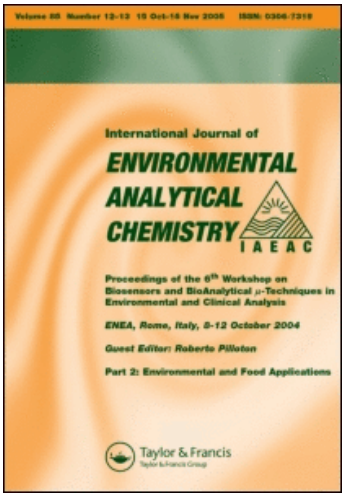

International Journal of Environmental Analytical Chemistry
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713640455

Monitoring of water quality in South Paris district by clustering and SIMCA classification

Michele De Luca[a]; Filomena Oliverio[a]; Domenica Ioele[a]; Gilles-Pascal Husson[b]; Gaetano Ragno[a]
[a] Department of Pharmaceutical Sciences, University of Calabria, Rende, Italy [b] Department of Public Health, Faculty of Pharmacy, University of Paris-Descartes, Paris, France

To cite this Article De Luca, Michele , Oliverio, Filomena , Ioele, Domenica , Husson, Gilles-Pascal and Ragno, Gaetano(2008) 'Monitoring of water quality in South Paris district by clustering and SIMCA classification', International Journal of Environmental Analytical Chemistry, 88: 15, 1087 — 1105
To link to this Article: DOI: 10.1080/03067310802428232
URL: http://dx.doi.org/10.1080/03067310802428232

Intern. J. Environ. Anal. Chem.
Vol. 88, No. 15, 20 December 2008, 1087–1105

Taylor & Francis
Taylor & Francis Group

Monitoring of water quality in South Paris district by clustering and SIMCA classification

Michele De Luca[a], Filomena Oliverio[a], Domenica Ioele[a], Gilles-Pascal Husson[b] and Gaetano Ragno[a]*

[a]*Department of Pharmaceutical Sciences, University of Calabria, Rende, Italy;*
[b]*Department of Public Health, Faculty of Pharmacy, University of Paris-Descartes, Paris, France*

(*Received 11 April 2008; final version received 22 August 2008*)

With the aim of obtaining a monitoring tool to assess the quality of water, a multivariate statistical procedure based on cluster analysis (CA) coupled with soft independent modelling class analogy (SIMCA) algorithm, providing an effective classification method, is proposed. The experimental data set, carried out throughout the year 2004, was composed of analytical parameters from 68 water sources in a vast southwest area of Paris. Nine variables carrying the most useful information were selected and investigated (nitrate, sulphate, chloride, turbidity, conductivity, hardness, alkalinity, coliforms and *Escherichia coli*). Principal component analysis provided considerable data reduction, gathering in the first two principal components the majority of information representing about 92.2% of the total variance. CA grouped samples belonging to different sites, distinctly correlating them with chemical variables, and a classification model was built by SIMCA. This model was optimised and validated and then applied to a new data matrix, consisting of the parameters measured during the year 2005 from the same objects, providing a fast and accurate classification of all the samples. The most of the examined sources appeared unchanged during the 2-year period, but five sources resulted distributed in different classes, due to statistical significant changes of some characteristic analytical parameters.

Keywords: water quality; multivariate analysis; PCA; clustering; SIMCA

1. Introduction

In the recent years, the interest in quality control of water for human use has increased considerably. An effective socioeconomic development of the communities depends much on the sustainability of the available water resources. Water of adequate quantity and quality is required to meet growing household, industrial and agricultural needs. It is largely influenced and determined by the natural processes and anthropogenic activities in the region. The excessive use of toxics in industry or the chemical fertilisers in agriculture can adversely affect the suitability of water resources.

The recent trends in environmental protection indicate that, in the immediate future, the treatment rules and the controls about water are tending to increase. Furthermore, the

*Corresponding author. Email: ragno@unical.it

ISSN 0306–7319 print/ISSN 1029–0397 online

DOI: 10.1080/03067310802428232
http://www.informaworld.com

ability to monitor parameters containing critical information on the analysis or treatment of water is very important for research and development.

One of the best approaches to this control is provided by use of multivariate analysis on the fundamental physical, chemical and biological properties of water. The main aim is to provide information useful for a fast and impartial water control management. The modern analytical techniques provide an enormous amount of data, which, if processed using descriptive univariate methods, is of little value for a decision-making process [1–2]. In contrast, multivariate procedures have been proven suitable for environmental quality assessment [3–6] and for pollution studies [7], offering great possibilities for management purposes in terms of aiding the decision makers [8]. The principal advantage of the multivariate techniques consists of the possibility to analyse a very high number of data from the investigated system, combining them to build a multivariate model. This model is then able to predict new unknown samples [9–12]. Grouping of data (objects) can be made by unsupervised methods which identify the natural clustering pattern and group objects on the basis of similarities between the samples. The most common methods of unsupervised pattern recognition are cluster analysis (CA) and classification methods, widely recognised as very powerful tools for getting better information about relations within data set. Application of these chemometric techniques offers a reliable and better understanding of the hydrochemistry and hydro-chemical processes in the study region.

In particular, classification methods are able to build a mathematic model to individualise the affiliation class of a new object, by using a limited number of independent variables (descriptors). The original data matrix is combined in a new matrix, named *distance matrix*. Classification of new objects provides to establish the better relationship between the variables describing the examined object and the class giving the best quantitative response. In the present article, a rapid and accurate methodological procedure based on CA and soft independent modelling class analogy (SIMCA) classification has been applied to the parameter data obtained from a high number of water sources in Paris and its neighbouring south region. Principal component analysis (PCA) provided to select the variables giving the optimal information [13–15]. SIMCA was used to perform the classification, providing to group the studied objects according to the similarity concept. The built model was tested to classify the same sources in the following year [16–19].

The objective of this study is the development of a methodology, based on advanced chemometric techniques, that makes it possible to assess and predict quality of the water to be used in a densely populated city. Multivariate techniques also permit identification of the possible analytical parameters or sources that influence the water systems and are responsible for the variations in water quality, which thus offers a valuable tool for developing appropriate strategies for effective management of the water resources. Since a continued and reliable monitoring of potable water is necessary to guarantee the health of people, the proposed method seems reliable to assess the water quality in the routine decision-making process.

2. Experimental

2.1 *Study area*

The study area is confined to a vast expand of land in the south of Paris, named Ile de France. Figure 1 shows the map of south Paris territory, expanded on a surface of

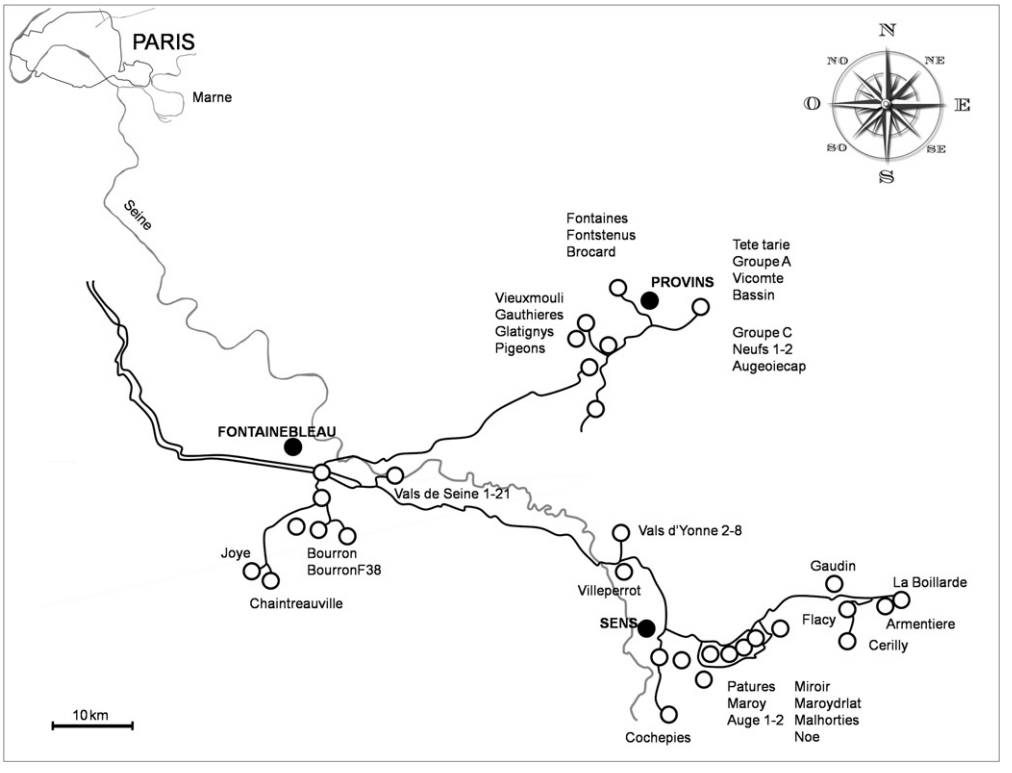

Figure 1. Study area showing monitoring network on the south region of Paris.

nearly 27.75 km^2, in which the collecting points of the studied samples are localised. Water arises from the basins of Vanne (Sens), Loing, Lunain and Voulzie (Provins). The basin of Voulzie is then divided into the three areas of Voulzie, Durteint and Dragon. The water has two origins: a 60% part derives from underground waters and the other part from superficial waters produced by the alluvial layers of rivers Seine and Yonne. The underground layers in which the water flows have a calcareous composition, showing a strong permeability due to a net of fissures. The flow speed of the water ranges from 100 $m\,h^{-1}$ in the region of Voulzie to 400 $m\,h^{-1}$ in the region of Vanne.

All the caught water is carried to Paris by means of an aqueduct system 600 km long and then carried to the depuration stations of Ivry, Orly on Seine and Joinville on Marna, where it is subjected to a slow filtration process, comparable to a natural soil filtration. After this treatment, the water is carried to five reservoirs from which it is distributed to the consumers.

2.2 *Sampling and methods of analysis*

The study was performed by using the data collected from 68 water sources. The investigation period was worked out along a 2-year period, 2004 and 2005. Analyses were performed every month for the year 2004 and mean data were used to build the

calibration set. A SIMCA classification model, defined by using the calibration set, was applied on a new data set (prediction set), measured on the same sources but carried out every 3 months during the year 2005.

All the samples were stored at 4°C until analysis. For each sample, 90 parameters including chemical, physical and microbiological ones, were carried out. Determination of pH, temperature, nitrite and ammonium was performed in situ. The other analyses were performed within 48 h after sampling. Analytical parameters were determined in triplicate with reference to official methods currently suggested [20].

2.3 *Chemometric software*

Application of the multivariate statistical algorithms has been supported by the software package *The Unscrambler 9.7*® (Camo Process As., Oslo, Norway). The software elaborates multivariate analysis and experimental design and is equipped with several methods, including CA, PCA, regression methods, SIMCA and PLS-DA. It also allows to optimise the calibration models and to develop validation procedures [21].

2.4 *Calibration set*

Water samples were collected during the year 2004 with a monthly frequency of sampling. Total of 68 water sources in the area of south Paris were examined. For each sample 90 analytical parameters (variables) were carried out, including chemical, physical and microbiological parameters. A wide set of data consisting of 73.440 values was collected (12 samples per station × 68 points × 90 parameters). The first selection of variables carrying the largest most useful information from the system was performed. This selection represents a critical step which should be carefully considered, because excluding important variables may lead to misleading results in building a classification model. The amount of relevant information does not necessarily increase when a higher number of variables is included. In contrast, it could increase random noise. The value of variance was adopted as a discriminating criterion to select the parameters, and those showing a value of relative standard deviation (RSD) under 10% were discarded. Anions seemed to play an important role in characterising the samples. In contrast, no cations showed significant variance. The selected parameters included six chemical parameters: conductivity (COND), alkalinity (AKM), chloride (Cl), sulphate (SO_4), nitrate (NO_3) and hardness (HRD); two microbiological parameters: *Escherichia coli* (ECOLI) and *Streptococcus faecalis* (STRF); one physical-chemical parameter: turbidity (TUR).

Most of the discarded parameters always showed a value under the detectable limit. Some of the selected parameters, as NO_3 or AKM can rise from the organic matter decay. External factors such as the temperature or the presence of other ions can play an important role in the decay process. Anthropogenic forces, as a massive use of toxics in industry or the chemical fertilisers in agriculture have the tendency to accelerate natural processes that affect water quality.

Table 1 lists the water sources used for the present investigation, the means of the selected parameters (12 measurements in 2004) and the SD values. The selected parameters were then normalised dividing them by SD for further elaboration.

Table 1. Calibration set.

No.	Sources	ECOLI UFC/100 mL	STRF UFC/100 mL	TUR NFU	COND $\mu S\,cm^{-1}$	AKM °F	Cl $mg\,L^{-1}$	SO_4 $mg\,L^{-1}$	NO_3 $mg\,L^{-1}$	HRD °F
1	ARMENTIERE	10.02 (0.58)	8.02 (0.47)	0.21 (0.01)	531.23 (30.95)	22.10 (0.90)	10.02 (0.56)	4.12 (0.18)	26.56 (1.12)	24.50 (1.43)
2	AUGE 1	1.03 (0.06)	4.00 (0.23)	0.20 (0.01)	726.56 (42.07)	24.70 (1.39)	31.64 (1.31)	21.23 (0.85)	57.12 (3.33)	34.65 (2.00)
3	AUGE 2	0.01 (0.00)	3.10 (0.18)	0.22 (0.01)	727.39 (43.28)	24.94 (1.41)	30.78 (1.72)	22.45 (0.98)	57.34 (3.35)	33.24 (1.98)
4	AUGEOIECAP	6.04 (0.25)	0.00 (0.00)	0.51 (0.02)	508.56 (21.21)	21.00 (0.89)	10.63 (0.57)	5.56 (0.21)	21.21 (0.86)	23.43 (0.98)
5	BASSIN	0.00 (0.04)	0.00 (0.00)	0.16 (0.00)	767.12 (34.11)	25.90 (1.11)	33.73 (1.85)	22.67 (1.30)	67.23 (3.77)	35.60 (1.58)
6	BIGNONS	0.02 (0.04)	0.00 (0.00)	0.10 (0.00)	521.46 (22.83)	22.33 (1.33)	19.54 (0.81)	20.77 (0.84)	28.43 (1.62)	26.42 (1.16)
7	BOUILLARDE	0.00 (0.00)	0.00 (0.00)	0.32 (0.02)	512.24 (28.89)	22.90 (1.37)	10.26 (0.60)	2.23 (0.12)	15.23 (0.85)	25.50 (1.44)
8	BOURRON	0.00 (0.00)	0.00 (0.00)	0.10 (0.01)	519.83 (29.65)	22.90 (0.97)	17.59 (0.96)	17.33 (1.01)	26.45 (1.04)	26.83 (1.53)
9	BOURRONF 38	0.00 (0.00)	0.00 (0.00)	0.24 (0.01)	523.48 (21.44)	22.65 (0.92)	17.43 (0.94)	17.45 (0.95)	26.56 (1.47)	27.05 (1.11)
10	BROCARD	0.02 (0.04)	0.00 (0.00)	0.10 (0.00)	822.45 (34.34)	28.25 (1.66)	38.73 (2.09)	29.56 (1.27)	63.67 (2.75)	37.60 (1.57)
11	CERILLY	1.00 (0.04)	0.00 (0.00)	0.33 (0.01)	535.47 (22.49)	22.97 (0.95)	5.94 (0.29)	5.67 (0.21)	21.50 (1.19)	24.90 (1.05)
12	CHAINTREAU	0.00 (0.00)	0.00 (0.00)	0.10 (0.01)	568.95 (32.47)	19.00 (1.08)	29.83 (1.24)	23.76 (0.95)	60.09 (2.59)	27.02 (1.54)
13	COCHEPIES	3.11 (0.13)	2.03 (0.09)	0.52 (0.02)	521.37 (22.50)	21.50 (0.96)	10.47 (0.45)	7.43 (0.40)	21.89 (1.19)	24.15 (1.04)
14	COIGNET	0.00 (0.00)	1.00 (0.04)	0.21 (0.01)	611.48 (24.82)	24.08 (1.05)	23.44 (1.00)	19.23 (0.76)	54.56 (3.04)	30.97 (1.26)
15	COIGNETFOR	0.00 (0.00)	0.00 (0.00)	0.10 (0.01)	610.48 (35.01)	24.20 (1.07)	23.57 (1.28)	20.33 (1.14)	50.67 (2.82)	30.84 (1.77)
16	DRAINFONT	1.01 (0.06)	1.01 (0.06)	0.25 (0.01)	824.74 (46.56)	28.12 (1.24)	35.49 (2.00)	28.44 (1.57)	60.83 (3.43)	37.98 (2.14)
17	FLACY	0.00 (0.00)	0.00 (0.00)	0.23 (0.01)	571.48 (24.60)	21.91(1.31)	11.03 (0.45)	7.76 (0.30)	31.87 (1.28)	26.73 (1.15)
18	FONTAINES	0.00 (0.00)	0.00 (0.00)	0.10 (0.00)	816.89 (33.36)	28.41 (1.24)	34.47 (1.38)	27.98 (1.58)	57.89 (3.39)	37.67 (1.54)
19	FONTSTENUS	0.00 (0.00)	1.00 (0.04)	0.11 (0.00)	738.00 (32.02)	24.72 (1.00)	30.16 (1.67)	26.09 (1.15)	53.09 (3.05)	35.30 (1.53)
20	GAUDIN	2.00 (0.09)	0.00 (0.00)	0.30 (0.01)	558.46 (23.79)	22.35(1.32)	11.83 (0.66)	5.40 (0.28)	26.07 (1.04)	24.50 (1.04)
21	GAUTHIERES	0.00 (0.00)	0.00 (0.00)	0.10 (0.01)	755.84 (44.17)	24.40 (0.98)	36.47 (2.01)	26.35 (1.07)	54.46 (2.20)	33.94 (1.98)
22	GLATIGNYS	1.03 (0.06)	0.00 (0.00)	0.10 (0.01)	832.24 (48.57)	26.30 (1.17)	41.82 (1.73)	28.48 (1.18)	71.23 (2.97)	38.71 (2.26)
23	GROUPEA	2.05 (0.08)	2.08 (0.08)	0.24 (0.01)	736.02 (30.52)	24.60 (1.07)	29.59 (1.63)	22.10 (0.90)	59.45 (3.28)	33.21 (1.38)
24	GROUPEC	2.03 (0.08)	3.05 (0.12)	0.20 (0.01)	735.00 (30.53)	24.90 (1.43)	30.59 (1.70)	21.23 (0.90)	59.51 (2.65)	34.36 (1.42)
25	JOYE	1.01 (0.04)	1.00 (0.04)	0.10 (0.00)	555.68 (22.79)	19.00 (1.07)	28.25 (1.63)	22.72 (1.30)	61.51 (2.72)	27.04 (1.11)
26	MALHORTIES	11.16 (0.61)	2.08 (0.11)	0.56 (0.03)	505.04 (27.80)	21.16 (0.86)	10.38 (0.45)	5.34 (0.28)	21.61 (0.89)	24.05 (1.32)
27	MAROY	0.00 (0.00)	0.00 (0.00)	0.10 (0.00)	579.03 (24.11)	22.20 (0.91)	14.65 (0.60)	6.53 (0.34)	36.23 (1.60)	26.56 (1.10)
28	MAROYDRLAT	0.00 (0.00)	0.00 (0.00)	0.10 (0.00)	578.60 (25.08)	23.10 (1.03)	14.97 (0.59)	8.46 (0.34)	33.48 (1.90)	27.67 (1.20)
29	MIROIR	15.07 (0.65)	2.00 (0.09)	0.42 (0.02)	497.60 (21.47)	21.30 (0.90)	9.36 (0.52)	5.67 (0.22)	18.65 (0.77)	24.19 (1.04)
30	NEUFS1	3.11 (0.13)	3.01 (0.13)	0.31 (0.01)	727.57 (31.36)	24.60 (1.03)	30.47 (1.71)	21.75 (1.23)	57.90 (3.17)	35.10 (1.51)
31	NEUFS2	5.10 (0.28)	4.02 (0.22)	0.30 (0.02)	736.29 (41.29)	24.50 (1.01)	30.25 (1.77)	21.12 (0.94)	57.78 (2.30)	34.90 (1.96)

(*Continued*)

Table 1. Continued.

No.	Sources	ECOLI UFC/100 mL	STRF UFC/100 mL	TUR NFU	COND $\mu S\,cm^{-1}$	AKM °F	Cl $mg\,L^{-1}$	SO_4 $mg\,L^{-1}$	NO_3 $mg\,L^{-1}$	HRD °F
32	NOE	14.08 (0.81)	1.00 (0.06)	0.41 (0.02)	527.78 (30.57)	22.10 (0.92)	10.46 (0.41)	5.14 (0.29)	22.23 (0.96)	24.97 (1.44)
33	PATURES	0.00 (0.00)	0.00 (0.00)	0.11 (0.00)	587.57 (24.83)	22.67 (1.01)	14.57 (0.59)	7.41 (0.40)	41.45 (1.81)	28.53 (1.21)
34	PIGEONS	0.00 (0.00)	0.00 (0.00)	0.10 (0.01)	822.00 (46.06)	26.60 (1.12)	39.93 (2.15)	28.23 (1.57)	72.51 (2.90)	37.32 (2.09)
35	RIVIEREFOR	3.07 (0.13)	1.03 (0.04)	0.20 (0.01)	726.04 (31.11)	24.90 (1.44)	30.03 (1.30)	21.32 (1.24)	66.31 (3.93)	34.61 (1.48)
36	SEL	0.00 (0.00)	0.00 (0.00)	0.10 (0.00)	525.79 (21.35)	22.30 (0.99)	18.02 (0.74)	17.78 (1.00)	25.41 (1.02)	27.04 (1.10)
37	STMARCOUF	3.00 (0.17)	1.00 (0.01)	0.23 (0.01)	548.09 (31.75)	22.07 (1.28)	11.26 (0.64)	5.96 (0.29)	28.51 (1.68)	25.86 (1.49)
38	STPHILDRA	0.00 (0.00)	1.02 (0.04)	0.20 (0.01)	591.03 (26.36)	23.57 (0.99)	13.48 (0.72)	8.50 (0.47)	35.67 (1.95)	28.54 (1.27)
39	STPHILSOU	5.04 (0.28)	4.22 (0.22)	0.20 (0.01)	549.07 (30.47)	22.35 (1.25)	11.58 (0.64)	6.09 (0.26)	28.45 (1.23)	25.33 (1.40)
40	STTHOMAS	1.05 (0.06)	1.00 (0.06)	0.23 (0.01)	607.04 (34.32)	23.80 (1.32)	23.36 (1.01)	20.89 (0.82)	55.90 (2.27)	30.82 (1.74)
41	STTHOMASFO	0.00 (0.00)	0.00 (0.00)	0.10 (0.01)	612.94 (36.29)	24.70 (1.39)	23.03 (1.35)	20.79 (0.87)	51.05 (2.88)	31.73 (1.88)
42	TETE	3.00 (0.18)	4.04 (0.23)	0.22 (0.01)	741.04 (43.49)	25.14 (1.40)	31.93 (1.32)	22.56 (1.27)	59.06 (2.40)	34.15 (2.00)
43	VICOMTE	7.26 (0.29)	2.00 (0.08)	0.20 (0.01)	736.07 (30.72)	24.80 (1.00)	30.86 (1.32)	24.24 (1.03)	59.34 (3.28)	34.17 (1.42)
44	VIEUXMOULI	0.02 (0.06)	2.00 (0.11)	0.10 (0.01)	767.83 (43.47)	25.60 (1.43)	35.38 (1.47)	29.57 (1.29)	55.54 (2.42)	35.32 (2.00)
45	VILLEPERRO	0.00 (0.00)	0.00 (0.00)	0.10 (0.01)	550.36 (31.44)	21.84 (1.26)	12.57 (0.68)	10.82 (0.43)	31.23 (1.36)	25.44 (1.45)
46	VALS DE SEINE 1	0.00 (0.00)	0.00 (0.00)	0.11 (0.01)	539.68 (31.31)	20.33(1.19)	20.46 (0.83)	38.48 (1.54)	9.35 (0.50)	26.45 (1.53)
47	VALS DE SEINE 10	0.02 (0.06)	0.00 (0.00)	0.10 (0.01)	580.38 (34.56)	19.63 (1.17)	31.37 (1.25)	34.34 (1.37)	7.68 (0.39)	24.42 (1.45)
48	VALS DE SEINE 12	0.03 (0.04)	0.00 (0.00)	0.09 (0.00)	622.85 (25.60)	20.52 (1.17)	27.97 (1.19)	26.56 (1.55)	37.13 (1.61)	26.78 (1.10)
49	VALS DE SEINE 13	0.00 (0.00)	0.01 (0.06)	0.10 (0.01)	694.25 (38.83)	23.60 (1.06)	26.81 (1.13)	23.78 (1.32)	46.24 (2.62)	30.94 (1.73)

50	VALS DE SEINE 14	0.00 (0.00)	0.01 (0.06)	0.10 (0.01)	699.00 (39.16)	24.00 (1.35)	26.82 (1.49)	23.90 (1.28)	48.35 (2.78)	31.00 (1.74)
51	VALS DE SEINE 15	0.01 (0.02)	0.00 (0.00)	0.09 (0.00)	580.04 (23.96)	20.50 (0.84)	21.63 (0.88)	26.09 (1.55)	23.47 (1.36)	25.20 (1.04)
52	VALS DE SEINE 16	0.00 (0.00)	0.02 (0.03)	0.33 (0.01)	574.05 (23.31)	21.00 (0.85)	20.45 (0.85)	33.87 (1.94)	14.80 (0.62)	26.26 (1.06)
53	VALS DE SEINE 17	0.00 (0.00)	0.02 (0.04)	0.10 (0.00)	598.50 (25.11)	21.30 (1.23)	20.13 (1.19)	40.65 (2.39)	10.31 (0.42)	25.94 (1.09)
54	VALS DE SEINE 18	0.01 (0.03)	0.01 (0.04)	0.40 (0.02)	547.70 (22.83)	20.68 (1.18)	18.14 (1.04)	29.45 (1.27)	9.52 (0.50)	24.36 (1.01)
55	VALS DE SEINE 19	0.00 (0.00)	0.00 (0.00)	0.22 (0.01)	532.46 (29.47)	19.70 (1.12)	18.25 (0.79)	29.45 (1.16)	9.64 (0.51)	23.43 (1.30)
56	VALS DE SEINE 20	0.00 (0.00)	0.00 (0.00)	0.40 (0.02)	634.08 (28.32)	20.67 (1.14)	30.53 (1.25)	22.32 (0.96)	33.85 (1.45)	26.20 (1.17)
57	VALS DE SEINE 21	0.00 (0.00)	0.00 (0.00)	0.20 (0.01)	570.46 (32.90)	20.20 (1.15)	19.46 (0.79)	34.13 (1.88)	20.47 (1.14)	25.60 (1.44)
58	VALS DE SEINE 3	0.00 (0.00)	0.00 (0.00)	0.10 (0.00)	565.21 (22.60)	20.30 (1.13)	22.75 (0.94)	26.14 (1.10)	18.83 (1.02)	26.07 (1.04)
59	VALS DE SEINE 4	0.00 (0.00)	0.00 (0.00)	0.11 (0.00)	505.23 (20.50)	20.30 (0.87)	17.48 (1.01)	20.54 (0.82)	6.67 (0.25)	23.83 (0.97)
60	VALS DE SEINE 7	0.00 (0.00)	0.00 (0.00)	0.20 (0.01)	534.03 (29.96)	21.24 (0.93)	15.94 (0.83)	10.67 (0.44)	12.37 (0.72)	24.51 (1.37)
61	VALS DE SEINE 8	0.02 (0.06)	0.03 (0.06)	0.10 (0.01)	556.72 (31.46)	21.53 (0.88)	17.25 (0.94)	21.90 (1.17)	15.23 (0.66)	25.60 (1.45)
62	VALS D'YONNE 2	0.00 (0.00)	0.00 (0.00)	0.10 (0.00)	498.06 (20.15)	19.26 (1.08)	12.23 (0.53)	20.43 (0.87)	17.45 (1.00)	21.85 (0.88)
63	VALS D'YONNE 3	0.01 (0.04)	0.00 (0.00)	0.10 (0.00)	537.00 (24.08)	20.88 (0.91)	15.45 (0.87)	18.12 (0.73)	25.69 (1.04)	25.10 (1.13)
64	VALS D'YONNE 4	0.00 (0.00)	0.01 (0.02)	0.10 (0.00)	587.09 (25.30)	22.06 (1.30)	18.67 (1.07)	20.34 (0.81)	27.63 (1.61)	26.22 (1.13)
65	VALS D'YONNE 5	0.00 (0.00)	0.00 (0.00)	0.16 (0.01)	540.62 (31.09)	20.85 (0.90)	15.87 (0.83)	16.21 (0.68)	25.21 (1.42)	25.22 (1.45)
66	VALS D'YONNE 6	0.00 (0.00)	0.00 (0.00)	0.10 (0.01)	538.04 (32.14)	19.77 (1.09)	16.65 (0.70)	17.54 (0.70)	30.23 (1.31)	23.51 (1.40)
67	VALS D'YONNE 7	0.00 (0.00)	0.00 (0.00)	0.10 (0.00)	617.05 (26.98)	24.01 (1.00)	18.34 (1.01)	17.65 (0.76)	31.11 (1.74)	28.26 (1.23)
68	VALS D'YONNE 8	0.00 (0.00)	0.00 (0.00)	0.10 (0.00)	650.53 (36.35)	24.59 (1.05)	19.12 (1.05)	16.34 (0.67)	40.44 (1.68)	29.64 (1.66)

Means and SD ($n = 12$) of the analytical parameters (U.I.) carried out in the year 2004: *Escherichia coli* (ECOLI), *Streptococcus faecalis* (STRF), turbidity (TUR), conductivity (COND), alkalinity (AKM), chloride (Cl), sulphate (SO_4), nitrate (NO_3) and hardness (HRD).

3. Multivariate analysis

3.1 *Cluster analysis*

Cluster analysis represents a series of multivariate methods which provide means for classifying a given population into groups (clusters), based on similarity or closeness measures. The objective principle of the distance is adopted for this aim. The agglomerative hierarchical clustering is nowadays the most cited method in literature [22], providing intuitive similarity relationships between any sample and the entire data set. The sample grouping is illustrated by a dendrogram that allows a global vision of the similarity between the objects. In this work, the hierarchic agglomerative cluster algorithm applied was the Weighted Average Linkage whereas the distance elaboration was performed using squared Euclidean distance as a measure of similarity.

$$\text{Weighted average linkage } s_{kf} = 0.5(s_{ks} + s_{kt})$$

where s_{kf} is the similarity of a new cluster, s_{ks} and s_{kt} represent the similarity of the starting cluster building the new cluster [23–25].

3.2 *Principal component analysis*

When an analytical system presents a high number of variables, the application of the chemometric techniques helps to have a global vision of the system, in such a way to appreciate the analytical weight of each variable and to single out possible relationships between the variables.

Principal component analysis is one of the most important data reduction method for a multivariate data set characterised by measurements on multiple variables. It reduces the significant dimensionality of a data matrix, allowing to retain most of the original information content. The original variables (X) are transformed by linear combination in a limited number of new variables, called principal components (PCs) [26]:

$$X = t_1 p_1' + t_2 p_2' + \cdots + t_A p_A' + E$$

where t are score values, p are loading values and E is the residual matrix.

The main aim of PCA is to explain as much as possible the total sum of square of the data matrix with a minimal number of PCs. The scores and loadings will be used to define the object classes. The number of PCs to be used is very important to increase robustness of the multivariate model. Including more PCs not necessarily increases the amount of relevant information and it could, on the contrary, increase noise [27].

In this work, PCA has been used to study the features of the defined classes and to enable the SIMCA classification.

3.3 *Soft independent modelling of class analogy*

Soft independent modelling class analogy is a chemometric technique in which new objects are classified with respect to their analogy with objects belonging to a class defined by PCA [19]. SIMCA has been reported to produce very high correct classification rate in the separation of very similar materials [28].

Each class is defined through a critical distance s_{crit}, which is a function of the distance of all the objects (training set) from the PCA model (calculated as SD of residuals s_0) and from *F*-test, measured in a confidence range of 95–99% [29].

$$s_{crit} = (Fs_0^2)^{1/2}$$

New samples are handled separately by each class model and its belonging is made on the basis of the distance to the class. If the distance of a new object is below the critical distance ($s_k < s_{crit}$) it will be considered part of that class. On the contrary, if this distance results to be higher, the object must be considered as an outlier.

4. Results and discussion

With the aim to define the calibration set, a selection process of the variables carrying good information was performed. The variables, carried out during the year 2004, showing a value of variance (expressed as RSD%) above 10% were selected, as above described. The mean values of the selected parameters are summarised in Table 1. The analytical parameters describing the water samples were chemical, chemical-physical or microbiological ones, presenting different units of measurements. A normalisation process then appeared necessary to obtain comparable data. The normalisation was applied to the analytical raw data, dividing the variables by SD [30–31], afterwards the data were centred and scaled before they were used in CA and PCA.

4.1 *Cluster analysis*

Cluster analysis was performed by using the data of the calibration set and then applying the weighted average linkage procedure. Results were reported in the form of dendrogram, depicted in Figure 2. On the basis of the connecting distances and in agreement with the PCA results, three distinctive clusters were defined.

Table 2 summarises for each cluster the mean values of the analytical parameters and the corresponding values of SD. The graphic in Figure 3 shows the percentage of the parameter means within a single cluster with respect to the mean values of all the clusters. The chemical parameters NO_3, SO_4 and Cl showed a high variation, providing very useful information to clustering. The microbiological parameters, ECOLI and STRF, also showed to be responsible for the cluster distribution.

The first group contained seven objects and was assorted with samples collected in a geographical area equipped with water softener systems. These sources came from the southeast region of the studied area, close to the town of Sens. The samples in this cluster were characterised by a relatively high bacterial content and a very low concentration of SO_4. The second cluster was formed by 18 samples, all localised in the territory around Provins, in the northern zone of the studied area. The most characterising parameters were found to be Cl and NO_3 which were almost twice the mean values. The last cluster was formed by the highest number of objects, 43, all characterised by very low levels of ECOLI and STRF. These sources were localised in a wide area including the regions of Fontainebleau and Sens.

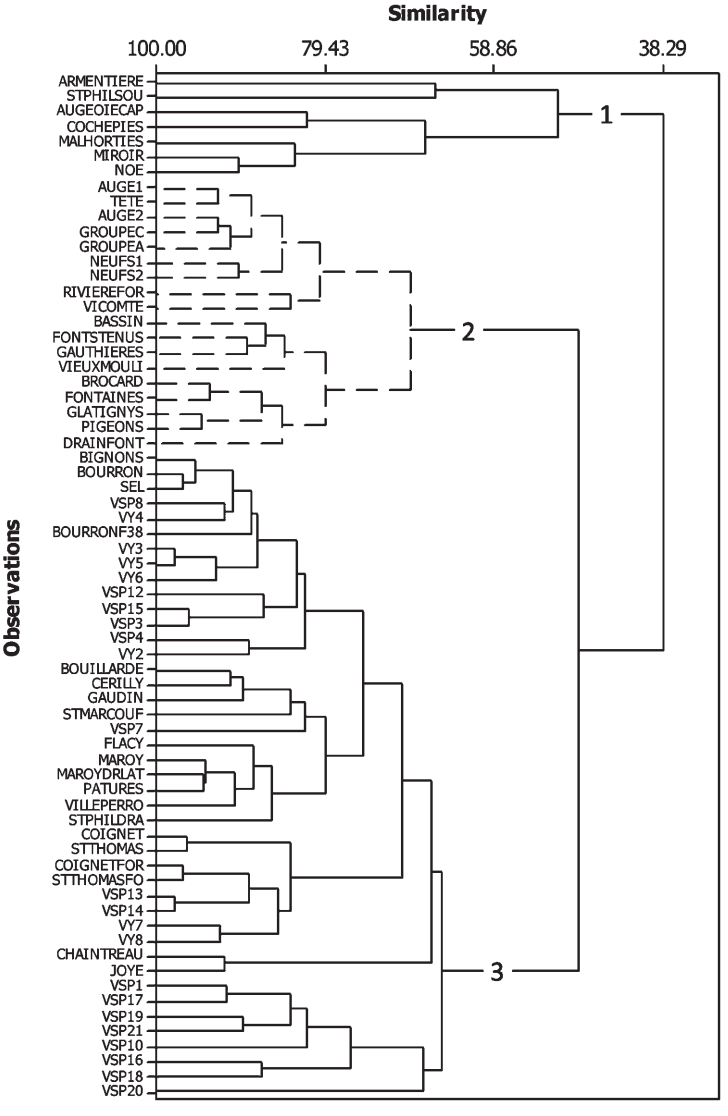

Figure 2. Dendrogram of the water sources by cluster analysis based on average linkage and euclidean distance.

Table 2. Statistical parameters of the clusters.

Cluster		ECOLI	STRF	TUR	COND	AKM	Cl	SO_4	NO_3	HRD
1	Mean	9.1	2.7	0.4	519.7	21.6	10.0	5.3	22.4	24.3
	SD	4.6	2.6	0.1	17.8	0.5	0.6	1.0	3.4	0.6
2	Mean	1.6	1.7	0.2	762.9	25.6	32.9	24.3	60.1	35.4
	SD	2.0	1.5	0.1	40.4	1.4	3.7	3.2	5.5	1.7
3	Mean	0.2	0.1	0.2	571.1	21.8	18.5	19.0	28.9	26.6
	SD	0.6	0.3	0.1	45.8	1.6	5.8	9.3	14.9	2.3

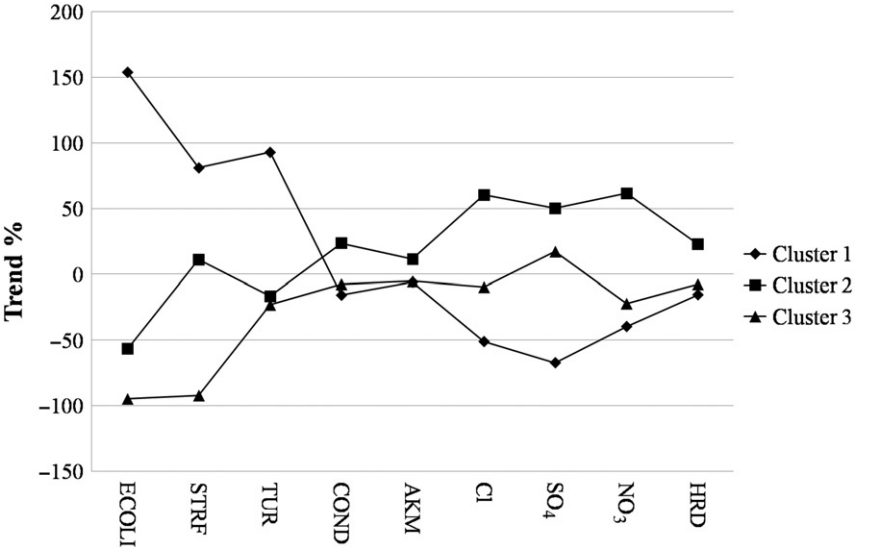

Figure 3. Variation of the selected parameters among the defined clusters.

4.2 *PCA-modelling*

Modelling procedure in SIMCA classification is articulated in a two-step process. The first step is based on a disjoint PCA-modelling, called *training stage*, where a model for each data class is built. In the second step, named *classification stage*, new objects are assigned to the defined class models.

In the training stage, PCA provided to extract and visualise the main information from the calibration data set in such a way as to examine qualitative differences between the three defined clusters. Figure 4 shows the PC1/PC2 score plot, especially useful because these two PCs summarised more variation in the data than any other pair of PCs. The first two components accounted for 92.2% of the variance of the data. The three classes resulted in perfectly separated, with positive scores on the second PC for the first class of samples and negative scores for the second one; in contrast, the third class presented positive values for the first PC and negative values for the second PC.

Bi-plot of the score and loading values (Figure 5) shows the most influential variables for each class. The third class resulted to have values close to the average for every variable. On the contrary, the parameters ECOLI, STRF, TUR in the first group or AKM, COND, HRD, NO_3, Cl in the second group were characterised by higher values.

4.3 *SIMCA classification*

Since in SIMCA technique each class is described by a PCA-model, three independent models were built. The obtained models were validated with full-cross validation approach. An optimal number of PCs was chosen for each model because the classes exhibited different shapes and structures (Table 3). All the objects were assigned to a single class according to a critical distance from the model at a given level of significance (5%).

In the classification stage, SIMCA was applied to the prediction data set, consisting of the measures collected in the year 2005 from the analysis of the same sources in the

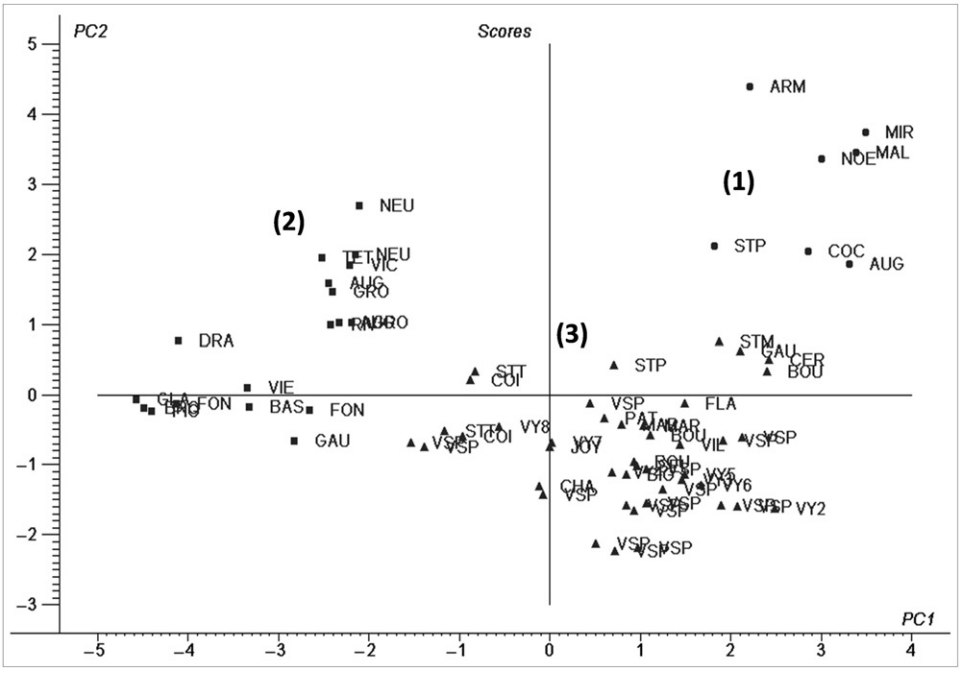

Figure 4. Score plot PC1 *vs.* PC2.

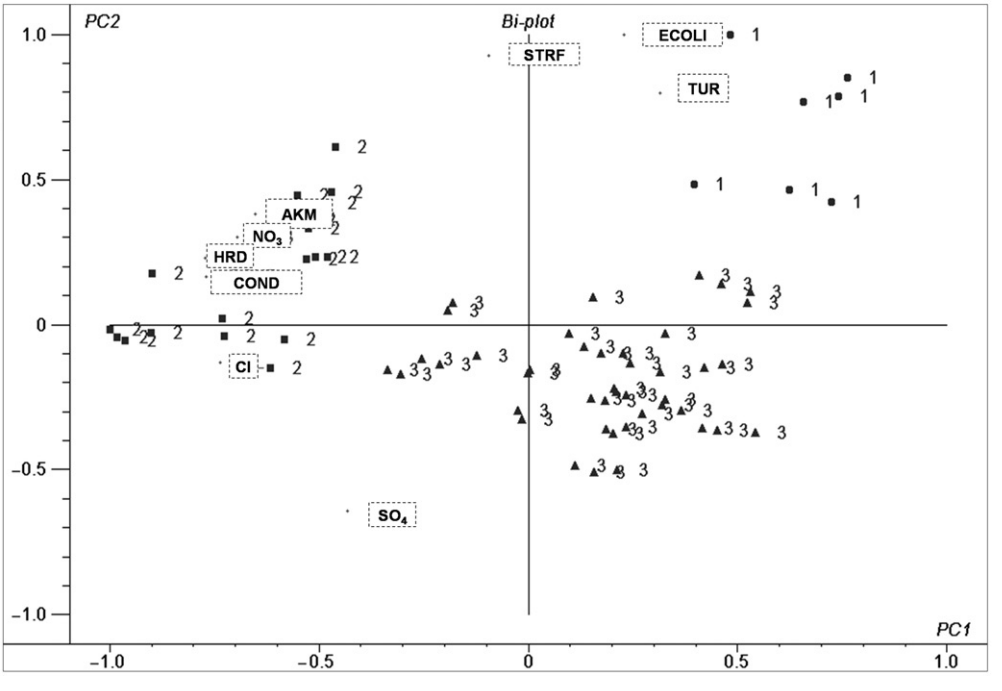

Figure 5. Bi-plot PC1 *vs.* PC2.

Table 3. Loadings of descriptors and variance for the class models.

PCA model	PCs	ECOLI	STRF	TUR	COND	AKM	Cl	SO_4	NO_3	HRD	Explained variance (%)	Cumulative variance (%)
1	1	−0.125	0.294	−0.379	0.419	0.398	0.333	0.014	0.416	0.368	83.13	83.13
	2	−0.578	−0.300	0.268	0.137	−0.120	0.340	0.589	0.040	−0.083	14.80	97.92
2	1	−0.251	−0.343	−0.320	0.393	0.341	0.375	0.359	0.216	0.360	65.41	65.41
	2	0.558	0.123	0.444	0.188	0.164	0.074	−0.165	0.538	0.303	13.29	78.71
	3	−0.106	−0.320	−0.300	−0.184	−0.421	0.090	−0.292	0.658	−0.243	8.80	87.51
	4	0.709	−0.304	−0.213	−0.035	−0.294	0.129	0.447	−0.196	−0.130	5.11	92.62
	5	−0.147	0.592	0.079	0.051	−0.451	0.597	0.236	0.054	0.014	3.70	96.32
3	1	0.009	0.224	−0.120	0.456	0.380	0.261	−0.037	0.476	0.536	35.56	35.56
	2	0.448	0.294	0.247	−0.170	0.270	−0.481	−0.559	0.076	0.005	25.33	60.89
	3	0.470	0.554	0.263	−0.015	−0.405	0.355	0.300	0.097	−0.118	14.81	75.70
	4	−0.069	−0.215	0.841	0.309	0.228	0.049	0.143	−0.257	0.068	10.38	86.08
	5	0.633	−0.203	−0.353	0.302	0.245	−0.105	0.303	−0.423	0.012	5.10	91.18
	6	−0.320	0.570	−0.059	−0.286	0.344	−0.175	0.369	−0.394	0.215	4.91	96.09

Table 4. SIMCA classification of the studied objects in the years 2004 and 2005.

Sources	Class 2004	Class 2005	Sources	Class 2004	Class 2005
ARMENTIERE	**1**	**0**	RIVIEREFOR	2	2
AUGE 1	2	2	SEL	3	3
AUGE 2	2	2	STMARCOUF	3	3
AUGEOIECAP	**1**	**3**	STPHILDRA	3	3
BASSIN	2	2	**STPHILSOU**	**1**	**3**
BIGNONS	3	3	STTHOMAS	3	3
BOUILLARDE	**3**	**1**	STTHOMASFO	3	3
BOURRON	3	3	TETE	2	2
BOURRONF 38	3	3	VICOMTE	2	2
BROCARD	2	2	VIEUXMOULI	2	2
CERILLY	3	3	VILLEPERRO	3	3
CHAINTREAU	3	3	VALS DE SEIN 1	3	3
COCHEPIES	1	1	VALS DE SEIN 10	3	3
COIGNET	3	3	VALS DE SEIN 12	3	3
COIGNETFOR	3	3	VALS DE SEIN 13	3	3
DRAINFONT	2	2	VALS DE SEIN 14	3	3
FLACY	3	3	VALS DE SEIN 15	3	3
FONTAINES	2	2	VALS DE SEIN 16	3	3
FONTSTENUS	2	2	VALS DE SEIN 17	3	3
GAUDIN	3	3	VALS DE SEIN 18	3	3
GAUTHIERES	2	2	VALS DE SEIN 19	3	3
GLATIGNYS	2	2	VALS DE SEIN 20	3	3
GROUPEA	2	2	VALS DE SEIN 21	3	3
GROUPEC	2	2	VALS DE SEIN 3	3	3
JOYE	3	3	VALS DE SEIN 4	3	3
MALHORTIES	1	1	VALS DE SEIN 7	3	3
MAROY	3	3	VALS DE SEIN 8	3	3
MAROYDRLAT	3	3	VALS D'YONNE 2	3	3
MIROIR	**1**	**0**	VALS D'YONNE 3	3	3
NEUFS1	2	2	VALS D'YONNE 4	3	3
NEUFS2	2	2	VALS D'YONNE 5	3	3
NOE	1	1	VALS D'YONNE 6	3	3
PATURES	3	3	VALS D'YONNE 7	3	3
PIGEONS	2	2	VALS D'YONNE 8	3	3

Sources changing class are reported in bold; sources excluded from the defined classes (outliers) are assigned to class 0.

calibration set. Whatever variation in the water characteristics could be so easily highlighted [32]. The same analytical variables selected for the calibration process were herein used.

The results of SIMCA classification are summarised in Table 4 and displayed graphically in Figures 6 and 7. In particular, Figure 6 shows the plots 'Distance *versus* Leverage' for all the classes, where the axes represent the distance of each sample from a specific class and Leverage is the distance of a single observation from the model centre. Leverage is very useful to identify outlying observations. If a sample presents a distance to the centroid greater than the critical distance, it is considered as an outlier and, as a consequence, rejected from the corresponding class model.

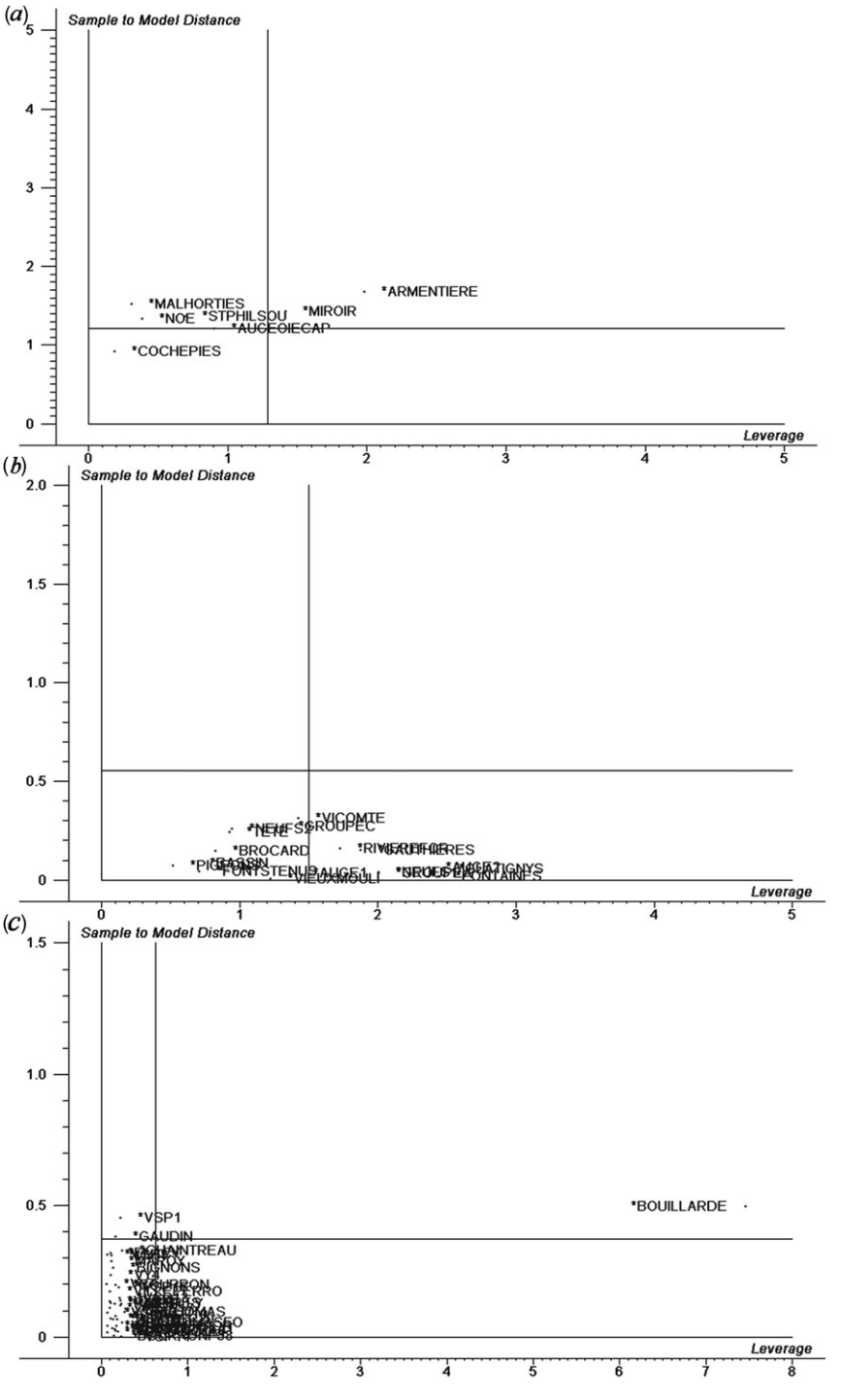

Figure 6. Distance *vs*. Leverage plots for class 1(a), class 2(b) and class 3(c).

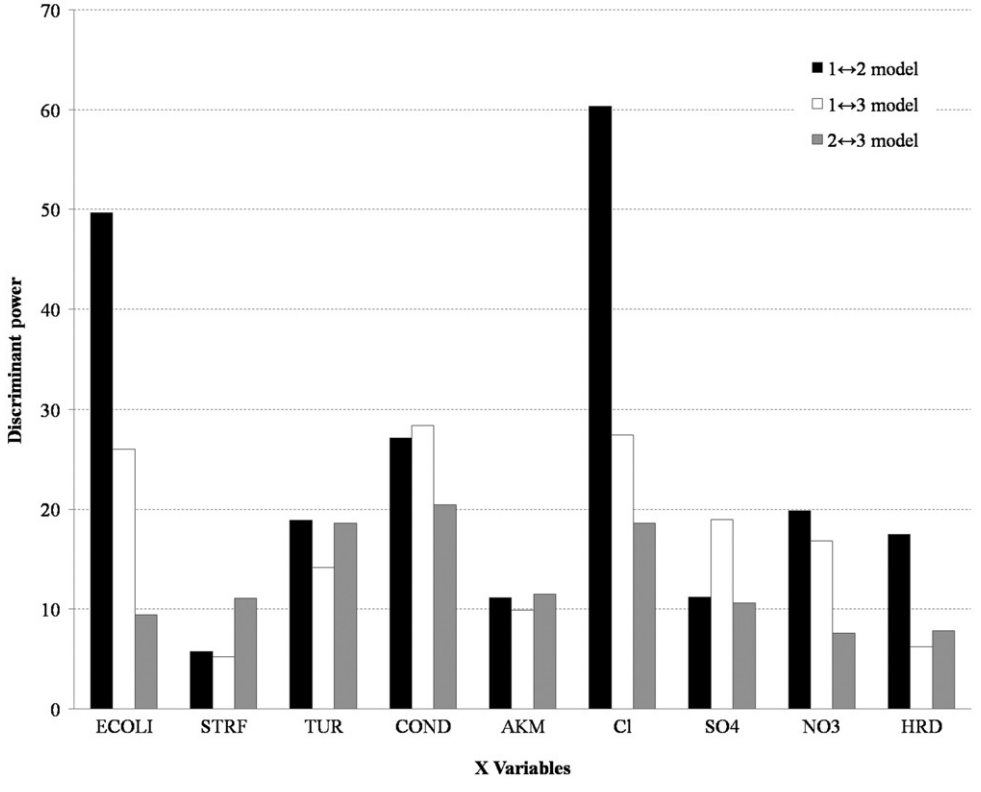

Figure 7. Discriminant power of the SIMCA model with respect to the studied variables.

The graph of Figure 7 displays for each class the discriminant power of the SIMCA model, showing the main responsible variables for the object distribution in the different classes. The parameters Cl, ECOLI and COND resulted clearly to be key parameters in allocating the water samples within the classes. The same parameters had resulted already in the previous CA, demonstrating they carry very useful analytical information.

Figure 8 shows the class distribution in the studied area. The first group comprised a good number of surface water sources with a collecting system devoid of water softeners [33]. This could explain the high presence of bacteria, almost 10 times higher than the samples within the other two classes. The second class composed by a series of sources scattered along a large region under cultivation. This class was characterised by considerable values of inorganic analytes. In particular, the high content of nitrate and chloride measured during the studied period, could be right justified by the use of nitrogenous fertilisers and chlorine-pesticides, which are the most common sources of nitrate and chloride ions in groundwater systems. The third class lists a large series of sources in a wide region in the neighbourhood area of Fontainebleau and along the Seine river. This area consists of heterogeneous areas, as it presents both urban and industrial areas encircled by woodlands and meadows. The collecting water system is supplied with water softeners [33] that explains the very low levels of ECOLI and STRF. This study demonstrated that the most of the tested water sources had kept the same analytical profile during the 2-year period investigated, without any shift through the classes.

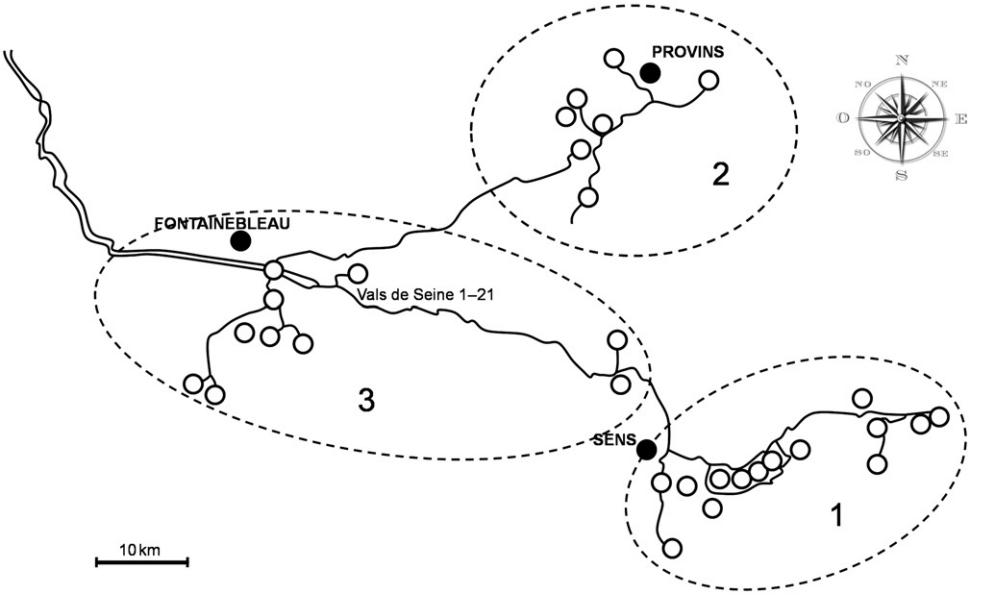

Figure 8. Class distribution in the studied area.

Table 5. Comparison between analytical parameters recorded in the years 2004 and 2005 for the water sources changing class.

Sources	Year	Class	ECOLI	STRF	TUR	COND	AKM	Cl	SO_4	NO_3	HRD
ARMENTIERE	2004	1	10.0	8.0	0.2	531.0	22.1	10.0	4.0	26.0	24.5
	2005	0	29.0	14.0	0.4	510.0	21.9	10.0	4.0	30.0	25.2
AUGEOIECAP	2004	1	6.0	0.0	0.5	508.0	21.0	10.0	5.0	21.0	23.4
	2005	3	1.0	0.0	0.2	529.0	21.5	12.0	6.0	29.0	24.1
BOUILLARDE	2004	3	0.0	0.0	0.3	512.0	22.9	10.0	2.0	15.0	25.5
	2005	1	9.0	4.0	0.3	509.0	23.5	9.0	3.0	21.0	24.8
MIROIR	2004	1	15.0	2.0	0.4	497.0	21.3	9.0	5.0	18.0	24.1
	2005	0	23.0	8.0	0.8	471.0	20.5	9.0	4.0	20.0	23.3
STPHILSOU	2004	1	5.0	4.0	0.2	549.0	22.3	11.0	6.0	28.0	25.3
	2005	3	0.0	0.0	0.2	547.0	23.4	11.0	5.0	22.0	25.8

Whatever variation of the analytical parameters occurred, it resulted not statistically significant so that water quality could be considered unmodified. Only five sources resulted shifting to a different class because of a significant variation of some analytical parameters.

The distribution within the classes in both the investigated years and the relative mean values of the used parameters are listed in Tables 4 and 5, respectively. It is plain that the class change for the sources *Bouillarde* (third-to-first), *Stphilsou* and *Augeoicap* (first-to-third) was caused by the considerable variation of ECOLI and STRF. Analogously, the high increase of both microbiological parameters caused a significant increase of distance from all the class models for the sources *Armentiere* and *Miroir*. In the second year, both objects resulted excluded from the defined classes, being therefore *outliers* (class 0).

The proposed procedure proved to be a useful tool to rapidly highlight any change occurring at the time about the water quality. It resulted very effective to handle a complex data matrix carried out from the analysis of a high number of water sources and could be implemented as a fast and efficient method in routine analysis.

5. Conclusions

Cluster analysis and SIMCA classification were applied to water chemical data from a large region to verify the water quality during a period. The procedure was applied on 68 analytical objects dealing with the potable water of Paris. CA and PCA provided to select just nine variables carrying the most useful information and to extract from the complex data matrix the principal factors causing the samples distribution. The combined use of these multivariate techniques enabled the classification of water samples into three distinct classes on the basis of their hydrochemical characteristics. The distribution of the water sources resulted to be controlled largely by the microbiological parameters. The classification model by SIMCA algorithm provided then powerful means of monitoring the quality of water during the time.

The treatment of data demonstrated to be rapid and easy to use and, above all, independent by a subjective interpretation of the analyst. The method could be used in water management to identify eventual risks from any water pollution, potentially dangerous to people's health or the environment.

Acknowledgement

The authors thank the Ministry of University and Research (MIUR) of Italy for financial support of this work.

References

[1] L. Ignatiades, A. Vassiliou, and M. Karydis, Hydrobiologia **128**, 201 (1985).
[2] L. Ignatiades, M. Karydis, and P. Vounatsou, Mar. Pollut. Bull. **24**, 238 (1992).
[3] M. Karydis, J. Environ. Sci. Heal. A **29** (4), 775 (1994).
[4] A. Moriki and M. Karydis, Environ. Monit. Assess. **33**, 1 (1994).
[5] V.D. Engle, K.K. Summers, and G.R. Gaston, Estuaries **17**, 372 (1994).
[6] H.M. Feder, U.S. Naidu, S.C. Jewett, J.M. Hamedi, W.R. Johnson, and T.E. Withledge, Mar. Ecol-Prog Ser. **3**, 171 (1994).
[7] R.M. Warwick, Mar. Pollut. Bull. **19**, 259 (1988).
[8] M. Karydis, Environ. Monit. Assess. **22**, 123 (1992).
[9] R. Reyment and K.G. Jöreskog, *Applied Factor Analysis in the Natural Sciences* (Cambridge University Press, London, 1993).
[10] E. Marengo, M.C. Gennaro, D. Giocosa, C. Abrigo, G. Saini, and M.T. Avignone, Anal. Chim. Acta **317**, 53 (1995).
[11] W.J. Krzanowski, Stat. Comput. **7**, 87 (1997).
[12] D. Brodnjak-Vončina, D. Dobčnik, M. Novič, and J. Zupan, Anal. Chim. Acta **2462**, 87 (2002).
[13] W.M. Jarman, G.W. Johnson, C.E. Bacon, J.A. Davis, R.W. Risebrough, and R. Ramer, Fresenius J. Anal. Chem. **359**, 254 (1997).
[14] E. Reisenhofer, G. Adami, and P. Barbieri, Water Res. **32**, 1193 (1998).

[15] W.D. Alberto, D.M. Del Pilar, A.M. Valeria, P.S. Fabiana, H.A. Cecilia, and B.M. De Los Angeles, Water Res. **35**, 2881 (2001).
[16] S. Wold and M. Sjostrom, in *Chemometrics, Theory and Application*, edited by B.R. Kowalski (American Chemical Society, Washington, DC, 1977), p. 243.
[17] M.A. Sharaf, D.L. Illman, and B.R. Kowalski, *Chemometrics* (John Wiley and Sons, New York, 1986).
[18] D.L. Massart, B.G.M. Vandeginste, S.N. Deming, Y. Michette, and L. Kaufman, *Chemometrics: A Textbook* (Elsevier, Amsterdam, Netherlands).
[19] M.J. Sáiz-Abajo, J.M. González-Sáiz, and C. Pizarro, J. Near. Infrared Spec. **12**, 207 (2004).
[20] S. Capri, R. Pagnotta, and M. Pettine, Manuali e linee guida, Apat-Irsa CNR, Rome (2003).
[21] V.T. Esbensen and H. Martens, Chemometr. Intell. Lab. **2**, 239 (1987).
[22] A.P. Reynolds, G. Richards, B. De La Iglesia, and V.J. Rayward-Smith, J. Math. Model. Algorithms **5**, 475 (2006).
[23] D.L. Massart and L. Kaufman, *The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis* (Wiley, New York, 1983).
[24] P. Willet, *Similarity and Clustering in Chemical Information Systems* (Research Studies Press, Wiley, New York, 1987).
[25] S. Sharma, *Applied Multivariate Techniques* (Wiley, New York, 1996).
[26] K.S. Booksh, *Chemometric Methods in Process Analysis*. edited by Meyers, R.A., Encyclopedia of Analytical Chemistry (John Wiley & Sons, Chichester, UK, 8145, 2000).
[27] H. Martens and T. Naes, *Multivariate Calibration*, 2nd ed. (Wiley, Ltd. Chichester, United Kingdom, 1989).
[28] G. Downey, P. McIntyre, and A.N. Davies, J. Agric. Food Chem. **50** (20), 5520 (2002).
[29] M. Blanco, S. Maspoch, I. Villarroya, X. Peralta, J.M. González, and J. Torres, Applied Spectroscopy **55** (7), 834 (2001).
[30] K.W. Johnson, P.M. Langdon, and M.F. Ashby, Mater. Des. **23**, 1 (2001).
[31] C. Hennig, J. Multivar. Anal. **86**, 183 (2003).
[32] K.P. Singh, A. Malik, D. Mohan, and S. Sinha, Water Res. **38**, 3980 (2004).
[33] Societè Anonyme de Gestion des Eaux de Paris, Rapport de délégation de service public – EAU DE PARIS (2006).