "score" for sample k, and B_{ai} (Beta) is designated as the "loading" for variable i in principal component a. The term $\bar{X}_i$ is the mean of variable X_i in all samples. The residual term (or unexplained part of the measurement not modeled) is designated E_{ik}, and "A" describes the number of principal components extracted from the data. A more detailed discussion of this approach was given by Dunn *et al.* (6,18).

The concentration data obtained from each sample analysis were expressed as fractional parts and normalized to sum to 100. The normalized data were statistically analyzed, and three principal components (A=3, Equation [1]) were calculated. The PCB constituents (varibles) are numbered sequentially and correspond to peak #1, peak #2, ... to peak #105. The structure and retention index of each constituent in the mixture were reported by Schwartz et al. (9). The tabular listing of the data is available from the present authors.

The Aroclor samples listed in Table I were modeled by principal components to illustrate how the result from principal components calculations can be used in describing PCB data. The sample scores (Figure 2, A.- Theta-1 vs. Theta-2; B. Theta-1 vs. Theta-2; and C.- Theta-2 vs. Theta-3) are plotted for the samples.

Results obtained from the plots of the variable loadings (Figure 2, A'- C') for the three components provide insight into the importance of the GC Peaks in separating the various Aroclors and their mixtures (Figure 2, A - C). The loading plots show a separation of variables that are tightly clustered, the groups of variables radiating outward from the center. They are clustered in groups that reflect the variables that are characteristic of the individual Aroclors.

The sample scores (Theta-1, Theta-2, and Theta-3) in each component were used to represent the samples in a 3-D

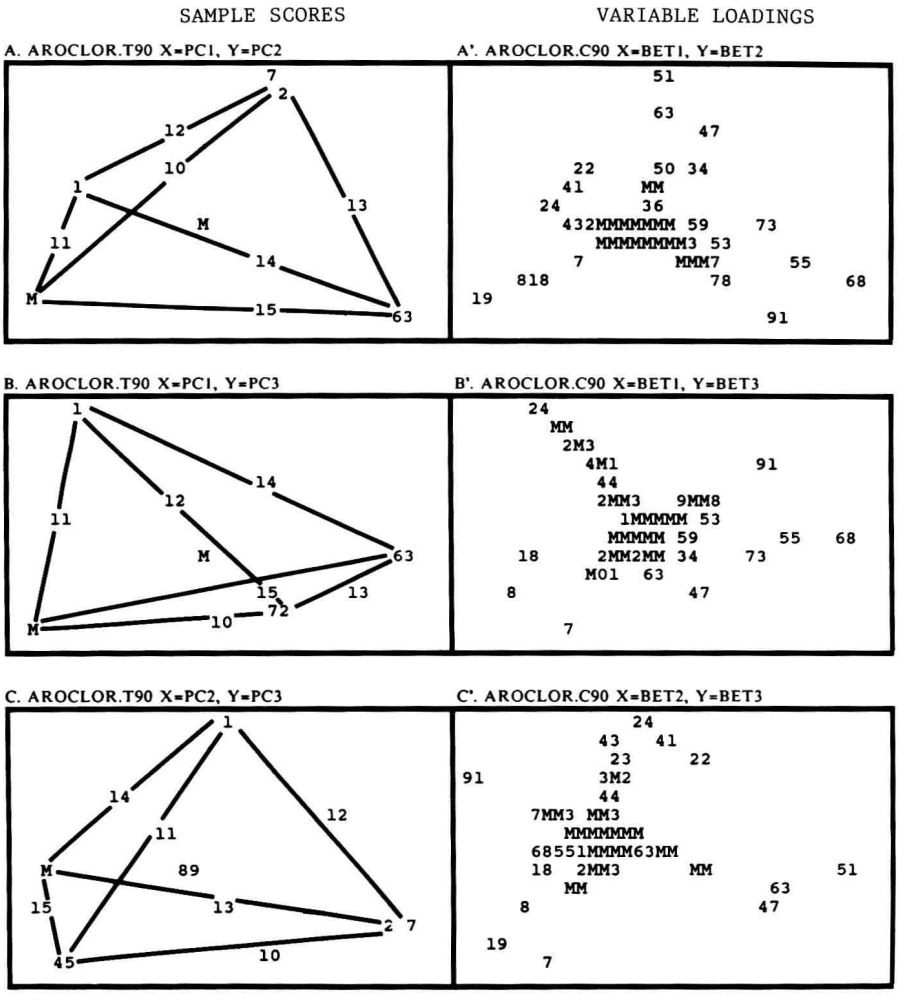

Figure 2. Principal Components Plots for Aroclor Samples (ref. Table 1 for Sample i.d.)