ELSEVIER

Analytica Chimica Acta 310 (1995) 15–25

ANALYTICA CHIMICA ACTA

Pollution of a river basin and its evolution with time studied by multivariate statistical analysis

Roberto Aruga [a,*], Daniela Gastaldi [a], Giovanni Negro [b], Giorgio Ostacoli [a]

[a] *Department of Analytical Chemistry, University of Turin, Via Giuria 5, 10125 Turin, Italy*
[b] *Assessorship for the Environment, Regione Piemonte, Via Principe Amedeo 17, 10183 Turin, Italy*

Received 3 October 1994; revised 1 February 1995; accepted 15 February 1995

Abstract

A set of quantitative analytical data for 13 rivers belonging to the basin of the Tanaro (Piedmont region, north-western Italy) has been processed by multivariate statistical techniques. The original matrix, which refers to the year 1990, consisted of 20 chemical and physico-chemical variables, determined at 44 sampling sites. The following methods have been used for the treatment of the data: cluster analysis (unsupervised pattern recognition), factor analysis, variable selection by Fisher weights. A comparison has also been made between the data referring to 1990 and to 1978 in order to investigate the evolution of the environmental situation for the Tanaro basin after such a lapse of time. This comparison has been carried out by processing the 3-D matrix which collects the data of the two different periods.

Keywords: Chemometrics; Clustering; Factor analysis; Multivariate analysis; Pollution; River pollution

1. Introduction

Multivariate treatment of environmental data and, in particular, of data referring to surface waters, has not attained a diffusion comparable to that concerning other kinds of analytical data, such as clinical or food data. It must also be noted that the most part of multivariate treatments deal with natural phenomena, of geochemical interest, and that a smaller number of studies deal with pollution phenomena deriving from anthropic activities [1–3].

In order to continue and to complete previous studies on the pollution of single rivers [4,5], it was thought of interest to extend this kind of treatment to the pollution phenomena of a whole fluvial basin. Precisely, the analytical data belonging to the Tanaro river and to a set of its tributaries, located in the Piedmont region (north western Italy) have been treated using statistical multivariate techniques. The data have been obtained in the course of a Survey of Hydrological Resources, which was carried out by "Regione Piemonte" (the official administrative organization of the Piedmont region) and refer to the year 1990 [6].

For several sampling sites and variables investigated in the 1990 survey, the corresponding data obtained in a previous survey performed in the year 1978 [7] were also available. Therefore, in this work, it has been thought of interest to make also a comparison between data referring to the two different surveys, with the aim of drawing some conclusions

[*] Corresponding author.

SSDI 0003-2670(95)00101-8

on the variation of pollution of a fluvial basin over a 12 year period.

2. Experimental data and methods

The sampling sites and methods have been fixed by the technical organization promoter of the survey (the Assessorship for the Environment of the Piedmont Region) and, for those concerning the 1978 survey, also by the Department of Analytical Chemistry of the University of Turin, in conformity with previously published criteria [6–9].

Samples were usually collected from the middle of the stream, at a depth of 15 cm. Drawing times were arranged in such a manner that any bias deriving from the periodicity connected with seasons (i.e., with the flow) or with the different hours of the day, was avoided [4].

Analytical parameters have been determined with reference to official methods currently suggested [8,9]. Metals, in particular, have been determined by flame atomic absorption spectroscopy. The chemicals used were always of Analytical Reagent Grade. The purity of standards was checked periodically, together with the purity of distilled water and that of the detergents used for glassware. Chemical analyses were carried out by Chemical Laboratories of the National Health Service of Piedmont Region. Tests to compare the within- and between-laboratory variance were made in the course of the survey.

The following 20 chemical parameters have been considered in the statistical treatment for the 1990 survey (the abbreviations used in the present text are in parentheses): aluminium, chloride, chemical oxygen demand (COD), conductance (CON), total chromium (CrT), copper, filtered copper (CuF), dissolved oxygen (DO), iron, manganese, ammonia, nickel, nitrate, nitrite, lead, pH, suspended matter (SM), sulphate, anionic surfactants (SUR), zinc. Only 12 of these parameters have been considered in the comparison between the two different surveys: these common parameters are reported with an asterisk in Table 1, where all the parameters are listed. The values have been expressed in terms of the following units (or subunits) of measure: μs for conductance; mg/l for Cl, COD, DO, NH_3, NO_2, NO_3, SM, SO_4, SUR; μg/l for the others, pH excluded.

Table 1
Mean and estimated standard deviation (s) for the experimental parameters [a]

Parameter	Mean	s	Mean	s	Mean	s
Al[b]	106	97	–	–	–	–
Cl[c]	34.7	37.3	–	–	–	–
* COD[c]	17.3	23.0	17.7	23.2	54.9	88.7
* CON[d]	558	322	569	300	447	227
CrT[b]	2.74	3.32	–	–	–	–
* Cu[b]	19.6	23.9	20.8	23.6	58.4	86.3
CuF[b]	11.4	12.8	–	–	–	–
* DO[c]	10.3	1.9	10.2	1.6	10.0	2.15
Fe[b]	516	540	–	–	–	–
* Mn[b]	146	268	171	328	126	246
* NH_3[c]	0.825	1.29	0.76	0.94	2.67	5.20
* Ni[b]	8.7	12.0	9.75	12.1	6.2	19.4
* NO_2[c]	0.027	0.029	0.032	0.030	0.124	0.179
* NO_3[c]	1.56	1.21	1.76	1.28	2.32	1.18
* Pb[b]	3.94	4.06	4.25	4.24	1.71	4.10
* pH	7.99	0.18	7.97	0.17	7.77	0.43
SM[c]	15.8	19.1	–	–	–	–
SO_4[c]	84.5	100	–	–	–	–
* SUR[c]	0.14	0.265	0.15	0.29	2.09	4.62
Zn[b]	101	138	–	–	–	–

[a] The parameters indicated with * have also been considered in the comparative study of the surveys of 1990 and 1978. At left: values for the 44 stations of the survey of 1990; centre: values for the 24 stations of the survey of 1990 compared with those of 1978; at right: values for the 24 stations of the survey of 1978.
[b] μg/l.
[c] mg/l.
[d] μS.

In the data treatment for the 1990 survey, 44 sampling sites have been considered in all. Thirteen of these sampling stations are located on the Tanaro river (symbol: TA). The remaining sites are located on 12 rivers confluent, directly or indirectly, into the Tanaro river. The rivers considered, besides the Tanaro, are the following (the corresponding abbreviation and the number of sampling sites are shown for each river): Corsaglia (C, 2); Ellero (E, 1); Pesio (P, 1); Stura di Demonte (SD, 6); Borbore (BB, 1); Versa (V, 1); Tiglione (TI, 2); Belbo (BL, 7); Bormida di Millesimo (BM, 5); Stura di Ovada (SO, 1); Lemme (L, 1); Orba (O, 3). In relation to the particular kind of pollution of the Bormida di Millesimo, some organic parameters were also determined for this river; their values were found to amount to few μg/l and have not been considered in this study.

The whole of 44 sampling stations on the 13 rivers is reported in Table 2, together with the symbol used for each of these and the distance in km from the ending point (i.e., from the confluence with another river). A distance of 0 km means that the sampling site lies immediately before the confluence point (distance < 0.5 km). Only 24 of these sampling stations have been considered in the comparison between the two surveys: they are reported with an asterisk in Table 2.

The course of the fore mentioned rivers, together with the sampling station position and the most important towns are showed in Fig. 1.

The original data matrix consisted of 44 rows and 20 columns (i.e., 880 data in all) for the 1990 survey and of 24 rows, 12 columns and two cases corresponding to the two periods (i.e., 576 values in all) for the comparison between the two surveys. Considering that the temporal distribution of sampling was arranged in such a way as to eliminate any influence by periodical phenomena, the arithmetical mean has been taken into account for each parameter and each station. The use of mean values, instead of all the experimental values obtained, leads to more immediate interpretations of graphical representations and to a simpler description of the system. A summary of the experimental data, including the mean value and the standard deviation for each variable, is reported in Table 1.

3. Statistical treatment of the data

A preliminary step of the treatment consisted of the normalization of the raw analytical data, so as to avoid misclassifications due to the different order of magnitude and range of variation of the analytical parameters. Autoscaling, or z-transform, was used for this purpose. The values of each variable, after this scaling procedure, are characterized by zero as the mean value and by unit variance [10].

As they are coming into widespread use, only a brief account will be given on the chemometric techniques used here, together with references to specific works for fuller details. The programs contained in the statistical package PARVUS [11] were used in the calculations. Hierarchical agglomerative clustering was performed on the scaled data by

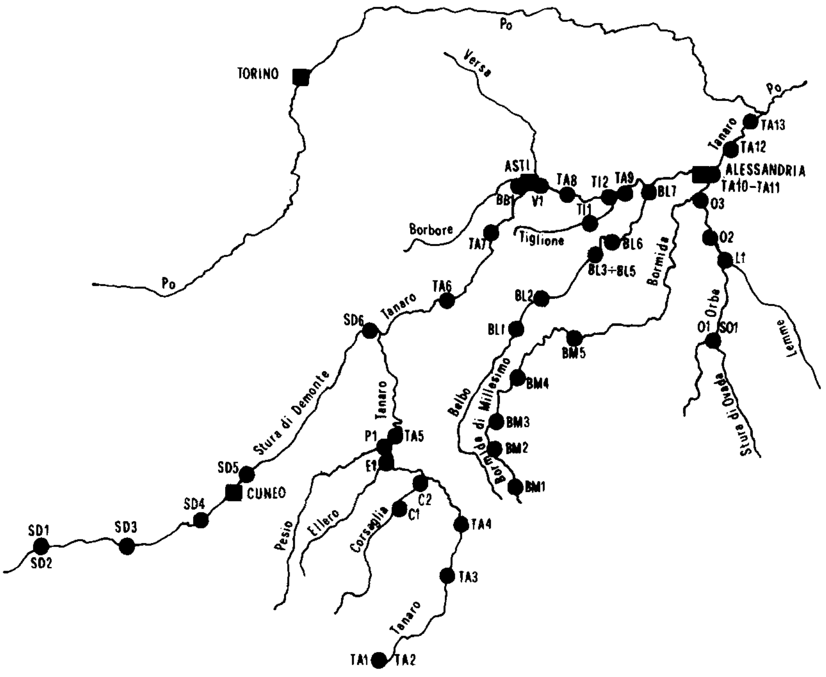

Fig. 1. Schematic representation of rivers, stations and main towns (see also Table 2).

Table 2
Sampling stations considered in the survey of 1990 (see text) [a]

Code	River	Site	km
C1	Corsaglia	S. Michele Mondovi'	10
* C2	Corsaglia	Lesegno	4
* E1	Ellero	Bastia	0
P1	Pesio	Carru'	0
SD1	Stura di Demonte	Vinadio 1	84
* SD2	Stura di Demonte	Vinadio 2	82
SD3	Stura di Demonte	Demonte	68
* SD4	Stura di Demonte	Borgo S. Dalmazzo	53
* SD5	Stura di Demonte	Cuneo 2	42
* SD6	Stura di Demonte	Cherasco	3
* BB1	Borbore	Asti 3	0
V1	Versa	Asti 3	0
TI1	Tiglione	Cortiglione 2	6
* TI2	Tiglione	Masio 2	0
BL1	Belbo	Cossano Belbo	51
* BL2	Belbo	S. Stefano Belbo	41
BL3	Belbo	Nizza Monferrato 1	31
* BL4	Belbo	Nizza Monferrato 2	28
BL5	Belbo	Nizza Monferrato 3	26
* BL6	Belbo	Castelnuovo Belbo	16
BL7	Belbo	Oviglio 1	3
* BM1	Bormida Millesimo	Saliceto	135
* BM2	Bormida Millesimo	Monesiglio	121
* BM3	Bormida Millesimo	Gorzegno	110
* BM4	Bormida Millesimo	Cortemilia	96
BM5	Bormida Millesimo	Monastero Bormida	76
* SO1	Stura di Ovada	Ovada 2	0
* L1	Lemme	Basaluzzo	0
O1	Orba	Ovada 2	29
O2	Orba	Casal Cermelli	9
* O3	Orba	Castellazzo Bormida	0
TA1	Tanaro	Ormea 1	212
TA2	Tanaro	Ormea 2	206
* TA3	Tanaro	Priola	188
TA4	Tanaro	Nucetto	176
* TA5	Tanaro	Clavesana	142
TA6	Tanaro	Alba 2	104
* TA7	Tanaro	S. Martino Alfieri	100
TA8	Tanaro	Castello di Annone	58
TA9	Tanaro	Masio 3	47
* TA10	Tanaro	Alessandria 1	32
* TA11	Tanaro	Alessandria 4	22
TA12	Tanaro	Montecastello	13
* TA13	Tanaro	Bassignana	6

[a] The sampling stations indicated with * have also been considered in the comparative study of the surveys of 1990 and 1978.

means of complete linkage, average linkage (weighted pair groups), average linkage (unweighted), and Ward's methods [10]. Square Euclidean distances were used for Ward's method and Euclidean distances for the others. Cluster analysis was also based on correlation coefficients in some cases.

Factor analysis was performed by a previous evaluation of principal components and computing the eigenvectors according to the method of Malinowski [12]. The rotation of the principal components was carried out by the Varimax method, using two different options: Raw Varimax and Eigenvalue Weighted Varimax [11]. The former performs a rigid (i.e., orthogonal) rotation, the latter does not keep the orthogonality of the axes. Since the differences between the two options were not appreciable, the results obtained by the first option have been considered in general. In one case, only, as the angles between the rotated axes were considerably different from 90° and different factors were identified using the two options, the results obtained by the Eigenvalue Weighted Varimax option have been considered.

Selection of variables was also carried out. It gives a sequence of variables on the basis of their decreasing information content (i.e., on the basis of their decreasing ability to discriminate objects belonging to different classes), and, in some instances, it allows to simplify the original data matrix by eliminating variables at the end of the sequence. Univariate selection was performed in this case; it

Table 3
Loadings of 20 experimental variables on four significant factors rotated according to the Varimax method (survey of 1990) [a]

F1	−0.0810	0.3254	0.3897	0.2415	0.0191
	0.0166	0.0161	−0.3381	0.2173	0.3696
	0.2911	−0.0723	0.0955	−0.1504	−0.0742
	−0.2651	0.1229	−0,0331	0.4015	−0.0585
F2	0.2749	−0.1115	−0.0336	−0.0340	0.2388
	0.4740	0.4872	−0.0473	0.0507	−0.0068
	−0.0036	0.0231	0.0077	−0.1967	0.4221
	−0.2738	−0.2016	−0.0427	0.0074	0.2256
F3	0.1176	−0.0001	−0.1436	0.0212	0.2173
	−0.0236	−0.0468	−0.0949	0.3722	−0.0685
	0.1179	0.5530	0.4608	0.4559	−0.0536
	−0.0983	0.0467	0.0101	−0.0863	0.0102
F4	0.1579	−0.1815	−0.0056	−0.4285	0.0812
	0.0204	0.0164	0.0628	0.1988	0.0757
	−0.0278	−0.0132	0.0464	−0.1970	−0.1539
	0.0275	0.0521	−0.6345	0.0983	−0.4703

[a] The order of the variables is the same as reported in Table 1.

classifies variables on the basis of the corresponding Fisher weights (or F-ratios) [13]. The value of the Fisher weight for each variable is the average of the values of all the possible pairs of categories into which the objects have been subdivided.

The elaboration of the 3-D matrix referring to the comparison between the two surveys was performed by the so-called "unfolding" method. Among the various techniques proposed for the treatment of 3-D matrix, the unfolding method is relatively simple and, moreover, it has proved to be suitable when techniques such as principal component analysis and its derivatives were used [14].

In short, given a 3-D matrix **X** with I rows, K columns and J cases (or ways), it can be sliced along one of the three dimensions and the slices can be rearranged into a 2-D matrix **X**. There are three possibilities for matrix **X**: $I \times (J \times K)$, $J \times (I \times K)$ and $K \times (I \times J)$. Matrix **X** (or matrix **X**$'$) can then be treated with usual multivariate techniques [14].

4. Discussion

4.1. 1990 Survey

The covariance matrix of the variables, as the raw data have been autoscaled, is equal to the correlation coefficient matrix. It gives the following main indications. The highest absolute values of the correlation coefficients have been found for the following groups of variables: Cl, CON ($r = 0.86$); Cl, COD ($r = 0.84$); Cl, Mn ($r = 0.80$); COD, SUR, Mn ($r = 0.80$ to 0.91); Cu, CuF ($r = 0.90$); DO, NH_3 ($r = -0.82$). The following variables appear to be poorly correlated with all the others ($r < 0.70$): Al, SM, Zn.

The following clusters of variables: Cu, CuF; Mn, SUR, COD; Cl, CON result also from dendrograms of the variables on the basis of correlations.

The four above cited algorithms of hierarchical clustering (see the previous section) give concordant classifications of the 44 sampling stations. The only exception regards the TA3 station, which is joined to TA2 by the complete linkage and to BL1 by Ward's method.

The dendrogram obtained by the complete linkage method is reported, as an example, in Fig. 2. (It must be noted that this method, compared with some others in a previous work [15], showed the greatest classification ability.)

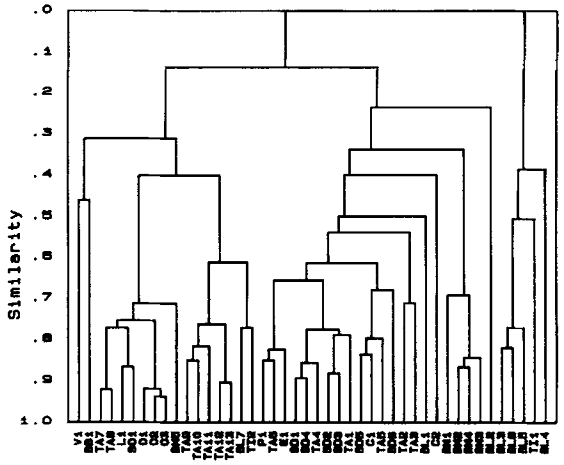

Fig. 2. Dendrogram by complete linkage (Euclidean distance) for the 44 sampling stations considered in the survey of 1990.

The river reaches (and the corresponding stations) belonging to the same cluster in the dendrogram are to be considered fairly uniform from the point of view of general pollution; on the contrary, the rivers subdivided in several different clusters, are to be considered quite variable as a pollution situation. The following clusters are evident in Fig. 2: the cluster from TA9 to TA13 (the Tanaro river shows a relatively steady pollution situation from Masio to Bassignana, on a distance of about 40 km); the cluster BM1–BM4 (Bormida di Millesimo from Saliceto to Cortemilia, about 40 km); the cluster SD1-SD4 (Stura di Demonte from Vinadio to Borgo San Dalmazzo, nearly 30 km) and the cluster O1–O3 (Orba from Ovada to Castellazzo Bormida, about 30 km).

The principal component analysis, and the subsequent evaluation of eigenvalues according to the "average variance" criterion, provides an identification of 4 significant physical factors, in addition to a possible, rather doubtful, fifth factor (in fact, its eigenvalue is very close to 1). The Indicator Function criterion [11] identifies four of them; therefore, four significant factors are considered. The values of the loadings for the Varimax rotation are reported in Table 3.

On the basis of Table 3, the four factors (F) can be associated with the following experimental vari-

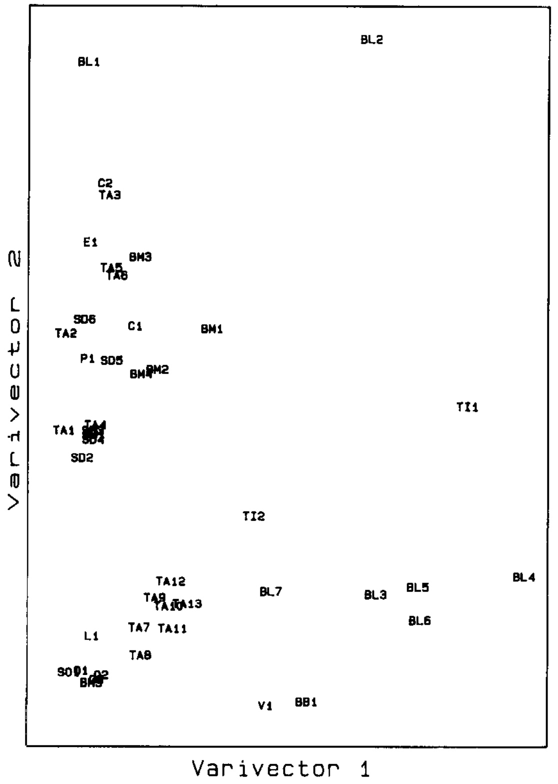

Fig. 3. Scores of the 44 sampling stations of the survey of 1990 on factors F1 and F2 rotated according to Varimax.

ables: F1 (variance of the rotated factor = 6.1): COD, Mn, SUR; F2 (variance = 3.5): Cu, CuF, Pb; F3 (variance = 3.0): Ni, NO_3, NO_2; F4 (variance = 2.2): CON, SO_4, Zn.

The scores of the 44 sampling stations are reported in Fig. 3 (on the F1 and F2 factors) and in Fig. 4 (F3 and F4 factors).

Each sampling station (or cluster of stations) shows, in these figures, a specific score (i.e., a specific coordinate value) on the axes of the factors. This indicates, for each station, the degree of pollution caused by the corresponding factor. Therefore it is possible, by means of a limited number of diagrams, to evaluate the general status of pollution of the rivers under investigation. Imperfect correspondence between clusters resulting from the dendrograms and those on the factors can be explained considering that all of the experimental variables were taken into account in the dendrograms, whereas each factor is related to some (and in general to a few) of these variables. It is also of interest to note, on the basis of the Figs. 2–4, that the clusters in the dendrogram are ordered from left to right following a sequence almost equal to that of increasing pollution. The cause of this fact can be defined as ‘‘indirect’’. More precisely, as observed in previous studies [4,16], the clusters of the more highly polluted samples are more scattered than the clusters of the less polluted ones. It should then be expected that a sequence of increasing Euclidean distances, such as that considered here in hierarchical dendrograms, is coincident with an increasing degree of pollution.

Selection of variables was then carried out. It allowed to sort the experimental variables according to their ability to discriminate the objects into categories, with the aim of considering only the variables showing the highest information content (see Section 3). In order to proceed to the selection it is necessary to subdivide the stations into categories. Two subdivisions of the stations into categories have been

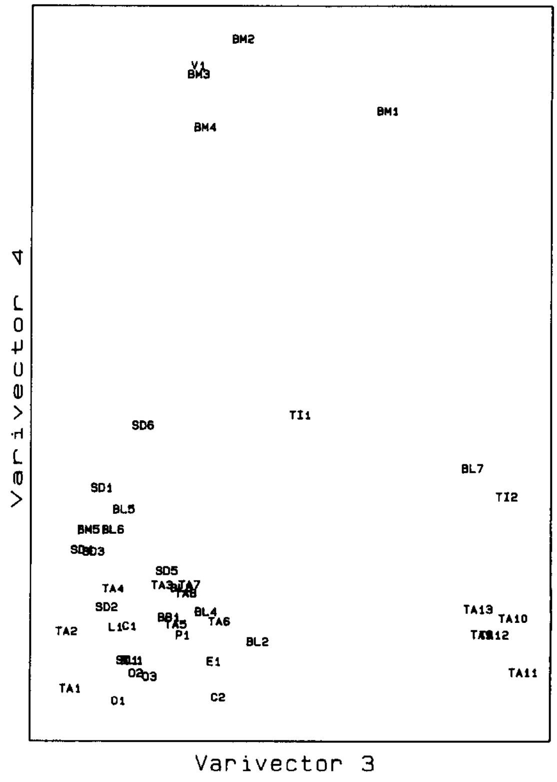

Fig. 4. Scores of the 44 sampling stations of the survey of 1990 on factors F3 and F4 rotated according to Varimax.

made in the present case, in the following way. In the first experiment the "complete linkage") dendrogram was considered and its branches were cut at a similarity level of 0.70. Considering only the groups of stations consisting of at least three objects, seven clusters were obtained, containing 31 stations in all. The second trial was carried out in an analogous way, but at a similarity level of 0.55. Five clusters containing 35 stations in all were obtained. The Fisher weights, arranged in decreasing order, gave the following sequences for the first 18 variables. Test 1: Zn, NO_3, Cl, COD, SO_4, CON, NH_3, Fe, DO, Al, SM, NO_2, Mn, CrT, Cu, pH. Test 2: SO_4, COD, Zn, Cl, CON, NO_3, Fe, CuF, DO, NO_2, SM, Mn, NH_3, Cu, CrT, pH. The above sequences appear to be strongly dependent on the selected conditions and so they are poorly reproducible.

4.2. Comparison between the surveys of 1990 and 1978

As already indicated, only 24 sampling stations and 12 experimental variables of the 1990 survey could be used for the comparison with the same stations and variables for 1978 (see Tables 1 and 2). It was therefore necessary, firstly, to verify if this reduced set gives a sufficiently exhaustive and accurate representation of the general situation in 1990, so as it results from the complete set of 44 stations and 20 variables. A factor analysis on the reduced and autoscaled matrix (24 × 12) was then performed. Three significant physical factors were identified by this treatment. From the loading values (Table 4) the factors can be associated with the following experimental variables: F1 (variance of the rotated factor = 4.9): COD, Mn, SUR; F2 (variance 2.6): Ni, NO_3, NO_2; F3 (variance 2.1): Cu, Pb. The inversion between F2 and F3 factors excepted, it is possible to conclude that the most important factors previously found on the complete matrix were also found on the reduced one (24 × 12). As a consequence, the latter can be used as a sufficiently representative matrix of the general situation in 1990.

Table 4
Loading of 12 experimental variables on three significant factors rotated according to the Varimax method (survey of 1990; 24 sampling stations considered) [a]

F1	−0.4398	−0.3173	−0.0175	0.3725	−0.4101
	−0.2823	0.0528	−0.1553	0.0957	0.1010
	0.2944	−0.4312			
F2	−0.1292	0.1663	−0.0459	−0.1113	−0.1240
	0.2644	0.5907	0.4151	0.5516	−0.0238
	−0.0775	−0.1525			
F3	0.0333	−0.1607	−0.6257	0.0770	0.1542
	0.0577	−0.0291	0.0037	0.1129	−0.6225
	0.3667	0.1167			

[a] The order of the variables is the same as reported in Table 1 (sign *).

Table 5
Loadings of 12 experimental variables on three significant factors rotated according to the Varimax method (survey of 1978)

F1	0.4022	0.2326	−0.0548	−0.3653	0.2591
	0.4009	−0.0457	0.3506	0.1883	0.0048
	−0.3625	0.3576			
F2	0.0651	0.4369	0.5874	0.0800	0.2459
	0.0103	0.5630	0.0238	−0.1742	−0.0106
	0.2113	−0.0197			
F3	0.0515	−0.0891	−0.0676	0.1979	−0.0615
	0.0413	0.1529	0.1080	0.3049	−0.8821
	0.1501	0.1069			

The comparison between the two surveys by factor analysis, was performed by means of two other analyses of this kind. The former was carried out on the analogous matrix (24 × 12) relating to the 1978 survey. The latter was performed on a (48 × 12) matrix; it resulted from an unfolding of the 3-D matrix (24 × 12 × 2), with a subsequent union at a constant number of columns (i.e., of variables), and allowed a direct comparison between the two surveys. The loading values obtained by the (24 × 12) matrix of 1978 (Table 5) suggested an association of each of the three significant factors with the following variables: F1 (variance 5.7): COD, DO, NH_3, NO_2, pH, SUR; F2 (variance 2.3): CON, Cu, Ni; F3 (variance 1.1): Pb.

An urban and industrial characterization can be attributed to F1 and F2, respectively, on the basis of the correlated experimental variables. The loadings obtained from the whole matrix for the two surveys (48 × 12) (see Table 6) suggested an association of each of the three identified significant factors with the following variables: F1 (variance 3.5): COD, NH_3, NO_2, pH, SUR; F2 (variance 1.6): Cu, Ni; F3 (variance 1.5): CON, DO, Mn.

It should be noted that in this case the Varimax

Table 6
Loadings of 12 experimental variables on three significant factors rotated according to the Varimax method, considering 24 sampling stations in the two surveys (1978 and 1990)

F1	0.4246	0.1040	−0.0125	−0.2931	0.0911
	0.4382	−0.0435	0.3867	0.1976	−0.0333
	−0.3947	0.4185			
F2	0.0190	0.3030	0.5765	0.1039	−0.0062
	0.0069	0.7066	0.0371	0.1838	0.0812
	0.1508	−0.0268			
F3	0.2125	0.5205	0.0547	−0.4139	0.6633
	0.1133	0.0354	0.0708	−0.1887	−0.0819
	−0.0901	0.0130			

rotation utilizing the "eigenvalue weighted" option (see the previous section) has been carried out. The series of variables associated with the three latter factors (i.e., the 1978 and 1990 surveys combined), compared with the analogous series for the two distinct surveys, leads to conclude that the general features resulting from the unified treatment are principally controlled by the situation in 1978. More precisely, the F1 factor resulting from the unified treatment agrees, for 5 out of 6 variables, with the F1 factor of 1978, and the F2 factor of the unified treatment agrees for 2 out of 3 variables with the F2 factor of 1978. This is a first, very general, indication of a greater degree of pollution existing during the first survey. More detailed indications about the changes during the period of 12 years can be derived from the examination of the corresponding station scores, for the two combined surveys, on the rotated factors (Fig. 5; in this figure, letters A and B indicate the surveys of 1978 and of 1990, respectively).

An examination of the scores of the sampling stations on the two factors considered in Fig. 5 shows a general improvement in the pollution status from 1978 to 1990. The only remarkable exception to this general improvement is represented by some stations on the Tanaro (TA10; TA11; TA13), which appear slightly more polluted in 1990 in comparison with 1978 (it must be noted that the stations TA11A and TA13A, referring to the 1978, are revealed at the left bottom of the graph, but they are not readily visible because of the overlapping with other symbols. Their position has been clarified by an enlargement of the graph). On the third factor no general and univocal changes between 1978 and 1990 can be observed.

On the same matrix combining the two surveys (dimensions: 48 × 12), cluster analysis using the four above described algorithms were also performed. The four corresponding dendrograms show essential agreement. In particular, the dendrogram according to complete linkage, reported as an example in Fig. 6, shows several quite narrow clusters.

It is important to note that, except few cases, the clusters are homogeneous from the temporal point of view; i.e., they consist only of samples taken in 1978 (A) or only of samples taken in 1990 (B). This means that the between-surveys distances are remarkably prevalent over the within-surveys distances. This fact, in other words, points out the remarkable quantitative difference between the pollution situation in 1978 and in 1990. A comparative study among sampling stations in 1978 and in 1990,

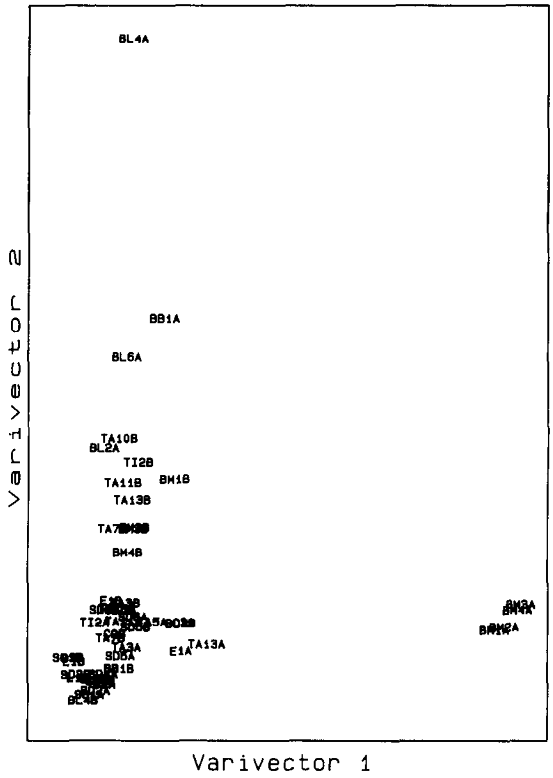

Fig. 5. Scores of 24 sampling stations in 1978 (A) and in 1990 (B) on the first two factors rotated according to Varimax.

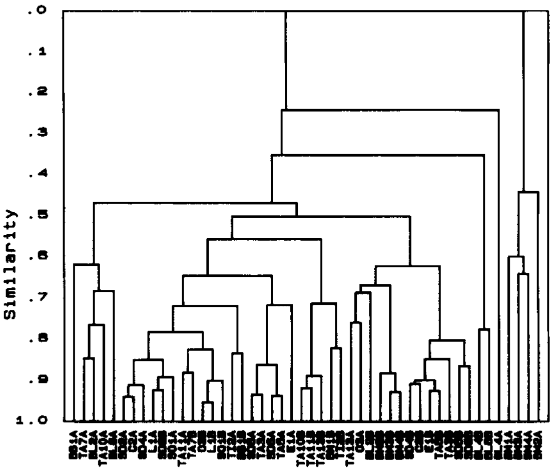

Fig. 6. Dendrogram by complete linkage (Euclidean distance) for 24 stations in 1978 (A) and for the same stations in 1990 (B).

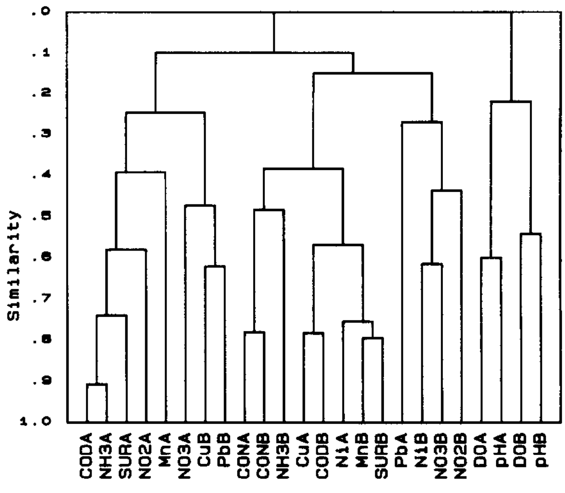

Fig. 8. Dendrogram by complete linkage, on the basis of correlation coefficients, for 24 variables, 12 of these referring to the survey of 1978 (A) and 12 to the survey of 1990 (B).

but from a qualitative (and not quantitative, as the previous case) point of view (i.e., on the basis of the mutual correlations) can be performed considering the corresponding transpose matrix (dimensions 12 × 48). The corresponding dendrogram according to complete linkage and based on correlation coefficients is reported in Fig. 7.

In this case, as well as in the previous one, it can be observed that the resulting clusters are homogeneous from the temporal point of view. The only exception is represented by the station L1, which appears qualitatively very similar in the two surveys (see the pair L1A and L1B in Fig. 7).

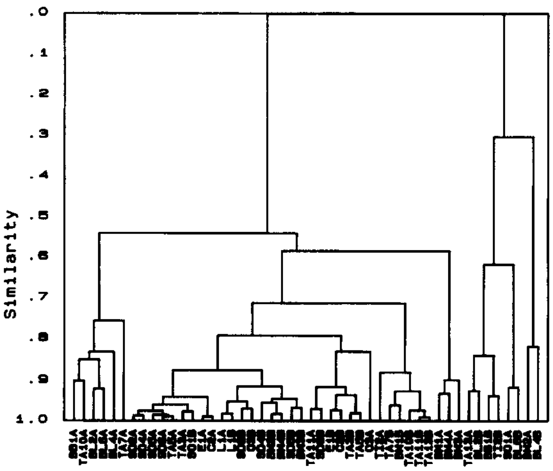

Fig. 7. Dendrogram by complete linkage on the basis of correlation coefficients for 24 stations in 1978 (A) and in 1990 (B).

The study of the correlations among the variables in the two surveys was performed on the basis of an unfolding of the original 3-D matrix (24 × 12 × 2) with a subsequent union at a constant number of rows (i.e., stations). A 2-D 24 × 24 matrix was obtained. Therefore, the obtained 24 column variables included the 12 variables for the 1978 survey (A) and the 12 variables for the 1990 survey (B). The corresponding dendrogram of the 24 variables, based on complete linkage and on correlation coefficients, is reported in Fig. 8.

Among all the variables, only the conductance (CON) provides a pair with high correlation between 1978 and 1990. In particular, from the examination of the correlation coefficient matrix, a correlation coefficient $r = 0.91$ is obtained for the pair CONA-CONB. The correlation coefficients for all of the other pairs result very low with a maximum value of just 0.4.

Taking into account that a correlation among conductances appears to be due to general causes and not to specific phenomena, the essential difference between the environmental situations in the rivers considered results evident, also from this point of view.

5. Conclusions

On the basis of the previous discussion, the following general conclusions can be deduced.

(1) The sequences of the variables according to their decreasing discriminant ability (and consequently to their usefulness) appear to be poorly reproducible, as they are strongly dependent on the conditions selected for the determination of their discriminant ability. It has been shown, in particular, that these sequences are determined by the different similarity levels used for the determination of the classes of objects. In a previous paper [4] it was demonstrated that the addition or the elimination of a few stations from the whole of sampling stations was sufficient for remarkable changes in the resulting sequences of variables. Therefore, it appears impossible, as a rule, to define some groups of a few suitable variables, which can be used for all the studies concerning the environmental systems of the same kind, in such a way as to make the experimental data collection faster and cheaper. The only general conclusion in this connection, derived from the present and previous studies [4,16] concerns the relative inefficiency of pH.

(2) Given a 3-D $(I \times J \times K)$ matrix (I objects, K variables and J cases or ways; in this case: two surveys referring to two different times), and considered the three matrices deriving from the unfolding, together with their corresponding transposes (6 possibilities in all, see Section 3), the following information, as a general rule, can be deduced.

(a) $(I \times J) \times K$ matrix: it is suitable to point out quantitative differences among the objects (in the present instance the different pollution degree among the sampling stations, both within a single survey and between the two surveys).

(b) $K \times (I \times J)$ matrix, i.e., the transpose of the previous one: useful to point out qualitative differences, that is different kinds of pollution, among the stations in the two surveys.

(c) $I \times (J \times K)$ matrix: it allows the study of the correlations (and therefore the analogies or the qualitative differences) among the variables for the two surveys. This kind of information assumes a geographical mean, because a variable correlated between the two surveys shows an analogous distribution of its values for the sequence of the stations in the two different periods.

(d) $(J \times K) \times I$ matrix, i.e., the transpose of the previous one: points out quantitative differences among the variables. However, this information appears to be redundant, since it already results from the Table of the mean values of the variables (see, for example, in this case, Table 1).

Finally, it must be noted that the two remaining matrices, namely $J \times (I \times K)$ and its transpose, appear to be less useful in the present study, as only two cases are considered ($J = 2$).

(3) The study of a 3-D matrix, in the present case, points out a remarkable difference in the pollution situation of the Tanaro basin from 1978 to 1990. This difference regards both quantitative and qualitative aspects as well as sampling stations and variables. More particularly, an appreciable decrease of the degree of pollution in 1990 compared to 1978 can be observed, both with regards to the factor characterized by COD, NH_3, NO_2, pH, SUR, of probable urban nature and for the factor characterized by Cu and Ni, of probable industrial nature. The more evident reduction of pollution concerns the Bormida (BM1–BM4), the Belbo and, to a certain amount, the Orba rivers.

This decrease of pollution is likely to be related to a network of depuration systems put into operation, at both industrial and urban levels, from the year 1978 to 1990 [6].

References

[1] J.I. Drever, The Geochemistry of Natural Waters, Prentice Hall, Englewood Cliffs, NJ, 1982.

[2] S. Geiss, J. Eimax and K. Danzer, Fresenius' Z. Anal. Chem., 333 (1989) 97.

[3] P.K. Hopke, J. Environ. Sci. Health, A11 (1976) 367.

[4] R. Aruga, G. Negro and G. Ostacoli, Ann. Chim. (Rome), 80 (1990) 34.

[5] R. Aruga, G. Negro and G. Ostacoli, Fresenius' J. Anal. Chem., 346 (1993) 968.

[6] Regione Piemonte (Assessorato all'Ambiente), III Censimento dei Corpi Idrici, Turin, 1992.

[7] Regione Piemonte (Assessorato all'Ambiente), I Censimento dei Corpi Idrici, Turin, 1980.

[8] APHA, AWWA, WPCF, Standard Methods for the Examination of Water and Wastewater, American Public Health Association, Washington, DC, 1975.
[9] CNR-IRSA, Metodi Analitici per le Acque, Consiglio Nazionale delle Ricerche, Rome, 1975.
[10] D.L. Massart and L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, Wiley, New York, 1983.
[11] M. Forina, R. Leardi, C. Armanino and S. Lanteri, PARVUS, Elsevier, Amsterdam, 1988.
[12] E.R. Malinowski, Factor Analysis in Chemistry, Wiley, New York, 1991.
[13] L. Kryger, Talanta, 28 (1981) 871.
[14] P. Geladi, Chemom. Intell. Lab. Systems, 7 (1989) 11.
[15] R. Aruga, P. Mirti and V. Zelano, Analusis, 18 (1990) 597.
[16] M. Baldi, C. Dacarro, B. Bonferoni and V. Riganti, Inquinamento, 27 (1985) 37.