Anal. Chem. 1995, 67, 2373–2378

Procrustes Rotation as a Way To Compare Different Sampling Seasons in Soils

A. Carlosena,† J. M. Andrade,*,† M. Kubista,‡ and D. Prada†

Analytical Chemistry Department of Universidade da Coruña, Campus A Zapateira s/n, E-15071, A Coruña, Spain, and Department of Biochemistry and Biophysics, Lundberg Laboratory, Chalmers University of Technology, Medicinaregatan 9C, S-41390 Gothemburg, Sweden

Procrustes rotation is a powerful technique to compare subspaces. In this paper, Procrustes rotation is used to select the minimum number of original variables that are important to describe a system and to statistically compare different sampling seasons. In the latter case, Procrustes rotation methodology has been applied to determine similarities and differences between sampling seasons in urban soils. Four seasons (during one year) were carried out to develop a monitoring scheme to control metal pollution in a medium-city area. Future efforts to monitor metal pollution will be based on analyzing only two or (at most) three metals, i.e., Pb, Co, and Cd.

When analytical studies (typically, environmental studies) are made, two questions are frequently posed by analytical chemists: First, should all initially considered variables be maintained in future samplings? Second, are all analytical tests necessary to describe the system?

Considering the latter question, what we look for (intuitively) is some (statistical) technique to find the minimum number of original variables that allows us to describe our system with sufficient accuracy. To answer the former question, some kind of intersampling seasonal comparison by means of some statistical technique is needed. Studying what happens in an interseasonal data sense, we should be able to describe which variables have a similar pattern/behavior along the different sampling seasons. Of course, the opposite will also be useful, i.e., to know what variables (if any) cause differences in the several data set patterns.

If we could apply some methodology to address both problems simultaneously, we could obtain not only a better understanding about the system but improvements in productivity, making use of lower resources, lower delay times, lower laboratory workloads, etc.

Several techniques have been considered with similar objectives. Good results have been reported by Grimalt et al.[1,2] using factor analysis and Krieg and Einax[3] and Krzanowski[4] using a discriminant way. Principal component analysis (PCA) and factor analysis are currently also used with similar purpose. These methods have a major disadvantage. They condense all the information in several abstract factors or principal components (PCs) which are combinations of the *overall* set of variables, and it is strictly not correct to delete (and select) variables taking into account their importance in these abstract factors.

PROCRUSTES ROTATION

Both topics can be solved by considering that the main idea is to compare two or more multidimensional data sets (e.g., one from each sampling season). This can be achieved using the (mathematically) well-known technique Procrustes rotation. In the Greek legend, Procrustes lodged travelers in his bed and during sleep either cut their legs or elongated them to fit them precisely into the bed.[5] In analogy with Procrustes himself, in Procrustes rotations the two sets of projections are rotated to a consensus target vector so as to match as closely as possible (including rotation, translation, stretching), in the least-squares sense.

This mathematical technique has been extensively studied by Krzanowski[6–8] and applied to the petrochemical industry field by Deane and MacFee[9] and, more recently, by Andrade et al.[10] Procrustes rotation is also extensively used in the spectroscopic area to identify pure spectra and to quantify analytes;[11,12] here the Procrustes technique is so powerful that typical calibration can be avoided. Recently, there has been one attempt to extend this interesting technique to the environmental field by Andrade et al.[13] The present work is one further step forward in that preliminary work where the interseasonal comparison is introduced. In this paper, we are not focused in presenting all the detailed algorithms but their general overview.

Selection of the Minimum Number of Original Variables To Describe One Sampling Season. Only a brief discussion is provided to center our work. A detailed mathematical treatment can be found in refs 8 and 13.

The main objective is to select redundant variables in raw data, which is equivalent to the identification of a subset of k variables that conveys the main structure of the raw data. Each variable is deleted in turn from the data set and two subspaces are compared by rotation, translation, and stretching by using the Procrustes rotation technique. The two compared subspaces are the scores

† Universidade da Coruña.

‡ Chalmers University of Technology.

(1) Grimalt, J. O.; Canton, L.; Olive, J. *Chem. Intell. Lab. Syst.* **1993**, *18*, 93–109.
(2) Grimalt, J. O.; Olive, J. *Anal. Chim. Acta* **1993**, *278*, 159–176.
(3) Krieg, M.; Einax, J. *Fresenius J. Anal. Chem.* **1994**, *348*, 490–495.
(4) Krzanowski, W. J. *J. Chemom.* **1992**, *6*, 97–102.
(5) Kubista, M.; Eriksson, S.; Albinsson, B. *Spectrosc. Eur.* **1992**, *4/6*, 28–30.
(6) Krzanowski, W. J. *Biometrics* **1987**, *43*, 575–584.
(7) Krzanowski, W. J. *Appl. Stat.* **1987**, *36*, 22–33.
(8) Krzanowski, W. J. *Principles of Multivariate Analysis: A User's Perspective*; Clarendon Press: Oxford, U.K., 1990.
(9) Deane, J. M.; MacFee, H. J. H. *J. Chemom.* **1989**, *3*, 477–491.
(10) Andrade, J. M.; Prada, D.; Muniategui, S. *J. Chemom.* **1993**, *7*, 427–438.
(11) Kubista, M.; Sjöback, R.; Albinsson, B. *Anal. Chem.* **1993**, *65*, 994–998.
(12) Scarminio, I.; Kubista, M. *Anal. Chem.* **1993**, *65*, 409–416.
(13) Andrade, J. M.; Prada, D.; Alonso, E.; López, P.; Muniategui, S.; de la Fuente, P. Quijano, M. A. *Anal. Chim. Acta* **1994**, *292*, 253–261.

0003-2700/95/0367-2373$9.00/0

from a PCA using all variables and series from a PCA obtained after deleting each variable. A statistic is defined, the "predicted residual error sum of squares", (PRESS)[6,14] (see eq 1). Important variables will produce larger PRESS values, and they should be maintained in the data set.

$$\mathrm{PRESS}(m) = \frac{1}{np}\sum_{i}^{n}\sum_{j}^{p}[\hat{X}_{ij}(m) - X_{ij}]^2 \qquad (1)$$

where m is the number of PCs considered, n the number of objects, p number of variables; X_{ij} one element of the original data matrix; and $\hat{X}_{ij}$ is the prediction of the X_{ij} element after "reconstruction" from the Procrustes rotation model, using m PCs.

An important problem here is to select the optimum number of PCs since the subspace comparison is made in a PCA base (to avoid noise and unuseful information). This is made by singular value decomposition and calculating the W_m statistic[6,10] (see eq 2). The W_m values are expressed in a cross-validation scheme and represent the increase in predictive information supplied by the mth component divided by the average predictive information in each of the remaining components. Important components should yield values of W_m greater than ~0.9.

$$W_m = \frac{[(\mathrm{PRESS}(m-1) - \mathrm{PRESS}(m))/D_m]}{(\mathrm{PRESS}(m)/D_r]} \qquad (2)$$

where m is the number of the PCs, PRESS was calculated above, and D_m and D_r are degrees of freedom ($D_m = n + p - 2m$; $D_r = (n-1)p - D_m$).

Using these tests, the subset of original variables will comprise an important percentage of the initial information and should preserve the principal patterns in the original data set.

Intercomparison of Sampling Seasons. First, a PCA must be made on each data set (one data set per sampling). The optimum number of factors has to be selected by using some standard method; e.g., the Malinowski test,[15-17] cross-validation, or the above-mentioned W_m statistic. Malinowski developed a test for determining the true dimensionality of a data set based on the Fisher variance ratio test. It is an empirical test which plot get a minimum when the optimum number of PCs is reached (see eq 3).

$$\mathrm{IND}(m) = \sqrt{\sum_{i=m+1}^{p}\frac{l_i}{n(p-m)^3}} \qquad (3)$$

where l is the corresponding eigenvalue, n and p are as above, and m is the corresponding number of the PCs.

It is our experience, that both the Malinowski test and the W_m statistic give good results and are adequate for most purposes. The Malinowski test is simpler and faster to implement in a program. Note also that it is not necessary to select the same number of PC (factors) in each data set.

The following treatment was originally developed by Krzanowski.[8] For simplicity, we will only consider two sampling seasons (two sample data sets) on which PCA has been made.

Let $\mathbf{X}$ and $\mathbf{Y}$ matrices $k \times p$, where k is the number of selected PCs and p is the number of original measured variables. $\mathbf{X}$ will contain the loadings from (e.g.) the first sampling season and $\mathbf{Y}$ the loadings from (e.g.) the second one.

Define $\mathbf{N} = \mathbf{XY'YX'}$ (=$\mathbf{TT'}$, where $\mathbf{T} = \mathbf{XY'}$ and the prime denoting transpose of matrix). Krzanowski proved the following two results:[8,18]

(1) The minimum angle between an arbitrary vector in the space of the first k principal components from the first sampling season and the one most nearly parallel to it in the space of the first k principal components of the second sampling season is given by $\cos^{-1}(\lambda_i)^{1/2}$, where λ_i is the largest eigenvalue of N.

(2) Define $\mathbf{b}_i = \mathbf{X'a}_i$ ($i = 1, ..., k$), where $\mathbf{a}_i$ is the λ_i associated eigenvector. Then $\mathbf{b}_i, ..., \mathbf{b}_k$ form a set of mutually orthogonal vectors embeded in the first subspace, and $\mathbf{Y'Yb}_i, ..., \mathbf{Y'Yb}_k$ form a corresponding set of mutually orthogonal vectors in the second season into which the differences between the subspaces can be partitioned. The angle between the ith pair $\mathbf{b}_i \leftrightarrow \mathbf{Y'Yb}_i$ is given by $\cos^{-1}(\lambda_i)^{1/2}$ ($i = 1, ..., k$).

The similarities between the sampling seasons can be exhibited solely by studying the pairs $\mathbf{b}_i \leftrightarrow \mathbf{Y'Yb}_i$ with λ_i being a measure of the contribution of the ith pair to the total similarity. Adding more pairs (new dimensions), planes and surface of similarities are obtained.

A quantification of the extent to which two r-dimensional portions differ is provided by the r critical angles $\cos^{-1}(\lambda_1)^{1/2}, ..., \cos^{-1}(\lambda_r)^{1/2}$. If the subspaces have very strong similarities there will be r angles close to zero. When the r values are large (close to 30 or greater) both spaces differ in that r-dimension.

A very important advantage of this approach is that although $\mathbf{b}_i$ and $\mathbf{Y'Yb}_i$ are mathematical vectors without simple meaning, we can define a *consensus vector* directly linked with the original measured variables. The consensus vector is a new vector that is closest to both $\mathbf{b}_i$ and $\mathbf{Y'Yb}_i$; it is the bisector of the angle between them. The bisector is given by eq 4, where $\mathbf{I}$ is the ($p \times p$) identity matrix.[8]

$$\mathbf{c}_i = [2(1 + \sqrt{\lambda_i})]^{-1/2}\left(I + \frac{1}{\sqrt{\lambda_i}}\mathbf{Y'Y}\right)\mathbf{b}_i \qquad (4)$$

The set of $\mathbf{c}_i, ..., \mathbf{c}_k$ defines a k-dimensional subspace that is the *average* or consensus of both sampling season data sets, and most important, it can be given a chemical interpretation. The sole restriction of the model is that all the considered data sets must have exactly the same variables.

The idea behind this mathematical treatment is to rotate, translate, and stretch one PC subspace to resemble the other subspace as much as possible. This is why Procrustes rotation is so useful. If similarities are found, the original subspaces (original seasonal data sets) will have corresponding similarities (and, of course, the same holds for differences).

(14) Osten, D. W. *J. Chemom.* **1988**, *2*, 39–48.
(15) Cela, R., Ed. *Avances en Quimiometria Práctica*; University of Santiago: Santiago, Spain, 1994.
(16) Brereton, R. G., Ed. *Multivariate Pattern Recognition in Chemometrics*; Elsevier: New York, 1992.
(17) Malinowski, E. R. *Factor Analysis in Chemistry*, 2nd ed.; John Wiley and Sons: New York, 1991.
(18) Krzanowski, W. J. *JASA, J. Am. Stat. Assoc.* **1979**, *74*, 703–707 (correction in **1981**, *76*, 1022).

Table 1. PC Loadings, Sampling Season 1

variable	PC1	PC2	PC3	PC4	PC5
Cd	−0.33	0.33	0.05	−0.20	−0.16
Co	0.41	0.28	-0.05	0.03	0.13
Cu	−0.31	0.34	0.21	−0.06	0.36
Cr	0.31	0.39	-0.07	−0.07	0.24
Fe	0.42	0.15	0.23	−0.09	0.01
Mn	0.19	0.03	0.65	0.13	−0.04
Ni	0.26	0.35	-0.01	−0.05	0.15
Pb	−0.28	0.40	0.08	−0.14	0.03
Zn	−0.35	0.27	0.16	−0.13	−0.22
humidity	−0.20	-0.31	0.32	0.06	0.74
LOI	0.12	-0.20	0.55	−0.35	−0.33
pH	−0.01	0.20	0.18	0.87	−0.19
% explained variance	35.0	22.1	12.8	8.0	5.5
% cumulative variance	35.0	57.1	69.9	77.9	83.5

SAMPLING AND ANALYTICAL VARIABLES

Four sampling seasons (1, fall; 2, winter; 3, spring; 4, summer) were organized to study soil contamination in a medium-size city and its neighborhood (~300 000 inhabitants). The studied area includes one highway, a medium-size city (La Coruña), an industrial area, and one main avenue with very high traffic levels.

Soils were taken in public gardens, uncultivated fields, and across several perpendicular transects along the highway. Several sampling sites were considered in each transect. Samples were taken from 0 to 5 cm depth, air-dried, ground, and sieved through a 2 mm mesh sieve in order to eliminate gravel, stones, and large fragments. After that, samples were dried by heating at 60 °C for 48 h and sieved to <0.2 mm. The fraction of soil that was <2 mm was used to determine moisture content (at 105 °C), pH (1: 2.5 in water), and organic material as loss on ignition (LOI) at 450 °C for 6 h.[19,20]

Exactly 0.3 g aliquots of <0.2 mm soil fractions were subjected to the chemical extraction procedure, using HNO_3(conc), and microwave-heated in Teflon vessels.[21]

Twelve variables were analyzed in each of the 95 samples. These were Cd, Co, Cu, Cr, Fe, Mn, Ni, Pb, and Zn concentrations, humidity, pH, and LOI. Metals were analyzed using flame or graphite furnace AAS when concentrations were low (Cd, Cr, Co, Ni). Accuracy was checked using certified reference materials (BCR, CRM141, calcareous loam soil; BCR, CRM277, estuarine sediment). Good agreement was obtained when certified and experimental values were compared. None of the studied variables had a standard deviation of zero.

RESULTS AND DISCUSSION

Principal Component Analysis. Tables 1−4, summarize the main results from the PCA studies (rotation has not been applied) for each sampling season. Loadings are given with only two significant digits, and five principal components are presented for each season. The cumulative percentage of explained variance lies between 75 and 80%. Note that the variances explained by each first, second, third, etc., PC are quite similar among the different sampling seasons.

(19) *M.A.P.A. Official Methods of Analyses*; Secretaría General Tecnica del Ministerio de Agricultura, Pesca y Alimentación: Madrid, Spain, 1986; Vol. III.

(20) American Association for Testing and Materials. *Annual Book of ASTM Standards*; ASTM: Philadelphia, 1993; Vol 04.08, pp 360−361.

(21) Carlosena, A.; Prada, D.; Muniategui, S.; Lopez, P.; Andrade, J. M.; Gonzalez, E. 2nd International Conference on Polluted Soils. IHOBE, Vitoria-Gasteiz, Spain, September 21−22, 1994.

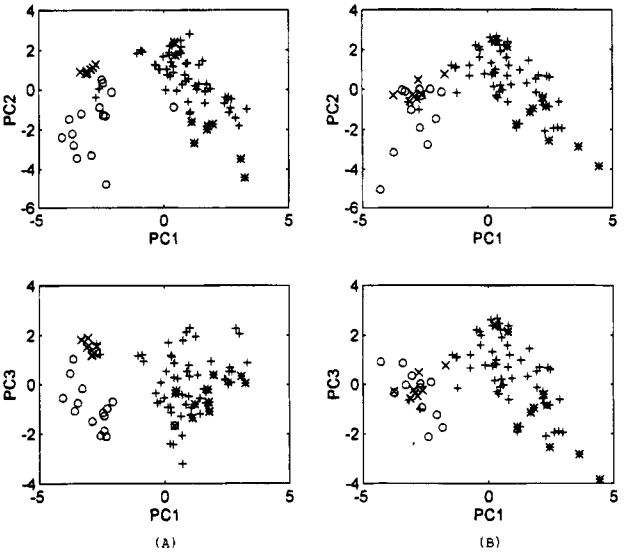

Figure 1. Soils distribution in the PC1−PC2 and PC1−PC3 score subspaces: (A) first sampling season; (B) second sampling season. Highway samples (○), transect samples (×), main avenue with high traffic levels (*), and city gardens, (+).

First Season. Table 1 shows that the first PC is related to Fe, Co, and Zn, being strongly associated to the natural soils' variability. The second PC, associated with Pb, Cr, Ni, Cu, and Cd, shows a clear relation to anthropogenic pollution sources. PC 3 is associated with Mn and LOI; the fourth PC is defined essentially by pH and the fifth PC by humidity. We will analyze the first sampling season in some more detail since it will be useful to understand the below described results and discussion.

The first PC shows differences beween two esential blocks of samples: gardens in the city, samples along the highway transects, and the highway itself. As illustrated by Figure 1A, a simple classification rule is to use the sample score. It is zero or close to zero for the gardens' samples. These have low values in Co and Fe and high values in Zn (~6 ppm Co, 20 g/kg Fe, and 300 ppm Zn). Clearly negative scores define samples from the second area (highway and its transects); these have high values in Co and Fe (~20 ppm and 35 g/kg, respectively), and lower values in Zn (~100 ppm).

The second PC discriminates between subareas in the two groups. Scores near zero or negative are related to samples taken very close to the highway (sampling distance lower than 0.3 m from the border). Samples from different distances from the transects form a group with clearly positive scores. Highway samples have values like [Pb] ≈ 300 ppm; [Cr] ≈ 60 ppm, and [Ni] ≈ 100 ppm or sometimes even higher. Samples from transects have lower values: [Pb] < 20 ppm; [Cr] ≈ 40 ppm, and [Ni] ≈ 20 ppm (see Figure 1).

Samples taken close to the main avenue with high traffic levels, have high, positive scores ([Pb] ≥ 500 ppm, [Cr] ≈ 30 ppm, [Ni] ≈ 35 ppm), whereas the rest of the points (see Figure 1) are from the city gardens (Pb variability, from 100 to 300 ppm; [Cr] ≈ 25 ppm, and [Ni] ≈ 20 ppm). PCs 3−5 do not add more differentiation.

So we can discriminate between four groups of samples: highway samples (○), transect samples (×), main avenue with high traffic levels (*), and city gardens (+).

Second Season. Table 2 shows the loadings from each of the five most important PCs obtained from second season data. The

Table 2. PC Loadings, Sampling Season 2

variable	PC1	PC2	PC3	PC4	PC5
Cd	−0.38	0.26	0.12	−0.12	−0.06
Co	0.37	0.35	−0.06	−0.07	−0.07
Cu	−0.34	0.32	0.21	0.05	0.04
Cr	0.27	0.43	0.04	−0.01	−0.12
Fe	0.41	0.27	0.17	0.12	0.03
Mn	0.22	0.13	0.15	0.69	0.11
Ni	0.09	0.42	−0.35	−0.23	−0.45
Pb	−0.33	0.35	0.07	0.07	0.12
Zn	−0.38	0.27	0.16	0.04	−0.01
humidity	0.19	0.18	0.11	−0.47	0.77
LOI	0.08	−0.02	0.69	0.07	−0.14
pH	−0.13	0.17	−0.49	0.44	0.37
% explained variance	34.7	19.8	12.3	9.4	6.6
% cumulative variance	34.7	54.5	66.8	76.2	82.8

Table 3. PC Loadings, Sampling Season 3

variable	PC1	PC2	PC3	PC4	PC5
Cd	−0.31	0.36	0.10	−0.05	−0.08
Co	0.41	0.22	−0.11	−0.06	−0.06
Cu	−0.24	0.32	0.08	0.41	0.45
Cr	0.30	0.43	−0.04	−0.15	−0.06
Fe	0.42	0.14	−0.12	0.16	−0.01
Mn	0.22	−0.01	−0.46	0.66	−0.12
Ni	0.25	0.43	−0.01	−0.24	−0.01
Pb	−0.25	0.43	0.02	0.14	−0.07
Zn	−0.32	0.35	0.00	0.06	0.02
humidity	0.22	0.09	0.30	0.26	0.86
LOI	0.08	−0.11	0.63	0.42	−0.40
pH	−0.29	−0.05	−0.50	0.13	0.14
% explained variance	36.8	21.4	11.3	7.9	6.4
% cumulative variance	36.8	58.2	69.5	77.4	83.8

first, associated with Fe, Zn, Cd, and Co, denotes a natural variability among soils. The second PC, related to Cr, Ni, and Pb, reveals the influence of human pollution. The third and fourth PCs are related to LOI and Mn, respectively.

Figure 1B shows the bidimensional projection scores PC1−PC2 and PC1−PC3, revealing a pattern very similar to the first sampling season, although it can be seen that the four groups of samples are not so clearly distinguished. This fact can be attributed to the climatological conditions linked to the winter season. Heavy winter rainfalls may homogenize the different polluted areas.

Third Season. Table 3 summarizes the loadings for the third season. Again, a natural pattern (first PC involves Fe, Co, and Zn) and a pollution pattern (second PC defined mostly by Pb, Ni, and Cr) can be observed. Figure 2 reveals a sample score distribution close to the first seasons. In spite of this, the main avenue vs city gardens and highway vs transects differentiation are not always distinguishable (we attribute this fact to the spring rainfalls in La Coruña, northwest of Spain).

Fourth Season. Table 4 shows the first five loadings for this sampling season. Good similarities are observed when compared to the first season (see Figure 2). The interpretation is as above.

Selection of the Minimum Number of Variables To Describe Each Sampling Season. The first problem to be addressed is to select the optimum number of principal components to describe each sampling season. As stated above, the W_m statistic or the Malinowski test is useful. In our studies, both statistics gave similar conclusions. From Table 5, it can be seen

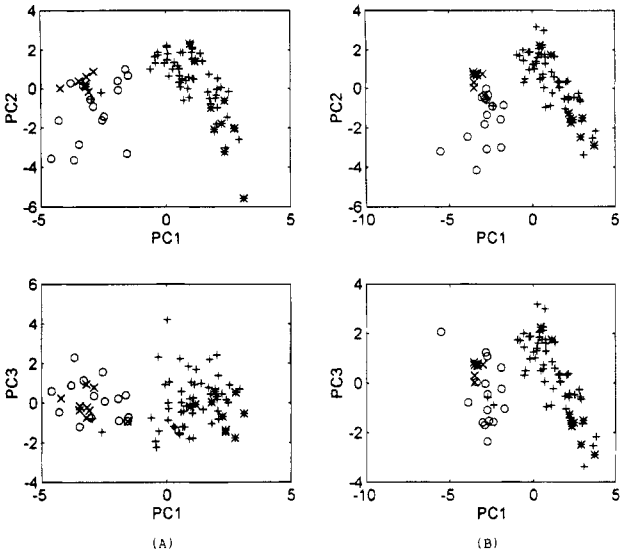

Figure 2. Soil distribution in the PC1−PC2 and PC1−PC3 score subspaces: (A) third sampling season; (B) fourth sampling season.

Table 4. PC Loadings, Sampling Season 4

variable	PC1	PC2	PC3	PC4	PC5
Cd	−0.35	0.29	0.10	0.06	0.02
Co	0.38	0.30	0.02	−0.02	−0.03
Cu	−0.30	0.30	0.22	0.01	−0.30
Cr	0.24	0.47	0.01	−0.14	0.23
Fe	0.40	0.19	0.07	0.07	−0.12
Mn	0.28	0.05	0.26	0.53	−0.58
Ni	0.22	0.44	0.05	−0.07	0.11
Pb	−0.30	0.32	0.20	−0.06	0.07
Zn	−0.35	0.25	0.24	0.02	0.03
humidity	0.06	−0.18	0.55	0.51	0.59
LOI	0.00	−0.25	0.60	−0.45	−0.32
pH	−0.29	0.10	−0.32	0.46	−0.19
% explained variance	41.2	20.2	11.3	6.8	6.2
% cumulative variance	41.2	61.4	72.7	79.5	85.8

Table 5. Selection of the Number of PCs Using the W_m Statistic

eigenvalue ordering	sampling season			
	1st	2nd	3rd	4th
1	1.87	2.19	2.51	3.87
2	2.72	2.16	2.77	2.19
3	0.84	0.32	0.44	0.67
4	0.39	0.26	0.08	0.74
5	−0.11	0.56	0.19	0.37
6	−0.32	0.14	0.23	−0.28
7	0.03	0.04	−0.37	−0.24
8	0.09	0.29	0.14	−0.21
9	−0.29	−0.41	0.08	−0.07
10	0.07	−0.35	0.09	0.28
11	−0.19	−0.06	−0.18	−0.12

that the W_m statistic strongly suggests two PCs as the most important ones for describing our data sets.

When the Malinowski's test is applied, a minimum is reached in about two or three. We found that two principal components are sufficient to describe the important information in all the seasonal data sets. In fact, this is a positive coincidence since there is no critical reason to suppose this equality, except that two PCs have been sufficient to adequately describe the different sample patterns in all four cases.

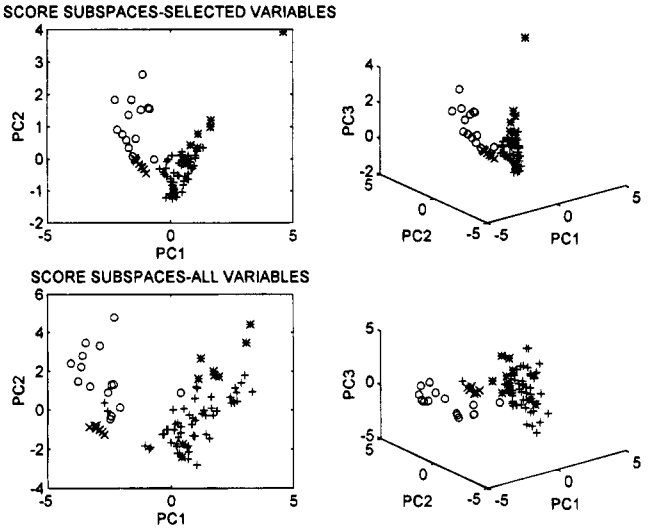

Figure 3. First sampling season: PC scores before and after selection of the most important variables were applied.

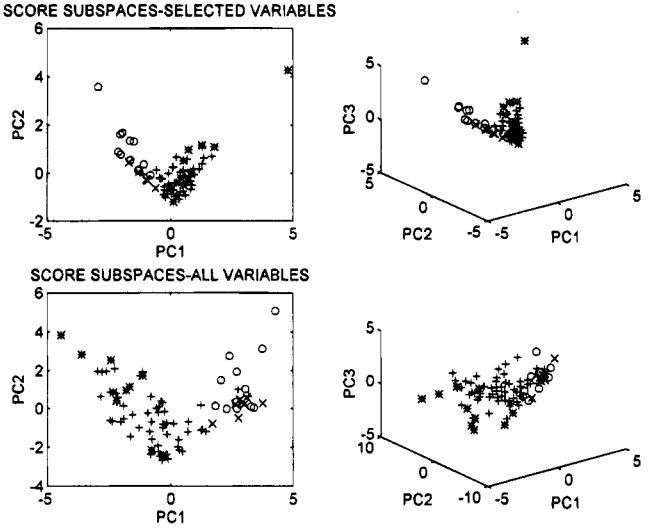

Figure 4. Second sampling season: PC scores before and after selection of the most important variables were applied.

Table 6. Selected Subsets of Variables

sampling season	selected variables	last retained variable	no. of PCs
1	Co, Pb	Cd	2
2	Co, Pb	Cd	2
3	Co, Pb	Cd	2
4	Pb, Cd	Co	2

In Procrustes analysis, if k PCs best describe our system, a minimum of k variables will be considered as the most important ones. Table 6 tabulates the variables selected for each sampling season. If we consider only the last variable being deleted in turn, a perfect agreement is observed, revealing that all the sampling seasons are, principally, characterized by the same variables. It is noteworthy that these selected variables characterize the artificial pollution.

Figures 3–6 compare the original PC1–PC2 and PC1–PC2–PC3 score subspaces (all variables) and those obtained after variable selection. A perfect agreement is observed. Even the "V" shape observed when all PCs are used becomes more

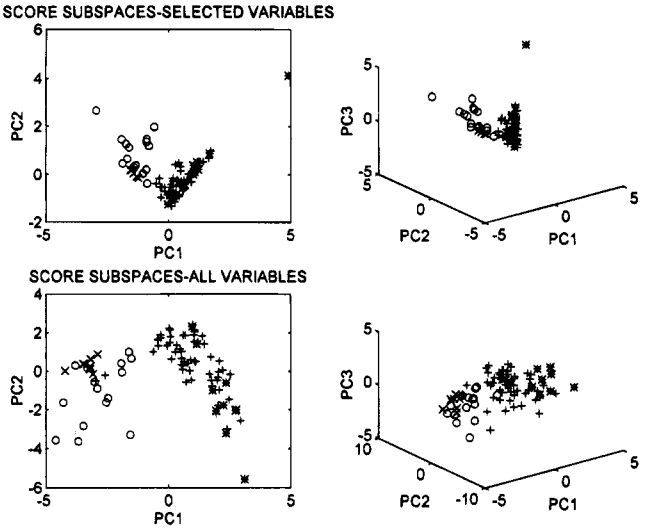

Figure 5. Third sampling season: PC scores before and after selection of the most important variables were applied.

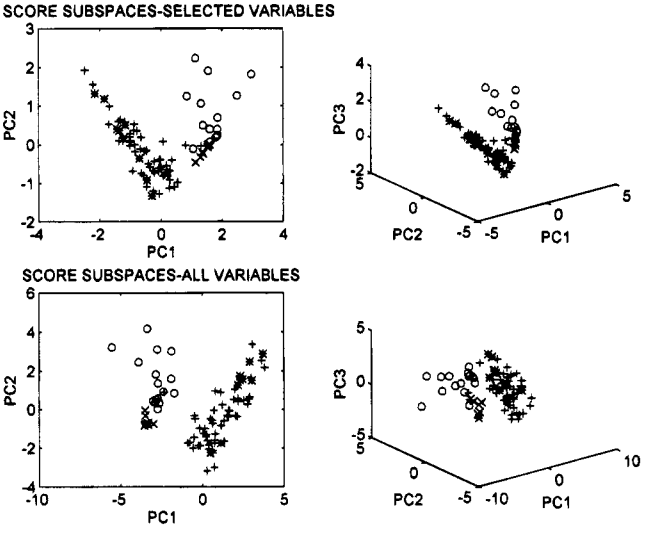

Figure 6. Fourth sampling season: PC scores before and after selection of the most important variables were applied.

pronounced in the subspaces obtained after variable selection. This is a property of the Procrustes rotation behavior, a reduction in the intracluster distance is obtained, probably because of noise reduction. In the second, third, and fourth season data (see Figures 4–6), a sign change in either PC1 or PC2 is observed upon variable reduction, but this has no relevance because it is caused by a lack of a unique mathematical origin of coordinates when different PCAs are made.

Of course, some information is lost when variables are excluded, but with only two variables, we still describe the samples' characteristics very well.

Intercomparison of Different Sampling Seasons. We are looking for some kind of consensus subspace that would comprise the main similarities (and/or differences) between four different data subspaces. The first step has been addressed in the preceding topic; the number of relevant principal components was found to be two. In spite of this, we have considered three and even four PCs on each data set to get a broad background for our comparison.

Table 7 summarizes these studies. For each dimension in the original subspaces (i.e., two, three, and four PCs) the consensus

Table 7. Consensus Subspaces and Vectors

	no. of PCs in the original subspaces and consensus vectors for each size								
	2 PC		3 PC			4 PC			
variable									
Cd	0.46	−0.06	0.46	0.00	0.11	0.46	−0.05	−0.07	−0.10
Co	−0.16	−0.49	−0.10	0.49	−0.05	−0.07	0.47	0.06	−0.17
Cu	0.43	−0.10	0.44	0.05	0.21	0.47	0.04	−0.16	0.12
Cr	0.02	−0.54	0.07	0.52	0.00	0.10	0.47	0.02	−0.26
Fe	−0.24	−0.40	−0.18	0.45	0.11	−0.12	0.48	−0.09	0.05
Mn	−0.17	−0.16	−0.13	0.26	0.19	−0.03	0.41	−0.05	0.70
Ni	0.07	−0.48	0.13	0.45	−0.12	0.13	0.38	0.11	−0.35
Pb	0.46	−0.15	0.47	0.09	0.12	0.49	0.06	−0.08	−0.03
Zn	0.45	−0.03	0.46	0.00	0.18	0.47	−0.03	−0.11	0.04
humidity	−0.10	0.01	−0.09	0.02	0.43	−0.03	0.07	−0.32	0.18
LOI	−0.15	0.10	−0.15	−0.07	0.72	−0.07	0.00	−0.71	0.12
pH	0.22	0.04	0.22	−0.02	−0.36	0.20	0.01	0.55	0.46
	angles								
	2 PC		3 PC			4 PC			
sampling 1	6.58	23.97	5.40	19.93	41.55	4.45	19.08	17.18	17.14
sampling 2	6.35	19.01	5.47	15.22	22.09	3.88	9.45	14.34	41.16
sampling 3	3.66	17.65	4.17	10.58	43.11	2.60	9.71	14.09	13.17
sampling 4	6.58	10.36	4.76	9.01	13.65	4.01	11.64	10.76	37.05

vectors are obtained, and one angle (expressed in degrees; see lower boxes in Table 7) is calculated to obtain a measure of the degree of similarity between each consensus vector (plane, surface) in the reduced set and each vector (plane, surface) in the original subspaces.

It can be seen that lowest angles are always associated with the first and second consensus vectors, thus comprising major similarities between the four sampling seasons. The third, fourth, etc., vectors reveal differences as expected. Once the consensus vectors have been obtained, a combined study between loadings from each season and the consensus loadings is needed to convert consensus loadings into something chemically meaningful. The first consensus vector is defined (principally) by Pb, Cd, Zn, and Cu. These are the variables for which the sample scatter is most similar in all four sampling seasons.

Pb and Cu were not in the individual season's first PCs although they are in the first consensus vectors. This shows they do not best explain the overall sample variability on each season (which is reasonable since the natural soil variability is too strong), but they are important when looking for similarities in the different sample patterns (distributions).

The second consensus vector, linked to Cr, Co, and Ni, also describes similarities between the four sampling seasons as it can be deduced from the angle values.

Consider now (artificially) three or four dimensions for the original data sets. The third consensus vector is principally defined by LOI; its values for the first sampling season are quite different from the other seasons (no clear reason has been found), so this is the (consensus) vector where differences arise. The fourth consensus vector, related to Mn, presents angles as large as 40°. This low correlation is in agreement with the high variability in Mn values observed from sample to sample and from one sampling season to another.

Finally, it is worth noting that the two selected variables agree with the variables that in the first consensus vector best describe similarities between sampling seasons. This reveals that Pb and Co (and even Cd) are the key variables when soil pollution is monitored in the areas covered by these studies. If environmental disasters do not occur, the present pollution sources do not change significantly, or both, these are the most suitable variables for a simple, fast pollution monitoring scheme.

Received for review January 18, 1995. Accepted April 15, 1995.[⊗]

AC950057H

[⊗] Abstract published in *Advance ACS Abstracts,* June 1, 1995.