Polycyclic aromatic hydrocabons in soil and air: statistical analysis and classification by the SIMCA method, Environ.Sci.Technol. 21, 35-44

ARTICLE *in* ENVIRONMENTAL SCIENCE AND TECHNOLOGY · JANUARY 1987

Impact Factor: 5.33 · DOI: 10.1021/es00155a003

Polycyclic Aromatic Hydrocarbons in Soil and Air: Statistical Analysis and Classification by the SIMCA Method

Nils B. Vogt,[*,†] Frode Brakstad,[‡] Karin Thrane,[§] Svein Nordenson,[†] Jostein Krane,[‖] Eli Aamot,[‖] Knut Kolset,[†] Kim Esbensen,[⊥] and Eiliv Steinnes[‖]

Center for Industrial Research, 0314 Blindern, Oslo 3, Norway, Department of Chemistry, University of Bergen, 5014 Bergen, Norway, Norwegian Institute for Air Research, 2001 Lillestrøm, Norway, Department of Chemistry, University of Trondheim, 7055 Dragvoll, Norway, and Norwegian Computing Center, 0314 Blindern, Oslo, Norway

■ Soil samples from 12 locations in Norway have been analyzed for 9 polycyclic aromatic hydrocarbons (PAH). The same unsubstituted PAH have been determined in air samples collected near an aluminum reduction plant. Analysis by high-resolution gas chromatography–mass spectroscopy in the selected ion mode showed concentrations in soil ranging from less than 1 ppb (detection limit) to 993 ppb for individual unsubstituted PAH. The highest concentrations are found close to aluminum plants. Correlation analysis and SIMCA pattern recognition show that the patterns of unsubstituted PAH in soil samples collected close to aluminum plants are different from those in soil samples collected from other areas. Soil samples from a bog environment show a somewhat different pattern of PAH than other soil samples.

Introduction

The environmental interest in the distribution of PAH has been directed toward determining concentrations and distributions in water (*1*), air (*2*), and sediment (*3*). Comparatively fewer investigations have been conducted to address the accumulation and distribution of PAH in soil (*4–6*). The reason for this reluctance to investigate PAH concentrations and distributions in soil might stem from the difficulties involved in analysis and interpretation of results.

The use of univariate methods for source identification and investigation of transport or distribution mechanisms (*6–8*) strongly inhibit the possibility of reaching an understanding in the complex system of air–soil interaction.

With the use of microcomputers (*9*) and multivariate statistical data analytical programs now available, new approaches to the problems of interpretation and source identification are being developed. Cluster analysis has been applied to identify sources of PAH in air (*10*), and principal component analysis has been used to classify sources of chemical compounds in air samples (*11*). Factor analysis and target transformation factor analysis have been used to investigate sources and origin of elements in urban aerosols (*12–14*); principal component and factor analyses have been used to interpret particulate composition data sets and characterize atmospheric samples (*15*, *16*) and to identify the relative contribution of two sources of PAH in air particulate material (*17*). Recently, Gunderson and Thrane (*18*) reported the use of Fuzzy C varieties of pattern recognition (Fuzzy clustering) in identifying sources for PAH in air samples exposed to different sources.

The SIMCA method of principal component analysis, which is the multivariate method used in this paper, has previously been used to investigate patterns of air pollutant analytical data (*19*).

As part of a national program to investigate long-range transport of trace metals and polycyclic aromatic hydrocarbons and the distribution of these in soils (*20*), a 3-year project has been initiated to study the distribution of PAH in Norwegian soils. The intention of this study is to investigate the possibility of identifying sources of input to Norwegian soils.

Experimental Section

Air Samples. Eight randomly selected 24-h air samples collected during the period July 1981 to December 1981 by high-volume sampler (*21*) and analyzed by high-resolution gas chromatography (HRGC) have been used to compare the pattern of nine PAH from air samples to that of soil samples. The sampling and analysis are described by Thrane (*22*) and Thrane et al. (*23*).

Soil Samples. Samples of the upper 0–10 cm of soil from 12 locations were collected and dried at 50–60 °C for 2–3 days. The soil samples were then crushed in a precleaned mortar and sieved through a 6 mesh sieve to remove larger particles and parts of organic debris. Table I gives a short description of the 12 sampling sites and possible input sources. Three parallel subsamples of sieved and mixed soil were analyzed for each location, and the average concentration was calculated. The samples were randomized before analysis.

Six of the samples are from within a distance of 1000 m from either Sunndalsøra or Lista aluminum plant or Fiskaa ferrosilicon works (P1–P6). These are described in the text as "polluted samples". Six of the samples are collected at different locations on the southern coast of norway and were expected to show different and somewhat varying degress of natural, local, and/or long-range input

[*] Author to whom correspondence should be addressed.
[†] Center for Industrial Research.
[‡] University of Bergen.
[§] Norwegian Institute for Air Research.
[‖] University of Trondheim.
[⊥] Norwegian Computing Center.

0013-936X/87/0921-0035$01.50/0

Table I. Sample Locations with Short Description of Area[a]

object no.	name	description
1	P1	200 m from Fiskaa ferrosilicon plant
2	P2	1 km S.E. of Lista aluminum plant
3	P3	1 km from Sunndalsøra aluminum plant
4	P4	Sunndalsøra Center, 500 m from Sunndalsøra aluminum plant
5	P5	1 km N.E. of Sunndalsøra aluminum plant
6	P6	same as P4
7	N1	1 km along the road to Høyland (long-range input)
8	N2	Tveita, heavy hill traffic
9	N3	Moi, heavy hill traffic
10	N4	Litledalen, 3 km from Sunndalsøra aluminum plant
11	N5	3.5 km from main road, 4 km from Sunndalsøra aluminum plant
12	N6	Høolsteid, 7.5 km from Sunndalsøra aluminum plant
13–16	M1–M4	2 km S.W. of Vanse
17–24	A1–A8	air samples collected in Sunndalsøra Center (22)

[a] All samples are collected from the upper 0–10 cm of top soil. Object numbering refers to SIMCA pattern recognition analysis.

of PAH. These are samples N1–N6. These samples are described in the text as "nonpolluted" samples. Four samples are from a bog on the southern coast of Norway. Of these, M1 was analyzed together with the N1–N6 and P1–P5 samples. Peat samples M2–M4 and polluted soil sample P6 were analyzed by a second person approximately 6 months after the first series.

Solvents and Adsorbents. All solvents used were either freshly distilled, p.a. quality, or extracted with freshly distilled solvents. Methylenechloride (DCM) was obtained as technical grade and distilled in all glass system, bp 39.8 °C. Hexane was obtained at HPLC-grade solvent from Rathburn Ltd., Scotland. Dimethyl sulfoxide (Me_2SO) was p.a. quality (Merck). Tests of solvents by evaporating 100-mL portions to 0.5 mL and by running blanks throughout the chemical analyses showed that they did not contain interfering contaminants. Water used in the extraction was double-distilled and freshly extracted with hexane. The Alumina (Woelm Alumina, neutral) was soxhlet-extracted with distilled DCM (rate 10–12/h) for 6–10 h and then activated at 410 °C for 24 h. The alumina was reactivated for 24 h if not used before 2 h after activation.

All glassware was cleaned by using Chromic acid, rinsing with water and then acetone before several rinses with distilled methylenechloride (DCM) just before use.

Standards. Polycyclic aromatic hydrocarbon standards were obtained from different sources and supplied by Dr. Rudolph Schmid of the Department of Chemistry, University of Trondheim.

The following nine PAH were selected for analysis: naphthalene, acenaphthene, biphenyl, fluorene, phenanthrene, fluoranthene, pyrene, chrysene/triphenylene, and benzo[*a*]pyrene. Deuterated biphenyl and 9-methylanthracene were used as internal standards.

Analysis of Soil Samples. The analytical procedure used for the analysis of soil samples is a modification of two methods (*24*, *25*). Ten to fifty grams of soil was transferred to a round-bottom flask. Organic material was extracted by using ultrasonic agitation twice for 30 min with an excess amount of distilled DCM (100–150 mL). The extract was filtered on a precleaned glass-sinter filter, and internal standard was added together with 3 g of fully activated alumina. The solvent was evaporated by using a Buchi rotavapor with slight vacuum (0.7–0.5 atm.) and a water-bath temperature below 40 °C. The dried alumina was transferred on top of 6 g of fully activated alumina in a glass column (i.d. 2 cm). A nonpolar fraction containing mostly hydrocarbons was eluted with 30–40 mL of hexane. To control that the aromatic fraction was not eluting, a UV lamp was used. If the fluorescent layer did show migration, the hexane elution was stopped. To elute the polar/aromatic fraction, 90–100 mL of distilled DCM was used. The DCM fraction was then carefully evaporated to dryness by first using the rotavapor to between 3 and 5 mL and then applying a slow stream of grade 3 nitrogen (O_2 content below 10 ppm) until dryness.

Partitioning between Me_2SO and hexane was done by vigorous shaking in a separatory funnel. A total of 10 mL of $MeSO_2$ and 10 mL of hexane in three portions was added to the residue, and the round-bottom flask was ultrasonically agitated. The dissolved material was transferred to a 100-mL separatory funnel. After vigorous shaking and settling, the hexane phase contained nonpolar compounds, this phase was discarded except for checks to evaluate the possible extraction of PAH in this fraction. The partitioning of the polar/aromatic fraction was done by extracting from a 2:1 volume ratio mixture of water/Me_2SO into hexane. The hexane was evaporated to approximately 1.0 mL in the rotavapor at a slight vacuum with a water-bath temperature at just above 50 °C. Finally, a slow stream of nitrogen was used to evaporate the hexane to between 0.2 and 0.5 mL.

The hexane residue was analyzed qualitatively and quantitatively by quadrapole HRGC–MS (HP 5985A) in the multiple ion and selected ion modes, respectively. The samples were analyzed by injecting 1 μL onto a 0.25-mm i.d., 25-m DB-5 (Chrompack) column with average phase thickness of 0.25 μm. The injector was operated at a temperature of 295 °C with an inlet flow of 66.3 mL/min and a pressure of 10.2 psi on the column. The transfer line (open split) was operated at 295 °C, and the ion source and mass analyzer were operated at 200 °C. The oven temperature program started at 50 °C (0.2 min), was raised to 100 °C (0 min) with an oven temperature program rate of 25 °C/min, and finally raised to 295 °C (5 min) with a temperature program rate of 8 °C/min.

Five mass groups were chosen for selected ion monitoring (SIM) during the quantitative analyses. These are given in Table II together with the retention times and the m/e used for SIM analysis. Dwell times and groups were chosen so that each ion was counted between 20 and 35 times per peak. Integration of the HRGC–MS(SIM) chromatograms has been done by using the SIMQT (selected ion monitoring quantification) program available on the updated HP 5985A HRGC–MS and checked by manually selecting integration areas. Quantification was only made if peaks had retention times within 1–0.5% of expected retention times. The reproducibility of the combined workup and HRGC–MS(SIM) quantification has been found to be between 5% for major, i.e., ppm level, and 20% for minor, i.e., ppb level, components, respectively. Tests of recovery from authentic samples were made by reextracting already extracted samples. No peaks with masses and retention times coinciding with the nine selected PAH were found in the reextracted residue when analyzed with HRGC–MS in the SIM mode.

The recovery of the workup method has been evaluated quantitatively by analyzing spiked test samples. Analyses during evaluation were made on a HP 5880 HRGC equipped with the same DB-5 column as for HRGC–MS analysis, on-column injection, FID, and electronic inte-

Table II. Groups of Ions Selected for HRGC-MS(SIM) Quantification of PAH Together with Retention Times and m/e Used for Quantification of Each Separate Compound

	group no.				
	1	2	3	4	5
start time	4.0	11.0	18.0	22.0	25.5
run time	7.0	7.0	4.0	3.5	8.0
total dwell time	300	300	300	300	400
mass	128.1	166.1	202.1	228.1	252.1
	153.0	178.1	206.1	252.1	254.1
	154.1	192.1	216.0	254.1	276.1
					278.1

compd	retention time, min	m/e (SIM)
naphthalene	6.15	128.1
acenaphthene	10.55	153.0
biphenyl	8.98	154.1
fluorene	12.13	166.1
phenanthrene	14.97	178.1
fluoranthene	18.62	202.1
pyrene	19.27	202.1
chrysene/triphenylene	23.13	228.1
benzo[a]pyrene	26.95	252.1

grator. The same temperature program was used as for HRGC-MS analysis. The recovery of the workup method has been found to vary between 69% and 80% for parent (unsubstituted) PAH and between 60% and 70% for alkyl-PAH with a relative standard deviation of between 4% and 10% for three real parallel spiked samples.

Statistical Analysis

Correlation Analysis. Pearson moment correlation analysis between the nine PAH in the air and soil samples has been carried out. The correlation coefficients for the correlation between the nine PAH are tabulated in Table IV. A confidence level of 99% has been used.

SIMCA Pattern Recognition. The SIMCA (*S*oft *I*ndependant *M*odeling of *C*lass *A*nalogy) (*26, 27*) pattern recognition method and applications in environmental and ecological chemistry have been described by Grahl-Nielsen et al. (*28*), Stalling et al. (*29*), and Vogt and Knutsen (*30*).

Although the SIMCA method is robust to nonnormal distribution of data, it has been found to work best if the residuals are approximately normally distributed (*31, 32*) and when strict statistical tests are used (e.g., *F* test) to define class boundaries, i.e., to identify outlier samples. Several possibilities for pretreatment (transformations etc.) of data are available in the SIMCA package.

To compensate for strongly skewed distributions and/or to avoid that high concentrations dominate the mathematical modeling, all data in the SIMCA analysis in this paper have been transformed to log $(x + 1)$ (*33, 34*).

Since the modeling performed in SIMCA analysis is scale-dependent, each separate class has been independently scaled to unit variance (class scaling) (*35*); this is not necessary if all variables have the same variance.

Results and Discussion

Analytical. The complex mixture obtained from extraction of soil by methylenechloride makes clean-up of samples necessary. The methods described by Jentoft (*24*) and by Lee et al. (*4*) show clean-up of samples for PAH analysis to range from simple partition chromatography through liquid–liquid partitioning to the use of sophisticated HPLC methods.

Ultrasonic extraction with DCM was found to give the best recovery although this method also extracts large amounts of other interfering material. Adsorption chromatography on its own was found not to be sufficient to remove interfering compounds. Liquid–liquid partitioning with Me_2SO and water has been reported (*4, 25*) to have favorable partition coefficients compared to the method described by Grimmer and Bøhnke using dimethylformamide (DMF) and water (*36*). On the other hand, liquid–liquid partitioning with Me_2SO (*4, 25*) alone, without preliminary adsorption chromatography, was found to pose problems when extracted samples contained high levels of colloid material after the ultrasonic DCM extraction. It was decided to use a combined approach with (i) adsorption chromatography first to remove (a) colloid material and polar humic material that might disturb the liquid-liquid partitioning and (b) most of the nonpolar compounds and then (ii) partitioning with Me_2SO to obtain the PAH fraction (*37*).

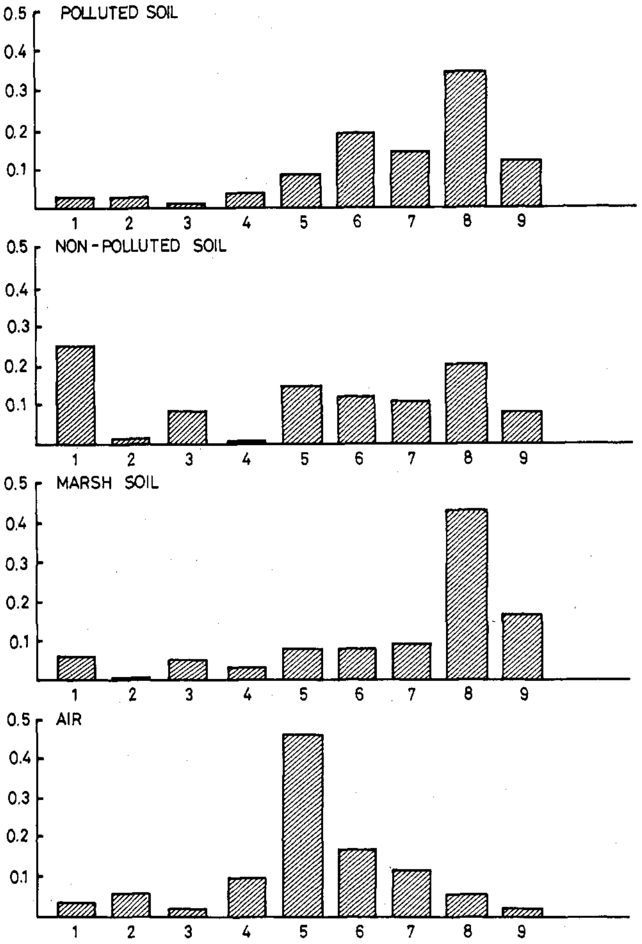

Figure 1. Bar diagram of normalized average concentrations for the four types of samples. Normalization has been done by dividing the average concentration for each PAH by the summed concentration of all nine PAH in each separate sample type. Numbering of PAH corresponds with Table III.

Quantitative. Table III shows average concentrations of PAH in the different soil samples. Figure 1 shows bar diagrams of the nine PAH as normalized average concentrations for the four types of samples analyzed. The bar diagrams show that the air samples have a different univariate pattern from all the soil samples. The soil samples collected far from aluminum or ferrosilicon plants contain higher concentrations of naphthalene compared to the polluted soil samples, whereas the polluted soil samples have higher concentrations of the relatively higher weight PAH's 6–9. The bog samples show a somewhat intermediate bar-diagram pattern between the polluted soil samples and the nonpolluted soil samples with a dominating

Table III. Average Concentration of Individual PAH in ng/g (ppb) for Soil Samples and in ng/m^3 for Air Samples[a]

variable no.	PAH	polluted soil	nonpolluted soil	bog soil	air
1	naphthalene	48.3	46.2	57.7	11.3
2	acenaphthene	53.6	1.7	3.8	32.5
3	biphenyl	23.0	14.8	31.7	8.2
4	fluorene	80.2	*b*	14.4	59.3
5	phenanthrene	352.9	30.0	77.7	286.2
6	fluoranthene	572.7	22.3	83.2	111.4
7	pyrene	459.1	19.7	89.7	64.0
8	chrysene/ triphenylene	993.8	38.3	379.5	29.6
9	benzo[*a*]pyrene	321.3	14.5	156.5	6.1

[a] Variable number refers to both the statistical analysis and the SIMCA analysis. [b] Fluorene was not detected in any of the samples collected from nonpolluted locations.

concentration of PAH 8 (crysene/triphenylene). The bog samples show high concentrations of PAH considering that these samples come from an area presumably not affected by a local pollution source (see Table III). There are two possible explanations: (a) the high PAH concentrations arise from sources in the area that have not been identified, or (b) the bog environment accumulates PAH from deposition or PAH are formed in the bog environment by some process.

The volatility of the PAH studied in this investigation suggests that PAH concentrations in soil will depend on the temperature and most likely be in equilibrium with the atmospheric concentration. The quantitative difference however suggests that the parent PAH are accumulated in soil around local input sources.

Statistical. Table IV shows the univariate correlation analysis of the nine PAH in the three sample types, air, polluted soil, nonpolluted soil, and also the combined correlation analysis for the polluted soil and air samples. The correlation coefficient for a 99% confidence level of rejecting H_0, $R(0.05,(2),\nu) = 0$, is also given (*38*). The correlation coefficient measures the degree of association between the PAH in the sample types analyzed. There are visual patterns in the correlation coefficients suggesting that correlations between PAH in polluted soil and air samples and also between the combined polluted soil and air data set are similar, whereas the nonpolluted soil samples show much lower correlations between PAH. This univariate analysis suggested that there was a possible connection between the pattern of PAH in the air samples from Sunndalsøra and the pattern of PAH in the polluted soil samples and indicated that some pattern recognition method should be tried.

SIMCA Principal Component Analysis. The intention of *principal component* analysis is that of dimension reduction through statistical investigation of systematic variation in data matrices. Principal components are constructed orthogonal to each other in the *m*-dimensional space by decomposition of the variance–covariance matrix. In the unsupervised model (*9*), principal component analysis may therefore be viewed essentially as a projection method where the intention is to preserve as much as possible of the systematic variation in the data set, while projecting onto as few principal axes as possible. Interpretation of object (sample) grouping and variable correlation are made on object score and variable loading plots, respectively (*33*). The SIMCA method is an extension of principal component analysis to include supervised modeling (*9, 30, 39*). The calculation of principal components depends on the data that are being analyzed. To investigate the difference between specific groups of samples, these may be analyzed together two and two classes at a time to obtain visual interpretation of object score and variable loading plots or the separate class models may be compared by modeling power and discrimination power (*31*). Coomans plots are used to investigate how good the modeling of individual classes are, i.e., to detect outliers and identify class models that are not resolved.

The SIMCA method of data analysis has been found to work very well with chemical data and is one of the few multivariate methods that allows data interpretation at several levels (*40, 41*).

Table IV. Correlation Analysis between the Nine PAR in the Samples Collected from Polluted Soil (Lower Left, Part A), Air (Upper Right, Part A), Nonpolluted Soil (Lower Left, Part B), and Polluted Soil and Air Together (Upper Right, Part B)[a]

Part A

	1	2	3	4	5	6	7	8	9
1		0.092	0.312	−0.217	−0.371	−0.336	−0.267	−0.400	−0.028
2	*0.958*		−0.169	0.739	*0.852*	*0.863*	*0.877*	*0.786*	*0.897*
3	0.569	0.743		−0.059	−0.218	−0.244	−0.261	−0.289	0.019
4	0.795	0.626	0.312		*0.839*	*0.857*	*0.844*	0.790	0.749
5	*0.995*	*0.976*	0.608	0.727		*0.997*	*0.986*	*0.952*	*0.848*
6	*0.992*	*0.985*	0.651	0.736	*0.995*		*0.994*	*0.959*	*0.836*
7	*0.995*	*0.979*	0.643	0.754	*0.996*	*0.999*		*0.959*	0.822
8	*0.939*	*0.970*	0.673	0.635	*0.943*	*0.966*	*0.953*		0.684
9	0.872	*0.952*	0.747	0.526	0.887	*0.920*	0.901	*0.983*	

Part B

	1	2	3	4	5	6	7	8	9
1		*0.859*	0.611	0.540	*0.845*	*0.977*	*0.984*	*0.943*	*0.887*
2	0.560		0.471	0.631	*0.947*	*0.918*	*0.903*	*0.864*	*0.822*
3	0.132	0.159		0.293	0.310	0.598	0.610	*0.677*	*0.763*
4	*b*	*b*	*b*		*0.676*	0.593	0.586	0.490	0.431
5	0.067	−0.369	0.757	*b*		*0.912*	*0.899*	*0.811*	*0.726*
6	0.492	0	0.798	*b*	0.880		*0.998*	*0.964*	*0.915*
7	0.438	−0.011	0.794	*b*	0.895	*0.993*		*0.957*	*0.907*
8	*0.998*	0.593	0.122	*b*	0.027	0.461	0.189		0.307
9	0.334	0	−0.225	*b*	0.164	0.189	0.238	0.307	

[a] Refer to Table III for numbering. Correlation coefficients larger than R are significant at the 99% confidence level and are italicized. R(poll. soil)(0.01,(2),4) = 0.917; R(air)(0.01,(2),6) = 0.834; R(nonpoll. soil)(0.01,(2),4) = 0.917; R(poll. soil and air)(0.01,(2),12) = 0.661. [b] PAH 4 (fluorene) was not detected in any of the samples from nonpolluted soils.

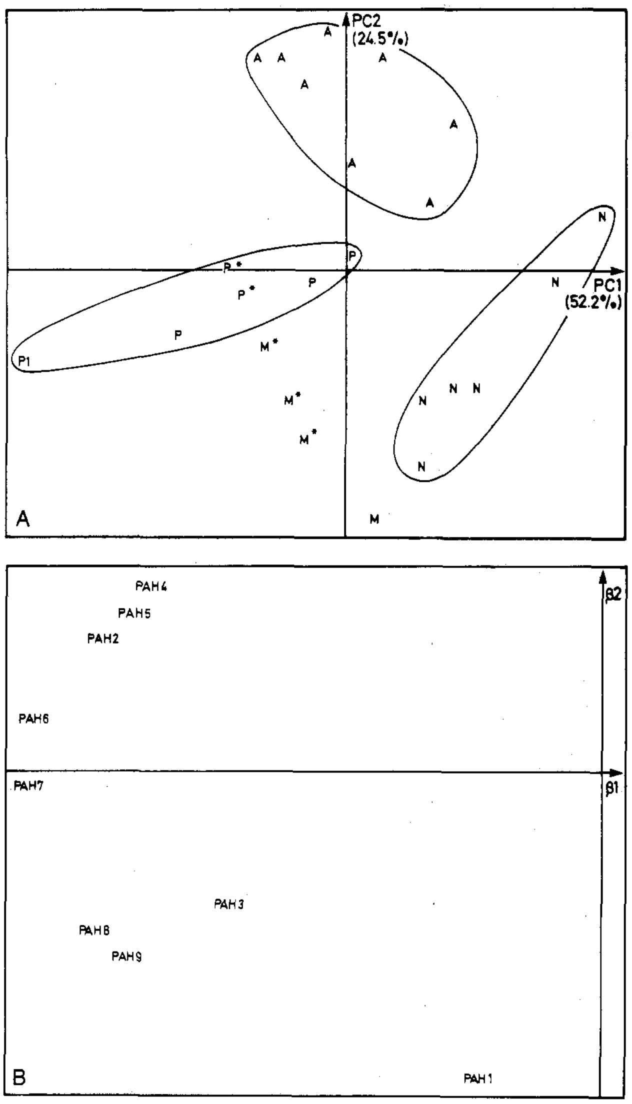

Figure 2. (A) Principal component score plot of all samples. Rings are drawn around the sample classes of air (A) and nonpolluted (N) and polluted (P) soils. Samples marked with an asterisk have been analyzed by a second person. Multiple samples (M) indicates that objects are not resolved in the plot. The percentage for each principal component tells how much of the variance is explained by that principal component. (B) Corresponding variable loading plot.

To remove skewness and large differences in numerical values between samples, the log transformation has been used in this paper. Another way to reduce the effect of samples with high concentrations, or quantitative correlation between variables, influencing the model construction is to normalize the data matrix objectwise to a sum. This has the disadvantage that subsequent interpretation of variable loading plots must be done with caution because object normalization leads to closure [erroneous negative correlation between large and small variable variances in the original matrix and erroneous positive correlation between small variable variances (*42*)]. If either a quantitative correlation between variables is present or samples with high concentrations are present, these are usually visualized along the first principal component in the object score and variable loading plots.

Figure 2A shows the object score plot of the unsupervised analysis of the whole data set. Two significant (cross-validated) principal components were found. Together they explain 76.7% of the variance in the data set. The plot suggests the presence of four different classes, nonpolluted soil (N), polluted soil (P), air (A), and bog (B). The polluted soil class (P) may seem to contain one outlier. The bog samples have a different pattern of the nine PAH compared to all other samples. This is seen by the direction of the bog sample class being perpendicular to the others. There are, however, too few bog samples to be considered a separate class.

Figure 2B is the equivalent variable loading plot for the total data set. The strong first principal component describes 52.2% of the variance in the total data set (Figure 2A). All variables, apart from variable 1 (naphthalene), show high loadings (correlation) along this component. As is seen in the univariate correlation analysis, there is relatively strong correlation between PAH for the air and polluted soil sample data sets. The first principal component is therefore interpreted as representing a "quantitative" correlation component, which implies that the main variance (52.2%) in the total data set is quantitative. The orthogonality of principal component number 2 suggests that this second principal component is the component containing "pattern" information. The main *pattern* difference, when the whole data set is analyzed in the unsupervised mode, is therefore between the air samples (A) and the soil samples (P, N, and B).

In order to investigate the similarities and differences between the sample classes, nonpolluted (N) and polluted (P) soil samples and the polluted soil (P) and air (A) samples, separate principal component analyses of pairs of sample types were done (*11*). Object score plots and variable loading plots are shown in Figure 3 [nonpolluted (N) and polluted (P)] and 4 [polluted (P) and air (A)], respectively. In both of these collected sample classes, one polluted soil sample (P1) is seen to be very well discriminated from the others. The sample contains very high concentrations of PAH and was collected 200 m from the Fiskaa *ferrosilicon* work plant. This sample may therefore have a different pattern of PAH. With only one sample it is not possible to assess the difference between this sample and the samples from around Sunndalsøra and Lista *aluminum* plants (P2–P6).

Nonpolluted vs. Polluted Soil. The object score plot in Figure 3A shows a first component explaining 69.6% of the variance in the data set of nonpolluted (N) and polluted (P) soil samples collected. The variable loading plot, Figure 3B, shows a strong correlation between variables 2 and 5–9 whereas variables 3 and 4 are relatively less correlated. Variables 1 and 4 are seen to be responsible for the direction of the second principal component. The second principal component is not statistically significant but is calculated for visual display purposes. The dominant first principal component is again interpreted as a quantitative correlation component. Principal component 1 describing the quantitative correlation between the variables accounts for the differences between average concentrations of the two types of samples. Following the discussion by Massart and Thielmanns (*11*), the main difference between the two types of samples, apart from the stronger quantitative correlation between the PAH in the polluted samples seen in Table IV and visually displayed as the strong correlation of variables along principal component 1 of Figure 3B, is that nonpolluted soil samples (N) contain higher concentrations of variable 1 (naphthalene) and polluted soil samples contain higher concentrations of variable 4 (fluorene), relative to the average concentration. This is seen as variables 1 and 4 are the variables responsible for the direction of principal component 2 along which the two classes are separated by pattern differences. This same tendency was seen in the variable loading plot of the unsupersized analysis of the total data set (Figure 2B) where variable 1 (naphthalene)

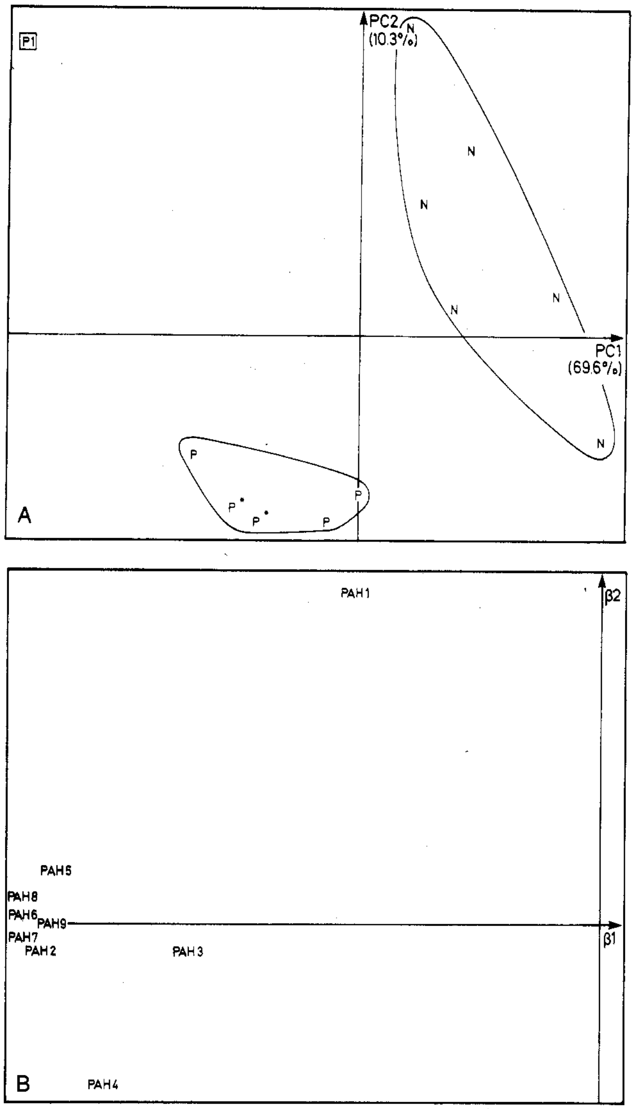

Figure 3. (A) Principal component score plot of the polluted (P) and nonpolluted (N) soil samples. One polluted soil sample (box) is visually identified as an outlier. (B) Corresponding variable loading plot.

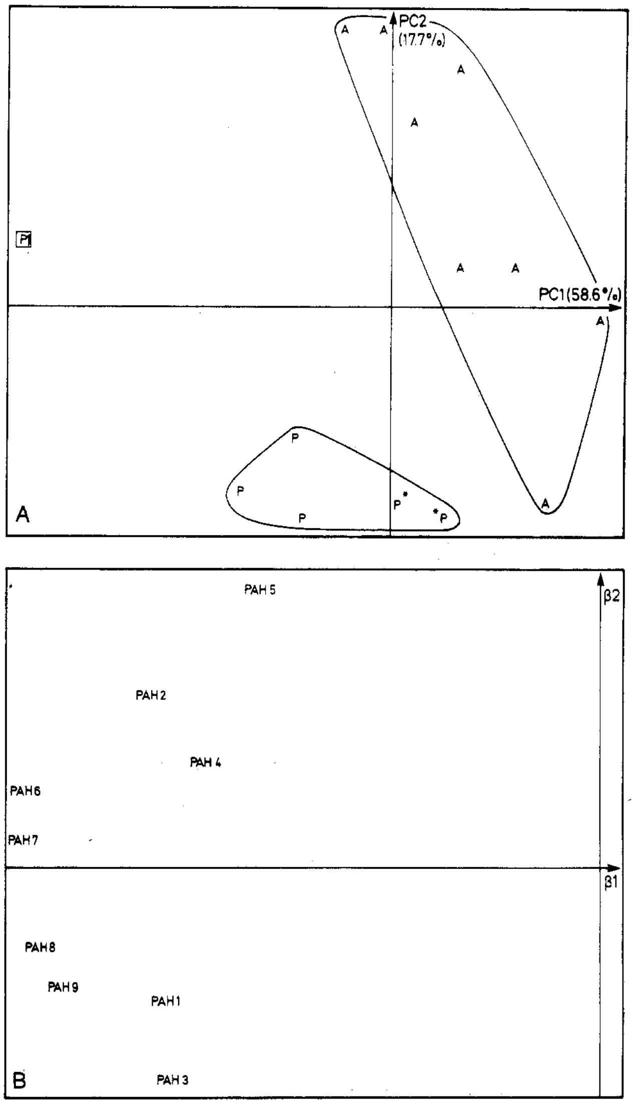

Figure 4. (A) Principal component score plot of the polluted (P) and air (A) samples. One polluted soil sample (box) is visually identified as an outlier. (B) Corresponding variable loading plot.

and variable 4 (fluorene) were the two variables that were responsible for the direction of principal component 2. The bar diagram, Figure 1, and Table III show that variable 1 (naphthalene) has a relative high concentration in nonpolluted soil samples and that fluorene was not identified in any of the nonpolluted soil samples.

Polluted Soil vs. Air. Figure 4 suggests a similar interpretation as the plots in Figure 3: a relatively large first principal component explaining 58.6% of the variance and a second principal component, which is not statistically significant, explaining 17.7% of the variance. The variable loading plot (Figure 4B) shows more structure than Figure 3B. The polluted soil and air samples show a similar, if not identical, pattern of univariate PAH correlation (Table IV). Interpreting the variable loading plot differences between sample types along principal component 2 (Figure 4B) shows that variables 3 (biphenyl) and 5 (phenanthrene) are responsible for the direction of the second principal component. Variables 3 and 5 are seen to be the variables that are most negatively correlated along principal component 2; i.e., they have the lowest and highest loadings, respectively. This is not seen in the correlation analysis (Table IVB, upper right). Comparing this to the bar-diagram representation in Figure 1 and the average concentrations in Table III, it is seen that variable 5 (phenanthrene) has a dominating difference in relative concentration in the two sample types air and polluted soil, whereas variable 3 (biphenyl) has a very similar numerical concentration in the two sample types and does not display the same univariate relative difference as variable 5 (see Figure 1). The orthogonal construction of principal components suggests that when the quantitative correlation, present in the first and only statistically significant principal component, is subtracted, there is a "left-over" pattern difference between nonpolluted and polluted soil and polluted soil and air samples. This is visualized by the separation along principal component 2 of the different sample classes in Figures 4A and 5a, respectively. The variable loading plots, Figures 4B and 5B, show that the variables responsible for the differences are negatively correlated. The dominating quantitative correlation found in the polluted soil and air samples makes it difficult to detect these differences with traditional univariate analysis. On the other hand, the decomposition of the data along orthogonally independent principal components allows this structure to be investigated, and in addition, the visual display possibilities present in principal component analyses greatly enhances the interpretation. The use of pattern recognition in environmental chemistry is therefore not restricted to problems of classification. Interpretation of the difference between PAH patterns in air samples and that in soil samples by decomposing the data structure into

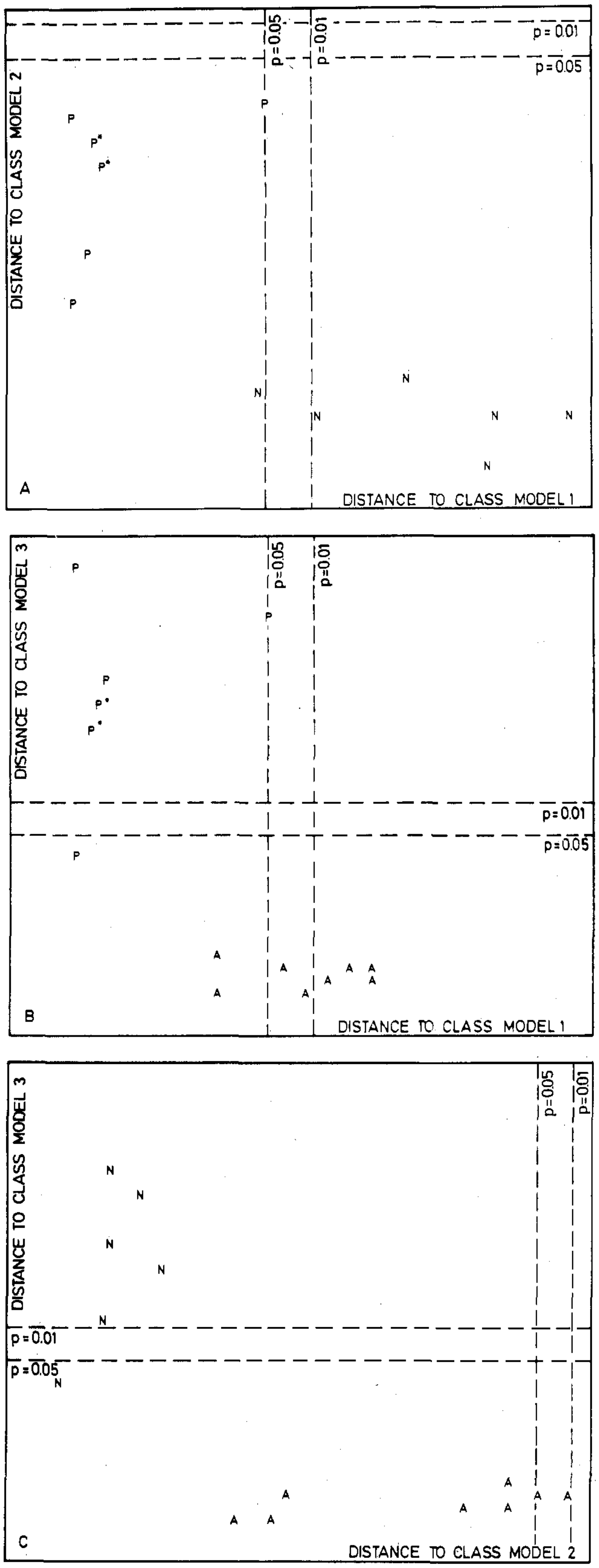

Figure 5. Coomans, or residual, plots of two and two classes. The axis represent distance to class model as defined in Table V. (P) Polluted and (N) nonpolluted soil samples and (A) air samples. Dotted lines represent 95% and 99% confidence intervals around the class models. (A) Coomans plot of class 1 (horizontal) and class 2. (B) Coomans plot of class 1 (horizontal) and class 3. (C) Coomans plot of class 2 (horizontal) and class 3.

principal components may allow the interpretation of processes governing the distribution of PAH between soil and air.

Table V. SIMCA Principal Component Models of the Individual Sample Classes[a]

	polluted soil	nonpolluted soil	air
no. of significant principal components	1	1	1
amount of variance explained, %	70.4	48.5	59.3
F1	8	8	8
F2	24	32	48
RSD max (P = 0.01)	1.078	1.351	1.130
RSD max (P = 0.05)	0.918	1.163	1.000

[a] F1 and F2 are the degrees of freedom, and RSD max is the maximum standard deviation. RSD max values have been used to draw dotted lines in the Coomans plots in Figure 5.

Classification analysis of the polluted soil class suggested that sample P1 had a very large effect on the modeling for this class. When left out of the class model, it was found that this object must be classified as an outlier. This sample has therefore been left out of the polluted soil sample class during classification analyses by SIMCA (*35*).

In order to investigate how well the classification of different sample types is, residual plots, or Coomans plots (*31*), may be used. In Coomans plots, tolerance intervals around separate class models are calculated and plotted on separate axes. In these plots, the residual distance of objects projected onto two different classes may be visually displayed. The tolerance interval is calculated from the residual variance in the class model by using an approximate F test with 95% and 99% confidence limits (*31*), see Table V. For all of the separate classes, i.e., polluted soil (n = 5), nonpolluted soil (n = 6), and air (n = 8), only one principal component was significant. Figure 5 shows three Coomans plots where the samples of polluted soil/nonpolluted soil, air/nonpolluted soil, and air/polluted soil are plotted. The tolerance interval around class model 2, the nonpolluted soil samples, encompassing objects from other classes, reflects that this class is not very well described with the one significant principal component. This is also seen in that only 48.5% of the variance in this class is described whereas for the polluted soil sample class 70.4% and for the air sample class 59.3% of the variance is explained. The low amount of variance described and the fact that the nonpolluted soil class consists of samples collected at different locations with different sources of PAH suggest that this class might be an example of the unsymmetric case in pattern recognition (*41*).

The unsymmetric case arises when a potential class found in unsupervised learning pattern recognition really consists of several types of different samples. The interpretation of this class as a case of unsymmetric classification is also supported by the univariate correlation analysis. Unsymmetric classes may not be modeled properly by simple principal component models (*41*). The one polluted soil sample left out of the polished class modeling and suspected of being an outlier of the polluted soil sample class is seen to be positioned on the 95% confidence interval line of Figure 5A,B. In Figure 5B it is found to be uniquely classified as different from air samples. This supports the classification of this sample as an outlier and suggests that it should be possible to differentiate between soil samples collected around ferrosilicium plants and around aluminum plants.

The discussion has shown an example where modeling of classes from environmental data sets with several sample types collected together makes it difficult to interpret variable loading plots. There are several reasons for this.

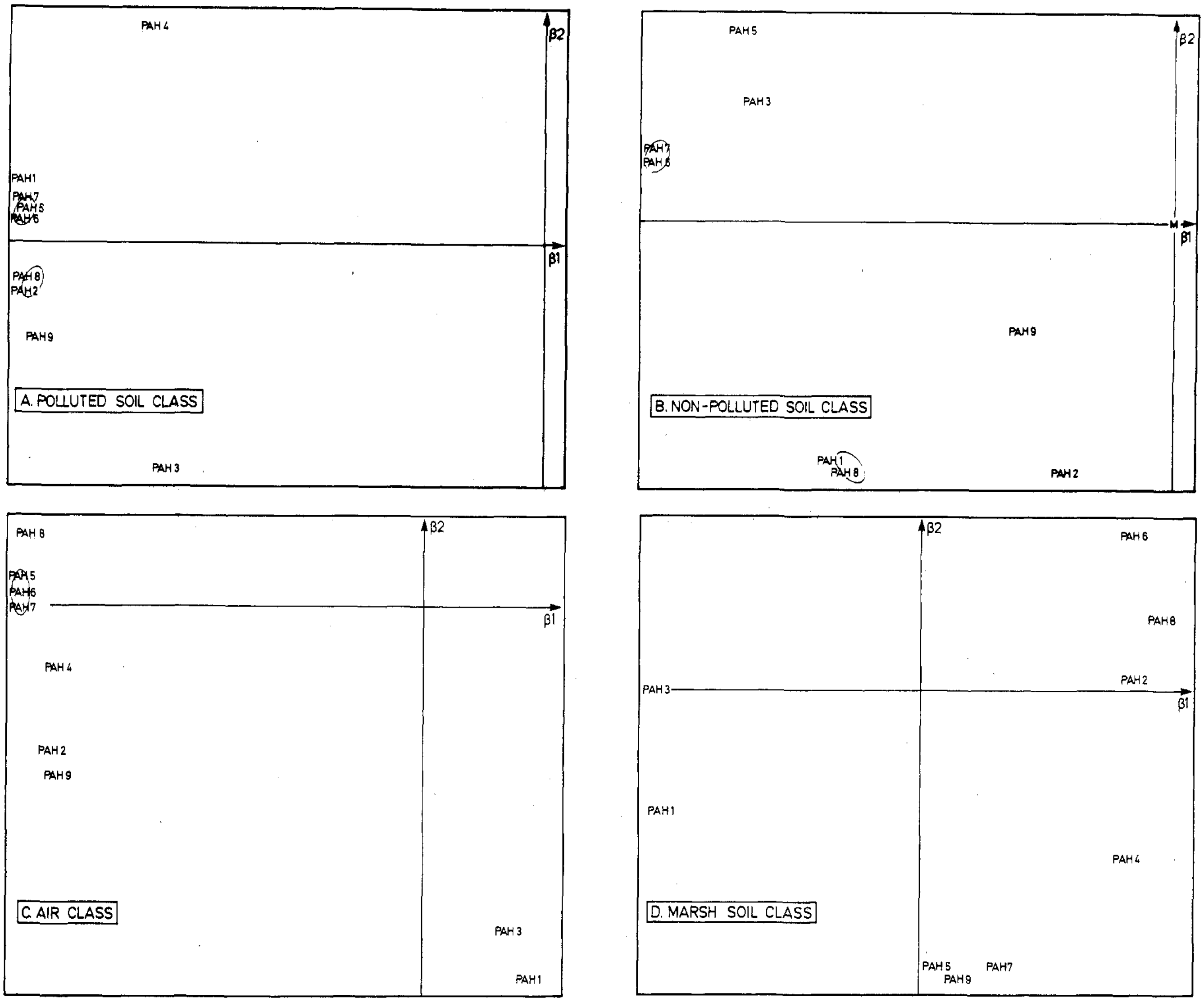

Figure 6. Variable loading plots of the separate classes. Correlation between variables is seen by close positioning in these plots. Compare with Table IV.

The most dominating effect in the data reported here is the quantitative correlation accounting for the strong first principal component. This quantitative correlation has traditionally been circumvented by applying some form of normalization objectwise preliminary to the principal component analysis. This has the effect that spurios correlations are augmented, and care should be exhibited when interpreting data that have been normalized objectwise. We have recently developed an approach where the information obtained from a preliminary principal component analysis is used to eliminate the effects that give uninteresting, and variancewise dominating, first principal components. The results from using this approach on several environmental data sets will be published elsewhere (*43*).

To obtain an idea of how the variables are grouped in the four classes, separately disjoint SIMCA models were made for each class (Table V). All classes had only one cross-validated significant principal component. The second principal component has been calculated for display purposes only. Variable loading plots of the separately modeled classes are given in Figure 6. For the polluted soil sample class, one object (P1) has been left out for the calculations in Table V, but is included for the visual analysis in Figure 6A.

The polluted soil (Figure 6A) and air samples (Figure 6C) show very much the same pattern of variable correlation, whereas the nonpolluted soil samples (Figure 6B) show a substantially different pattern. The polluted soil sample class (n = 6) clearly shows a quantitative first principal component with variables 1, 2, and 5–9 strongly correlated and variables 3 and 4 somewhat less correlated along this component. Comparing this to Table IVA (lower left), it is seen that the correlation along principal component 1 is present in the pattern of significant correlations (italicized). Again, the sample collected close to the Fiskaa ferrosilicon plant (P1) was found to be different from the other samples. The second principal component, although not significant, might, if more samples are collected from the ferrosilicon location, show that the two types of locations may be classified as separate on the basis of their unsubstituted PAH pattern.

The nonpolluted sample class has a less dominant first principal component describing only 48.5% of the variance. Variables 6 and 7 (fluoranthrene and pyrene) dominate and direction of the first principal component. The other variables are more distributed along principal component 2. There are two groups of two variables each that are strongly correlated. There are variables 1 and 8 (naphthalene and chrysene/triphenylene) with negative loadings along principal component 2 and the previously mentioned variables 6 and 7 responsible for the strong first principal component. The strong correlation of these variables is also seen in Table IVB (lower left), suggesting that there is a quantitative correlation between these two groups of variables in the nonpolluted soil samples. The correlation may result from either specific soil processes or from two different sources of input.

The air class loading plot, Figure 6C, shows a positive correlation along principal component 1 of variables 2 and 4–9 with a specially strong correlation between variables 5–7. The negative correlation along principal component 1 between variables 1 and 3 and the previous mentioned group of variables is also seen in Table IVA (upper right). This negative correlation along the first principal component being present in the univariate correlation analysis supports the interpretation of this first component as being a quantitative component.

The input of PAH to the air samples is most likely dominated by the Sunndalsøra aluminum plant. The negative correlation indicates either that there are atmospheric processes influencing the distribution of PAH or that possibly, if the PAH are present in different particulate fractions, that the sampling procedure may have some discriminating effect. The variable loading plot for bog samples, although too small to be considered a class, is given in Figure 6D. The first point to notice is that the variables are all positioned on a semicircle. This is because calculating two principal components describes approximately 90% of the variance in the data set. The second point is the relative large amount of structure in the variable positioning, paralleling that of the nonpolluted soil sample class, compared to the polluted soil and air classes where there is a dominating quantitative correlation between most of the variables. The structure indicates that there might be several processes responsible for the pattern of PAH in these bog samples.

The results suggest that the soil around the aluminum and ferrosilicon plants is directly influenced by input from the air and that there are more complex sources for the PAH in both the nonpolluted and the bog classes. The complexity of the data analyzed is this investigation and the scope of the project do not allow speculation of possible processes responsible for the differences between loading plots in the nonpolluted soil and bog classes. The results do show that there is a potential in using pattern recognition methods to investigate and interpret complex chemical systems where different interacting processes and sources may be present.

Restricting the study to unsubstituted PAH may be seen to have advantages. The first is that analyses of soil samples with liquid–liquid extraction will pose problems if alkyl-substituted PAH are used. There are two reasons for this. The first is that extraction efficiencies will influence the quantitative result through the different mechanisms responsible for soil adsorption and liquid extraction. The second point is that studies of literature show that alkyl-substituted PAH are infrequent and quite variable in environmental samples. Although this may lead to proposing selected PAH as identifyers of specific pollution sources in air samples, the processes responsible for the formation and distribution of these compound groups between soil and air are not well understood. This prohibits the use of this approach for soil samples.

The project is presently being continued to develop data on the distribution of PAH from areas that have different input sources. By investigating the patterns of these different input sources, it may be possible to use the pattern recognition approach in source identification of pollution in soil samples.

Conclusions

Point sources of PAH are evident in the concentrations, the correlation analyses, and the SIMCA patterns of unsubstituted PAH in soil samples. The patterns of unsubstituted PAH in soil around aluminum and ferrosilicon plants are different from the pattern found in relatively unpolluted soil. comparing the pattern of unsubstituted PAH in soil to the pattern of unsubstituted PAH in air shows that there are clearly differences in these patterns. The potential amount of information available from interpretation of principal component modeling by the SIMCA method is illustrated through the differences found in variable loading plots of the different sample types. Analyzing soil samples that come from areas with a dominating local input source and comparing the pattern of PAH in these samples to that of air samples from the same area may allow interpretation of the mechanisms that govern the distribution of PAH between soil and air. Bog samples show specific PAH patterns.

Acknowledgments

Suggestions put forward by Dr. S. Wold at Umeå University and from two anonymous referees are gratefully acknowledged.

Registry No. Al, 7429-90-5; naphthalene, 91-20-3; acenaphthene, 83-32-9; biphenyl, 92-52-4; fluorene, 86-73-7; phenanthrene, 85-01-8; fluoranthene, 206-44-0; pyrene, 129-00-0; chrysene, 218-01-9; triphenylene, 217-59-4; benzo[*a*]pyrene, 50-32-8; ferrosilicon, 8049-17-0.

Literature Cited

(1) Acheson, M. A.; Harrison, R. M.; Perry, R.; Wellings, R. A. *Water Res.* **1976**, *10*, 207.

(2) Avery, M. J.; Richard, J. J.; Junk, G. A. *Talanta* **1984**, *31*(1), 49.

(3) Hites, R. A.; LaFlamme, R. E.; Farrington, J. W. *Science (Washington, D.C.)* **1977**, *198*, 829.

(4) Lee, M. L.; Novotny, M. V.; Bartle, K. *Analytical Chemistry of Polycyclic Aromatic Hydrocarbons*; Academic: New York, 1981; pp 78–156.

(5) Zander, M. In *The Handbook of Environmental Chemistry, Anthropogenic Compounds*; Hutzinger, O., Ed.; Springer-Verlag: New York, 1980; Vol. 3, Part A.

(6) Windsor, J. G.; Hites, R. A. *Geochim. Cosmochim. Acta* **1979**, *43*, 27.

(7) Neff, J. M. *Polycyclic Aromatic Hydrocarbons in the Aquatic Environment*; Applied Science: London, 1979; pp 29–43.

(8) Sporstøl, S. P.; Gjøs, N.; Lichtenthaler, R. G.; Urdal, K.; Oreld, F.; Skei, J. *Environ. Sci. Technol.* **1983**, *17*, 282.

(9) Derde, M. P.; Massart, D. L. *Z. Anal. Chem.* **1982**, *313*, 484–495.

(10) Thrane, K. "Applications of Cluster Analysis to Identify Sources of Airborne Polycyclic Aromatic Hydrocarbons"; Proceedings of the 77th APCA Annual Meeting and Exhibition, June 24–29, 1984; APCA: Pittsburgh, PA, 1984; Vol. 1, p 13.

(11) Massart, D. L.; Thielemans, A. *Chimia* **1985**, *39*, 7–8, 236.

(12) Hopke, P. K. *Ann. N.Y. Acad. Sci.* **1980**, *338*, 103.

(13) Hopke, P. K.; Gladney, E. S.; Gordon, G. E.; Zoller, W. H.; Jones, A. G. *Atmos. Environ.* **1976**, *10*, 1015.

(14) Alpert, D. J.; Hopke, P. K. *Atmos. Environ.* **1980**, *14*, 1137.

(15) Gaarenstroom, P. D.; Perone, S. P.; Moyers, J. L. *Environ. Sci. Technol.* **1977**, *11*, 795.

(16) Roscoe, B. A.; Hopke, P. K.; Dattner, S. L.; Jemks, J. M. *J. Air Pollut. Control Assoc.* **1982**, *32*(6), 637.

(17) Cretney, J. R.; Lee, H. K.; Wright, G. J.; Swallow, W. H.; Taylor M. C. *Environ. Sci. Technol.* **1985**, *19*, 397.

(18) Gunderson, R. W.; Thrane, K. In *Environmental Applications of Chemometrics*; Breen, J. J.; Robinson, P. E., Eds.; ACS Symposium Series 292; American Chemical Society: Washington, DC, 1985; pp 130–147.

(19) Scott, D. R. "Application of SIMCA Pattern Recognition to Air Pollutant Analytical Data"; Technical Report. U.S.

Environmental Protection Agency, U.S. Goverment Printing Office: Washington, DC, 1984; EPA-600/D-84-271, p 40.
(20) Solberg, W.; Steinnes, E. "Heavy Metal Contamination of Terrestrial Ecosystems from Long Distance Atmospheric Transport"; Proceedings of the Acid Rain and Forest Resources Conference, Quebec, Canada, 14–17 June 1983.
(21) Thrane, K. E.; Mikalsen, A. *Atmos. Environ.* **1981**, *15*, 909.
(22) Thrane, K. "Luftkvalitet i et boligområde"; Report No. 1/83; Norwegian Institute for Air Research: Lillestrøm, Norway, March 1983; p 86, Reference 22981.
(23) Thrane, K.; Mikalsen, A.; Stray, H. *Int. J. Environ. Anal. Chem.* **1985**, *23*, 111–134.
(24) Jentoft, N.-A. Cand scient Thesis, University of Trondheim, Norway, 1982.
(25) Natusch, F. S.; Tomkins, B. A. *Anal. Chem.* **1978**, *50*, 1429.
(26) Wold, S.; Albano, C.; Dunn, W. J., III; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjøstrøm, M.; skageberg, B.; Wikstrøm, C.; Øhman, J. "Multivariate Data Analysis: Converting Chemical Data Tables to Plots"; Presented at the VIIth International Conference on Computers in Chemical Research and Education, Garmisch-Partenkirchen, June 10–14, 1985.
(27) Wold, S.; Sjøstrøm, M. In *Chemometrics, Theory and Application*; Kowalski, B. R., Ed.; ACS Symposium Series 52; American Chemical Society: Washington, DC, 1977; pp 243–282.
(28) Grahl-Nielsen, O.; Kvalheim, O.; Øygard, K. *Anal. Chim. Acta* **1983**, *150*, 145.
(29) Stalling, D. L.; Dunn, W. J., III; Schwartz, T. R.; Hogan, J. W.; Petty, J. D.; Johansson, E.; Wold, S. In *Trace Residue Analysis, Chemometric Estimates of Sampling, Amount, and Error*; Kurtz, D. A., Ed.; ACS Symposium Series 284; American Chemical Society: Washington, DC, 1985; pp 195–234.
(30) Vogt, N. B.; Knutsen, H. *Mar. Ecol.: Prog. Ser.* **1985**, *26*, 145.
(31) Albano, C.; Blomquist, G.; Coomans, D.; Dunn, W. J., III; Edlund, U.; Eliasson, B.; Hellberg, S.; Johansson, E.; Norden, B.; Sjøstrøm, M.; Søderstrøm, B.; Wold, H.; Wold, S. *Proceedings of the Symposium on Applied Statistics*; NEUCC, RECKU & RECAU, DTH: Lyngby, Denmark, 1981; pp 183–218.
(32) Christie, O. H. J.; Wold, S. *Anal. Lett.* **1979**, *12*(A9), 979.
(33) Kvalheim, O. M.; Aksness, D.; Brekke, T.; Eide, M. O.; Sletten, E.; Telnaes, N. *Anal. Chem.* **1985**, *57*, 2858–2864.
(34) Wold, S., Umeå University, personal communication, 1985.
(35) Wold, S.; Albano, C.; Dunn, W. J., III; Esbensen, K.; Hellberg, S.; Johansson, E.; Sjøstrøm, M. In *Food Research and Data Analysis*; Martens, H.; Russwurm, H., Eds.; Applied Science: New York, 1983; pp 147–189.
(36) Grimmer, G.; Bøhnke, H. *J. Assoc. Off. Anal. Chem.* **1975**, *58*, 725.
(37) Vogt, N. B.; Aamot, E.; Krane, J.; Steinnes, E.; manuscript in preparation.
(38) Zar, J. H. *Biostatistical Analysis*; Prentice-Hall: Englewood Cliffs, NJ, 1974; pp 236–243.
(39) Wold, S. "Analysis of Chemical Data in Terms of Analogy and Similarity"; First International Symposium on Data Analysis and Informatics, 7–9 September 1977, Versailles; Textes des Communications: Paris, 1977; Vol. 2, p 683.
(40) Wold, S.; Albano, C.; Dunn, W. J., III; Edlund, U.; Esbensen, K.; Geladi, P.; Hellberg, S.; Johansson, E.; Lindberg, W.; Sjøstrøm, M. In *Proceedings of the NATO Advanced Study Institute on Chemometrics*, Cosenza, Italy; Kowalski, B. R., Ed.; Reidel: Dordrecht, Holland, 1983; pp 17–97.
(41) Albano, C.; Dunn, W. J., III; Edlund, U.; Johansson, E.; Norden, B.; Sjøstrøm, M.; Wold, S. *Anal. Chim. Acta Comput. Tech. Optim.* **1978**, *103*, 429.
(42) Johansson, E.; Wold, S.; Sjødin, K. *Anal. Chem.* **1984**, *56*, 1685.
(43) Kolset, K.; Nordenson, S.; Vogt, N. B.; Thrane, K. "An Approach to Interactive Normalization in Principal Component Analysis"; Presented at the IIIrd CAC Meeting of the International Chemometrics Society, S. Terenzo di Lerici, May 25–30, 1986.

Received for review May 29, 1986. Revised manuscript received August 7, 1986. Accepted August 18, 1986. Part of this work has been supported by the Royal Norwegian Council for Scientific and Industrial Research (NTNF).