Chemometric modeling of organic contaminant sources in surface waters of a mediterranean river basin (Catalonia) district

Alejandro G. García-Reiriz,[1] Alejandro C. Olivieri,[1] Elisabeth Teixidó,[2] Antoni Ginebreda[3] and Romà Tauler[3]

[1] *Department of Analytical Chemistry, Faculty of Biochemistry and Pharmaceutical Sciences, National University of Rosario, Rosario Institute of Chemistry (IQUIR-CONICET), Suipacha 531, Rosario, S2002LRK, Argentina.*

[2] *Agència Catalana de l'Aigua, Provença 204-208, 08036 Barcelona, Spain.*

[3] *Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Catalonia, Spain.*

Abstract

Chemometric methods are applied to the analysis and interpretation of large multivariate data sets obtained in environmental monitoring studies. Concentrations of multiple organic compounds were measured in river samples taken from several sampling sites, at various geographical locations, during a number of campaigns and/or sampling time periods. Samples were collected and analyzed as part of an extensive multi-annual monitoring program from a mediterranean river basin (in Catalonia, at the northeast of Spain) by the Water Catalan Agency. Due to the great amount of multivariate data stored in environmental databases and to their complexity, chemometric modeling methods like Principal Components Analysis (PCA) and Multivariate Curve Resolution with Alternating Least-Squares (MCR-ALS) coupled to appropriate mapping representations are proposed for the evaluation of the environmental quality of the studied rivers. Results achieved in this study are intended to be a contribution to water quality assessment and evaluation of contamination of surface waters in Catalonia, and to support public policies of environmental control and management in the region under study.

Keywords: Chemometrics, Multivariate curve resolution alternating, Principal Components Analysis, Environmental Monitoring, Surface Water.

1. Introduction

Environmental data bases constitute a suitable option for monitoring and control of water systems. A potential disadvantage associated with large data bases is the difficulty in their interpretation for decision making. A good alternative for analysis is to resort to multivariate methods. They allow extract information on the behaviour of the variables involved in the several different studied dimensions (e.g., time and space). When combined with geo-positioning tools, identification of the main sources of contamination is possible, whether they are point or diffuse, or from anthropologic or geologic origin.

In this work, a data base from the Agencia Catalana de Agua (Catalonian Water Agency), containing information on a large number of potential contaminants, was studied using different chemometric techniques. Chemometrics provides powerful tools for the modeling and interpretation of large environmental multivariate data sets generated within environmental monitoring programs.[1,2] The goal of these studies is the computation, screening and graphical display of patterns in large data sets, looking for possible contamination sources and their distribution. Principal Component Analysis (PCA) is one of these multivariate methods for data analysis, which is frequently used in environmental exploratory studies.[3,4] PCA allows the transformation and visualization of complex data sets into a new and simpler perspective, in which the more relevant environmental information can be easily perceived. Using PCA, contamination patterns may be identified and their geographical and temporal distributions may be investigated. PCA has been applied in previous studies by several authors to various types of environmental data sets, such as those stemming from waters, biota and sediments.[5-9] Another method here applied is Multivariate Curve Resolution Alternating Least Squares (MCR-ALS), a powerful

chemometric tool with an increasing application for the analysis of environmental monitoring data sets.[10] It has been recently validated for the identification of environmental pollution patterns in surface water.[11] This latter study was intended to model pollution in surface water of the Ebro River delta (a smaller area of around 300 km^2), during the main growing-season of the rice crop. Other chemometric methods have also been applied to the investigation of environmental data, such as partial least-squares (PLS),[7,8] parallel factor analysis (PARAFAC) and Tucker3 models.[12] The use of multivariate factor analysis, such as those proposed in the present work has also been discussed in several books.[1,13] In the present work, the research is focused on a large environmental data set, obtained during a study of natural surface waters from the rivers of Catalonia (northeast of the Iberian Peninsula), including the analysis of multiple organic contaminants. In the framework of this extensive multi-annual environmental monitoring program from the Water Catalan Agency, organic contaminant compounds in the entire geographical area of Catalonia were analyzed during the years 1997-2004. The occurrence of organic compounds in natural surface waters is attributed to the presence of several industrial, agricultural and urban wastewater points and to diffuse contamination sources. The Catalonia geographical area is one of the most industrialized areas of Spain, and it is of interest to evaluate its environmental situation. Although this investigation provides results which could be considered only of concern for the particular area under study, the obtained results and conclusions are of general interest from an environmental point of view to other river basin areas, especially those which are close to the Mediterranean coasts, which have the same type of climate, hydrology, vegetation and human activities (industrial, agricultural, urbanization) operating over the river water systems. This work is also of interest from a

chemometric point of view, specifically for the comparison of the results furnished by PCA and MCR-ALS, which are often used in environmental studies.[1,14] Other recent publications are concerned about the presence of persistent organic compounds in Catalonia,[15-18] (ACTUALIZAR)which were also analyzed in several types of environmental compartments. Additionally, other recent examples exist proposing rather similar approaches for the resolution and interpretation of major contamination sources of surface waters operating in several river basins over the world.[19] The two main objectives of this work are thus: 1) the investigation of main long-term diffuse contamination sources of organic contaminants in the Catalonia river basin area, and 2) the estimation of their geographical distribution, in order to contribute to the evaluation of the environmental health of the surface waters of the region under study. To achieve these two goals, multivariate data methods of analysis based on PCA and MCR-ALS are applied and compared.

Diffuse and point pollution in the Catalonia River basin area arising from agriculture, industry and human sewage, is an issue of great concern, since changes in climatic conditions and land use practices have produced large scale adverse impacts on both water quality and quantity. Through the environmental monitoring program performed at several sampling sites and environmental compartments of the network, a large amount of concentration values of chemicals spread into the Catalonia river basin were obtained. In order to derive useful environmental information from the data, the application of modern chemometric methods based in new multivariate factor analysis[20] tools is proposed. The basic assumption of these methods when they are applied to environmental data tables is that each value of a measured variable in a particular sample is due to the sum of

contributions from individual independent sources of different origin. Each one of these sources is characterized by a particular chemical composition profile and is distributed among samples in a different way. As a result of the application of chemometric methods, the main point and diffuse sources of contamination in the environment and their origin may be identified and their distribution profiles among samples (geographical, temporal, among environmental compartments) are characterized.

The distribution of contamination sources and their impact over the territory can be assessed by the use of geographical information systems,[21,22] by means of cartographic techniques of symbols and pollution prediction maps. Geo-statistical methods[23-25] based on mathematical and statistical functions are used, which allow the estimation of continuous surfaces using the measured variables to predict unknown value by interpolation and, at the same time, give an estimation of the errors associated to these predictions.

Finally, it is worth mentioning that the proposed techniques and tools can contribute to the management of the river basins under the application of the Water Framework Directive (2000/60/EC)

2. Experimental data

Samples were not taken in a special monitoring design in the zone under study. Sampling was performed from locations previously decided by the Catalan Water Agency, and they were taken once per year, although not every site was studied every year. They were obtained over a period of eight years, from 1997 to 2004. The geographical area under study where these compounds were analyzed covered several small and medium size rivers in the Catalonia region, such as (from North to South Catalonia coast): Muga, Fluvià, Ter,

Daró, Riudaura, Tordera, Besòs, Llobregat, Foix, Gaià, Francolí, Riu de Canyes, Noguera Pallerasa, Noguera Ribagorçana, Segre, Ebre, and Garona rivers (see location of these rivers in Fig. 1). With the exception of the last five rivers, the remaining ones are typical Mediterranean rivers, characterized by short length and small catchment areas, steep slopes and drastic flow variations between the dry summer season, and sudden flow increases after the fall and spring rains that often cause floods and damages.

These data sets have been analyzed as they were provided by the Catalan Water Agency, and no attempt was made to have an optimal design of the best sampling sites for the purpose of environmental source identification. It should thus be noted that because of these sampling limitations, information about temporal evolution of the river contamination sources in Catalonia rivers could only be obtained in a limited way. These provisional results should be confirmed with new data obtained using a better designed monitoring sampling plan, including more recent years. Some work is pursued at present in this direction.

Water samples were collected from the already numbered points of the quality network established by the Catalan institution of water ('Agencia Catalana de l'Aigua'), indicated in Fig. 1. Samples were kept in 1 liter glass bottles fitted with Teflon-lined caps, leaving no headspace. After sampling, they were preserved in cold (not exceeding 5ºC) until the moment of performing the analysis.[35]

The following volatile compounds were analyzed in the samples: ethylbenzene, *m,p*-xylene, and toluene using headspace analysis with GC-FID[36-37]. Headspace analysis was performed with a Varian Genesis headspace autosampler connected to a Varian Star 3600 gas chromatograph. Samples were equilibrated at 70 ºC for 4 min, mixed at 80% of

full power for 7 min, and, after mixing, stabilized for 1 min. The sample loop volume was 1 mL, line and valve were maintained at 150ºC and vials were pressurized at 7 psi. These conditions resulted in the highest sensitivity and reproducibility. Compounds were separated on a 75 m × 0.53 mm i.d. × 3 μm film DB-624 fused-silica column from J&W. The GC operating temperatures were: injector 160 ºC detector 300ºC oven 40ºC (5 min) programmed at 5ºC/min to 250ºC. Helium at 9 psi was the carrier gas.

Other volatile compounds: 1,1,1-trichloroethane, 1,2-dichloropropane, 1,2,4-trichlorobenzene, 1,2-dichlorobenzene, bromodichloromethane, bromoform, chloroform, dibromochloromethane, tetrachloroethylene, carbon tetrachloride, and trichloroethylene were analyzed by headspace with GC-ECD.[36-37] Headspace analysis was performed with a Varian Genesis headspace autosampler connected to a Varian Star 3600 gas chromatograph. Samples were equilibrated at 70 ºC for 4 min, mixed at 80% of full power for 7 min, and, after mixing, stabilized for 1 min. The sample loop volume was 1 mL, line and valve were maintained at 150ºC and vials were pressurized at 7 psi. These conditions resulted in the highest sensitivity and reproducibility. Compounds were separated on a 30 m × 0.32 mm i.d. × 1.8 μm film DB-624 fused-silica column from J&W. The GC operating temperatures were: injector 160 ºC detector 300ºC oven 40ºC (5 min) programmed at 6ºC/min to 140ºC (1 min) and at 15ºC/min to 220ºC (5 min) Helium, at 7 psi, was the carrier gas.

The following compounds: chlorpyrifos, diazinon, phenitrotion, malathion, acenaphthene, acenaphthylene, anthracene, phenanthrene, fluoranthene, fluorene, pyrene, 4,4'-dichlorodiphenyldichloroethane (DDD), 4,4'-dichlorodiphenyldichloroethylene (DDE),

4,4'-dichlorodiphenyltrichloroethane (DDT), α-, β-, δ- and γ-hexachlorocyclohexane, endosulfan I, endosulfan II, endosulfan sulfate, hexachlorobenzene, were analyzed by liquid-liquid extraction according to the method 625 from the U.S. Environmental Protection Agency[38]. One liter of sample was added with internal standards (anthracene-d10 and decachlorobiphenyl) and extracted twice with dichloromethane (150 mL and 100 mL) by stirring for 10 min. The organic extracts were combined and dried with anhydrous Na_2SO_4. Dichloromethane was removed under reduced pressure, first in a round bottom flask and further in a conic flask, until a volume of 0.5 mL. The concentrate was transferred to a 1 mL conic vial, washing the flask with isooctane, and dried under N_2 stream until a final volume of 100 µL for HRGC/MS and/or HRGC/ECD analysis. Surrogate standard mixture (nitrobenzene-d5, 2-fluorobiphenyl and 4-terphenyl-d14) was added to the extract.

Final extracts were analyzed by HRGC. Organochlorine pesticides (4,4'-dichlorodiphenyl-dichloroethane (DDD), 4,4'-dichlorodiphenyldichloroethylene (DDE), 4,4'-dichlorodiphenyltrichloroethane (DDT), α-, β-, δ- and γ-hexachlorocyclohexane, endosulfan I, endosulfan II, endosulfan sulfate, hexachlorobenzene) were quantified by HRGC/ECD and their structure identity was confirmed by HRGC/MS. The other pesticides and PAH (chlorpyrifos, diazinon, phenitrotion, malathion, acenaphthene, acenaphthylene, anthracene, phenanthrene, fluoranthene, fluorene, pyrene) were identified and quantified by HRGC/MS.

The HRGC/MS were performed on an integrated quadrupole GC/MS MD-800 from Fisons (Manchester, UK). Helium was used as carrier gas (at a constant flow of 1.2 mL/min) in a DB-5MS column (30 m × 0.25 mm i.d., 0.25 µm film thickness). The

program was from 90ºC (held 5 min) to 240ºC at 6ºC/min and to 310ºC (maintained for 10 min) at 10ºC/min. Injector temperature was 280ºC, and the injection mode was splitless for 90 s. The scanning was 40-500 m/z at 1 scan/s. MS spectra were compared with NIST spectra library (62,235 spectra) and with Wiley spectra library 5th ed. (138,111 spectra), and each compound was identified according to its best mass spectrum fitting. The HRGC/ECD analysis was performed on a Shimadzu GC-9A model gas chromatograph equipped with an ECD-9 model detector of the same firm. Helium was used as carrier gas at a flow of 2.6 mL/min in a DB-5 column (30m × 0.2 mm i.d., 0.25 µm film thickness). The program was from 130ºC (held 1 min) to 140ºC (maintained for 1 min) at 12ºC/min, and from 140ºC to 310ºC at 4ºC/min; the final temperature was further held for 10 min. Injector and detector temperature were 290ºC and the injection mode was splitless for 1 min.

Pentachlorophenol was derivatized to its acetyl pentachloro derivative by treatment with 0.7 mL of acetic anhydride. For the extraction, 1 mL of hexane was previously added to 100 mL of sample, and 0.7 g of $NaHCO_3$ were added to the water sample as buffer. The organic extract was transferred to a 1 mL vial to analyze by HRGC/MS. Final extracts were analyzed by HRGC/MS with an integrated quadrupole GC/MS MD-800 from Fisons (Manchester, U.K.). Helium was used as carrier gas (at a constant flow of 1.2 mL/min) in a DB-5MS column (30 m × 0.25 mm i.d., 0.25 µm film thickness). Chromatograms were recorded under time-scheduled selected ion monitoring (SIM) using acquisition windows from 5-28 min, and 266, 264, 268, 308 m/z values. The dwell time was set at 0.08 s. The program was from 90ºC (held 2 min) to 130ºC (0 min) at 15ºC/min, and from 130ºC to

310ºC at 10ºC/min; the final temperature was further held for 5 min. Injector temperature was 280ºC, and the injection mode was splitless for 90s.[39-40]

3. Chemometric methods

3.1. Data pre-treatment

Data sets were organized in one data table or data matrix. The rows of this data matrix identified samples at the various geographical sites and sampling dates, while columns (variables) represent the analyzed chemical compounds. Dimensions of the data matrix were 303×37, corresponding to 303 observations (sampling sites and times) and 37 analyzed compounds. Prior to the application of chemometric data processing, the different variables contained in the data matrix were evaluated. Variables whose concentrations were only 5% over the limit of detection were removed, because they did not provide useful information. In other cases, for other variables having a significant amount of values above their detection limit, the remaining values below the limit of detection were replaced by half of this limit. The distribution of variables was studied, and values which were very far from the mean of the distribution were replaced by the maximum value of the specific variable in the same sample site when this value was excluded.

Scaling the elemental values over the sample is generally recommended, because the different compound concentrations can have large variations among them. Normalizing the concentrations will provide more equal weight to chemical species with substantially different concentrations. However, it should be kept in mind that scaling will lose information on the relative size and relative errors associated with the various data variables.

In determining the best data pretreatment method, a compromise was sought to find the method which provided the easiest and more optimal interpretation of possible contamination sources. The following data pretreatment methods were investigated: 1) concentration values were log-transformed, i.e., the decimal logarithm of all data matrix values were calculated; this transformation of experimental data has been recommended for skewed data sets,[26,27] such as those usually found in environmental studies, where a large amount of the values are low, with a minor global contribution of high values, 2) log concentration values of each compound in the several samples were mean centered, i.e., the mean of the log concentration values of the same compound in the several samples (mean of each column variable of the data matrix) was subtracted from each log concentration value, 3) log concentration values of each compound in the several samples were scaled, i.e., each log concentration value was divided by the standard deviation of the log concentration values of the same compound in the several samples (standard deviation of each column variable of the data matrix), 4) log concentration values of each compound in the several samples were auto-scaled, i.e., previous mean centering and unit variance scaling pretreatment methods were combined, 5) data were scaled based on either the whole set of values or on a yearly basis, 6) MinMax transformation, and 7) MinMax of log concentration values (the two last ones explained in more detail below).

Of all the data pre-processing methods mentioned above, the MinMax with logarithmical transformation was the most successful one, hence some additional details are provided below. The specific expression for the MinMax transformation is:

$$x_{\text{transf}} = \frac{x - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})} \quad (1)$$

where **x** is a vector of log-values, max(**x**) and min(**x**) are the maximum and minimum of **x** respectively, and x and x_{transf} are the raw and transformed elements. MinMax was applied in two different ways: 1) taking as minimum and maximum values those corresponding to the complete data set, and 2) taking the minimum and maximum of each yearly campaign. In the first MinMax mode, the differences among years are superimposed to the differences within each year, while in the second mode the scaling differences among years are decreased. Therefore, the information provided by these two different transformation modes is not the same: the first one allows observe scale/intensity differences in the scores among years, while the second allows the detection of specific variations within each year. In this work, more interpretable results were obtained using the first approach.

3.2. PCA

PCA assumes a bilinear model to explain the observed data variance using a reduced number of components, which are orthogonal. For a detailed description of this well-known methodology in chemometrics and other multivariate statistical data analysis methods see previous references.[3,4] The bilinear decomposition may be written by the element wise equation:

$$d_{ij} = \sum_{n=1}^{N} x_{in} y_{in} + e_{ij} \qquad (2)$$

where d_{ij} is one of the entries of the experimental data matrix (concentration of one organic compound) from the *i*th row (a particular sample) and the *j*th column (a specific organic compound), x_{in} is the corresponding *n*th score element for the sample *i*, y_{jn} is the corresponding *n*th loading element for the variable *j* and e_{ij} is the residual not modeled by

the sum of *N* components or contributions. The same bilinear equation can be written in matrix form as:

$$\mathbf{D} = \mathbf{X}\mathbf{Y}^{T} + \mathbf{E} \quad (3)$$

where **D** is the experimental data array expressed as a data matrix. Eq. (3) describes the decomposition (matrix factorization) of matrix **D** into two matrices, the loading matrix $\mathbf{Y}^{T}$ and the score matrix **X**. The loading matrix $\mathbf{Y}^{T}$ identifies the main sources of the data variance by means of their chemical composition (composition loadings), which eventually may be related to the main patterns and sources of contamination. The score matrix **X** provides sample scores for these data variance patterns, indicating the geographical and temporal sample distribution of these patterns. PCA solves Eq. (3) under orthogonal constraints. Each successively extracted principal component explains maximum variance. The determination of the complexity of the model in PCA (i.e., the number of principal components) is performed as a compromise between several goals: model simplicity (few components), maximum variance explained by the model (more components), and model interpretability.

3.3. MCR-ALS

MCR-ALS[28,29] works with the data array arranged in a column-wise augmented data matrix $\mathbf{D}_{aug}$, such as that for PCA (described in the previous Section). The bilinear decomposition of the augmented matrix $\mathbf{D}_{aug}$ is performed according to the same expression already given for PCA [i.e., Eq. (3)]. Although only the recovered information in $\mathbf{Y}^{T}$ appears to be explicitly related to one of the three modes, the matrix **X** implicitly contains the information related to the matrices **X** and **Z** in the remaining two modes, and

they can be recovered by appropriate refolding followed by singular value decomposition (SVD) analysis, as described before for PCA. In contrast to PCA, however, during the ALS optimization phase of MCR-ALS, the selected constraints were non-negativity for the profiles in both modes (for the augmented scores mode and for the loadings in the second mode), while the loadings in the second mode were normalized to equal length.

3.4. MCR-ALS for trilinear models

Trilinear models can be implemented iteratively as a constraint during ALS optimization in the MCR-ALS method.[11,30,31] The application of MCR-ALS using this constraint should not be considered to be equal to a standard bilinear decomposition of the augmented two-way data matrix $\mathbf{D}_{aug}$. During the ALS optimization, each individual profile of the augmented scores matrix **X** is constrained to fulfill the trilinearity condition independently and iteratively. The same procedure used previously for the recovery of the loadings in the three modes from the augmented scores matrix obtained by PCA or MCR-ALS is applied now inside/during the ALS optimization instead of at the end of the optimization as in PCA. Each column of the **X** matrix is appropriately folded at each ALS iteration step to give a matrix with a number of rows equal to the number of sampling sites and eight columns corresponding to each of the years (1997-2004). SVD of this folded scores matrix gives the loadings in the first and third modes for the considered component. These two loadings describe the common variation captured by ALS in the two modes (sampling sites and years) for that particular component. The Kronecker product[32-34] of these two new loading vectors gives the new augmented scores vector which substitutes the corresponding column of the **X** scores matrix. When this constraint is inserted during each

step of the ALS iterative optimization procedure, it forces the shape of the loadings vector in the first mode (describing the sampling site variation of the considered component) to be the same for the eight years. Moreover, it captures the intensity (scale) variation of this component in the loadings of the third mode, showing the scale differences of this component among the eight years. This is precisely what is implied by the trilinear PARAFAC model described by Eq. (3) and, in practice, results obtained by MCR-ALS with the trilinearity constraint applied to all the components of the system should give practically the same results as the application of PARAFAC model based methods. However, the main advantage of the trilinearity constraint in MCR-ALS over the PARAFAC model based methods is that this procedure in MCR-ALS is applied independently for each component and that it is not compulsory to apply it for all the resolved profiles in **X**. Actually, several columns of **X** matrix can be constrained in several manners during MCR-ALS. This makes a clear distinction with PARAFAC where all resolved components should fulfill the sought trilinear condition.

4. Results and discussion

In Table 1, a summary of the descriptive statistics obtained in the analysis of the previously referred compound using the procedure described above is given. In this Table, for every compound analyzed, the minimum (usually at the limit of detection), the maximum value, the mean, the median, the standard deviation and the % of values above the detection limit are given. This database contains 37 compounds measured in several sampling sites from Catalonia rivers between the years 1997 and 2008, during a total of 8 years. The sampling sites were 35 but samples were not taken from all sites in all campaigns. For this reason, the size of the matrix was 303×37, where the first mode includes all samples taken in different sites and times, and the second mode contains the measured chemical compounds. In this table, variables are given in their own different scales, which can be very different in some case, being necessary the use of an appropriate scaling preprocessing method to give them a similar weight during their analysis. When minimum, maximum, average and median values are observed, it is easily concluded that data distribution is not following a normal distribution of values and that probably are better described by a log normal distribution of values. This indicates the possibility of using log transformation of data to better investigate the data variance and to decrease the weight of extreme variable values,

Since in this work source apportionment was intended, data were not initially mean centered. In multivariate curve resolution and source apportionment and receptor modeling studies, the main interest is in actual values and not in their deviations from the mean. MinMax was the finally preferred data pre-processing tool, because it minimized differences in variable scales, allowing the comparison of results without giving more

weight or importance to a particular variable. MinMax was applied separately to samples of each campaign rather than jointly to the entire data base, decreasing in this way the differences among the several campaigns and thus providing a better comparison of the variation of the compound concentrations within them. A first approximation to estimate the number of components was obtained by PCA, which indicates the number of possible major independent sources of pollution affecting measured data. The number of components was estimated by examining the size of the changes in explained variance in PCA as a function of the number of principal components. Three components were proposed to model the MinMax pre-processed data matrix, which allowed to explain 62.9% of the overall variance.

In Fig. 2, loadings obtained by PCA are shown. It can be observed that the first component (% var expl) describes the average contamination affecting the geographical region under study over the investigated years, and the other two are components describing the contrast with more specific contamination sources. The second component (% var expl) highlights the contamination coming from some pesticides like hexachlorcyclohexane (alpha, beta and delta isomers, as well as the gamma isomer lindane), endosulfan (I, II and sulfate) and diazinon. Finally, the third component (% var expl) describes the different behavior of the halomethanes (bromodichloromethane, chloroform, dibromochloromethane), chlorinated ethenes (tetrachloroethylene, trichloethylene), carbon tetrachloride and chlorobenzenes and halopropanes. The corresponding PCA scores describe the geographical distribution of these contamination patterns, marking what sites were more highly contaminated on the average (PC1 scores) and what sites were more affected by more specific agricultural contamination sources

(PC2 scores) and by more industrially related contamination sources (PC3 scores). Because PCA defines the same space vector space as the one obtained by MCR-ALS decomposition using the same number of components (see below), PCA score plots have been omitted for brevity. An advantage of MCR-ALS over PCA is the possibility of applying natural constraints like non-negativity, making easier the physical interpretation of the results. For this reason, the discussion about the possible sources or patters was mainly focused on MCR-ALS results.

4.1 MCR-ALS results of the complete data set applying non-negativity constraints

MCR-ALS was first applied to the complete data set (37 compounds in all sampling sites and in 8 years) with non-negativity constraints. The trilinearity constraint could not be applied in this case, since not all the sites were sampled in all the campaigns and therefore the data set could not be arranged as a three-way data array. Explained variance was 61.1 % for three components. These three components are interpreted in environmental terms as follows (see Fig. 3).

The first component (Figs. 3 and 4) (33.8% of the total variance explained) is dominated by PAHs (polycyclic aromatic hydrocarbons, i.e., acenaphthene, phenantrene, fluorantene, fluorene, pyrene, etc.), THMs (trihalomethanes, chloroform, bromoform, bromodichloromethane) and minor contributions of other compounds such chlorinated ethenes (tri and perchloroethylene). The former group of compounds reflects diffuse contamination related to combustion engines, characteristic of areas with heavy traffic and industry. The second group can be associated to disinfection by-products generated during the chlorination treatment of drinking-water and returned to the environment through discharges from WWTP. As a whole, the contamination pattern described by this first

component corresponds to areas with heavy industrial and urban pressure. It is mainly located in Barcelona and its surrounding metropolitan area.

The second component (Figs. 3 and 5) (14.8% of the total variance explained) is dominated by DDT related compounds (DDT and its metabolites DDD and DDE), hexachlorobenzene, halomethanes (bromodichloromethane, chloroform, dibromochloromethane), chlorinated ethenes (tetrachloroethylene, trichloethylene), carbon tetrachloride, and minor contributions of chlorobenzenes and halopropanes. Such a profile is specifically related to the chloro-alkali industry located in the low Ebro (Flix), which at present manufactures these chlorinated solvents, but was also a major producer of DDT in the past. Even though the production of DDTs was discontinued after their banning in Europe, the reported presence of polluted sediments in the river still generates downstream Flix a background contamination of DDT (and mostly of its metabolites DDE and DDD). Hexachlorobenzene is also generated in the same chloro-alkali chemical plant as by-product during the electrolysis process.

The third component (Figs. 3 and 6) (12.5% of the total variance explained) is dominated by pesticides like diazinon, chlorpyriphos, phenitrotion, malathion, hexachlorocyclohexan (alpha, beta and delta isomers, as well as the gamma isomer lindane), endosulfan (I, II and sulfate), and other minor contributors such as pentachlorophenol and chlorinated solvents. The strong presence of these pesticides is consistent with its occurrence in agriculture dominated areas, like Lleida or the rural areas in the neighborhood of the Barcelona metropolitan area (Maresme, Anoia etc.) or the floodplains and deltas of the main rivers, all of them characterized by intensive agriculture.

Figures 4, 5 and 6 summarize the geographical distribution of these three major contamination patterns previously described. These plots were obtained averaging the scores of all the campaigns in a single matrix to simplify their visualization.

4.2 MCR-ALS results of the reduced data set applying non-negativity and trilinearity constraints

From the 37 compounds included in the original data base, 24 (see Table 1) were removed because the corresponding concentrations were not measured in all campaigns and all sites. The remaining 13 compounds were: diazinon, phenanthrene, fluoranthene, fluorene, pyrene, pentachlorophenol, γ-hexachlorocyclohexane (lindane), 1,1,1-trichloroethane, bromodichloromethane, chloroform, dibromochloromethane, tetrachloroethylene and trichloroethylene. A new data set was built with only these 13 compounds measured in 17 sampling sites at 8 campaigns. The whole data set gave a data table or matrix of size 136×13, i.e. concentrations of the 13 compounds at the 136 different samples.

MCR-ALS was then conducted in two manners: (1) only applying the non-negativity constraint to loadings and scores, and (2) applying non-negativity and also trilinearity constraints (give references here). This latter constraint is more restricted, leading to a decreased percentage of explained variance, but it has the advantage of separating the between year campaigns patterns of the resolved components. It demands some data reorganization, in such a way that all campaigns display the same number of sampling sites, leaving only 17 studied locations. Results obtained by these two MCR-ALS analyses were rather similar, in terms of explained variances, 61.2% and 54.4% for the two

approaches respectively, and also resulted rather similar in relation to the composition of the resolved components. This suggested that the data could be approximated by the trilinear model, giving more easily interpretable component profiles, especially in terms of the distribution and geographical representation (mapping) of the resolved components describing the different contamination patterns under study. For brevity only the results obtained using the trilineraity constraint are finally given in this discussion.

Figure 7 shows the results corresponding to non-negativity/trilinearity constrained MCR-ALS study. If the loading profiles corresponding to the different variables are compared, some patterns are observed always within the different components. Three different patterns grouping different compounds were identified (total explained variance 54.4%): (1) A first component (30.1% of the total variance explained) is dominated by diazinon, phenanthrene, fluoranthene, fluorene, pyrene, lindane, tetrachloroethylene and trichloroethylene; (2) a second component (13.2% of the total variance explained) is dominated by phenanthrene, fluoranthene, fluorene, pyrene, bromodichloromethane, chloroform, dibromochloromethane; and (3) a third component (11.1% of the total variance explained) is dominated by pentachlorophenol, 1,1,1-trichloroethane, tetrachloroethylene and trichloroethylene. Once identified the chemical composition of the main contamination patterns, the localization of these patterns and corresponding possible sources are investigated

The first MCR-ALS component or contamination pattern (30.1%) defined by the first group of compounds (see above) is mainly localized in the following sampling sites (see Fig. 7): (1) Anoia river, Vilanova del Camí, (2) Foix river, Castellet i la Gornal, and (3) Clamor de les Canals, Lleida. All these locations correspond to rural and middle size

villages and according also to the composition of this possible source identifies a general contamination source of mixed agricultural and population sources. Second (13.2%) and third (11.1%) components are focused in regions near Barcelona (Fig. 7) and give patterns corresponding to industrial and heavy population sources. Specifically, the second component corresponds to: (1) Congost river, Montornès del Vallès, (2) Besòs river, Montcada i Reixa and Santa Coloma de Gramenet, Barcelonès, and (3) Riera de Rubí, Castellbisbal Finally, the third component is localized in: (1) Mogent river, Montornès Del Vallès, (2) Besòs river, Montcada i Reixac, (3) Besòs river, Santa Coloma de Gramenet, Barcelonès, and (4) Llobregat river, Abrera. As regards the time evolution of these components, it can be concluded that both the first and third component have a growing trend over time, while the second one appears to be decreasing.

Results obtained with trilinearity and non-negativity constraints do agree with previous results obtained modeling the whole data set with MCR-ALS bilinear modeling. Again three MCR components were used to justify the observed data variance. Interpreting the composition and location of each component we can conclude that: the first component can be associated to the presence of several pesticides related to agriculture activities, the second component can be associated to combustion engines characteristic of areas with heavy traffic and industry, and third component profile can be specifically related to the chloro-alkali industry.

5. Conclusions

In this work, MCR-ALS is applied to investigate major contamination patterns affecting river basins of a particular geographical region (in Catalonia, NorthEast Spain)

over several years of monitoring and analysis. Using MCR-ALS with non-negativity and with or without trilinearity constrains resulted to be an efficient tool to resolve the major contamination patterns explaining the measured data variance. Three major contamination patterns were detected, which were respectively related to agriculture activities, to industrial activities and to the chlorination treatment of drinking-water. Areas where these major contamination patterns were more relevant were then displayed using appropriate mapping tools.

An additional conclusion of this work is the demonstration of the data summarizing and interpretation possibilities obtained by the application of chemometric methods to large environmental data sets stored by official environmental agencies for their improved quality management and interpretation.

Acknowledgements

The authors acknowledge the following institutions for financial support: Agencia Española de Cooperación Internacional, Universidad Nacional de Rosario, CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica) and Catalan Water Agency (*l'Agència Catalana de l'Aigua*) for providing the data.

References

1 D. L. Massart, B. G. M. Vandeginste, L. M. C. Buydens, S. de Jong, P. J. Lewi, J. Smeyers-Verbeke, Handbook of chemometrics and qualimetrics. Elsevier, Amsterdam, 1998.

2 J. W. Einax, H. W. Zwanzinger, S. Geiss, Chemometrics in environmental chemistry, VCH, Weinham, 1997.

3 I. T. Jolliffe, Principal component analysis, Springer, New York, 2003.

4 S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst 2 (1987) 37-52.

5 M. Manz, K. D. Wenzel, U. Dietze, G. Schüürmann, Persistent organic pollutants in agricultural soils of central Germany, Sci. Total Environ. 277 (2001)187-198.

6 C. Backe, I. T. Cousins, P. Larsson, PCB in soils and estimated soil–air exchange fluxes of selected PCB congeners in the south of Sweden, Environ. Pollut. 128 (2004) 59-72.

7 S. P. Mujunen, P. Minkkinen, B. Holmbom, A. Oikari, PCA and PLS methods applied to ecotoxicological data: ecobalance project, J. Chemometrics 10 (1996) 411-424.

8 U. Dietze, T. Braunbeck, W. Honnen, H. R. Köhler, J. Schwaiger, H. Segner, Chemometric discrimination between streams based on chemical, limnological and biological data taken from freshwater fishes and their interrelationships, J. Aquat. Ecosyst. Stress Recovery 8 (2001) 319-336.

9 E. Peré-Trepat, M. Petrovic, D. Barceló, R. Tauler, Application of chemometric methods to the investigation of main microcontaminant sources of endocrine

disruptors in coastal and harbour waters and sediments, Anal. Bioanal. Chem. 378 (2004) 642-654.

10 R. Tauler, D. Barceló, E. M. Thurman, Multivariate correlation between concentrations of selected herbicides and derivatives in outflows from selected US midwestern reservoirs, Environ. Sci. Technol. 34 (2000) 3307-3314.

11 M. Terrado, D. Barceló, R. Tauler, Quality assessment of the multivariate curve resolution alternating least squares (MCR-ALS) method for the investigation of environmental pollution patterns, Environ. Sci. Technol. 43 (2009) 5321-5326.

12 R. Tauler, S. Lacorte, M. Guillamon, R. Cespedes, P. Viana, D. Barceló, Chemometric modeling of main contamination sources in surface waters of Portugal, Environ. Toxicol. Chem. 23 (2004) 565-575.

13 E. D. Malinowski, Factor analysis in chemistry, 3rd Ed., John Wiley & Sons, New York, 2002.

14 R. B. Cattell, The scientific use of factor analysis in behavioral and life sciences, Plenum, New York, 1978.

15 E. Eljarrat, J. Caixach, J. Rivera, M. de Torres, A. Ginebreda, Toxic potency assessment of non- and mono-ortho PCBs, PCDDs, PCDFs, and PAHs in northwest Mediterranean sediments (Catalonia Spain), Environ. Sci. Technol. 35 (2001) 3589-3594.

16 M. Calvo, J. Caixach, M. Guerra, M. Om, C. Planas, A. Ginebreda, Comparison of polychlorinated biphenyls (PCBs) and organochlorine pesticides concentration levels

in sediment and biota. Analysis by HRGC/HRMS, Organohalog. Compd. 57 (2002) 353-356.

17 D. Larrazabal, E. Eljarrat, B. Fabrellas, D. Barceló, Assessment of the toxic potency of PCDDs, PCDFs and PCBs in marine sediments from Catalonia Spain, Organohalog. Compd. 62 (2003) 148-151.

18 E. Teixidó, L. Olivella, M. Figueras, A. Ginebreda, R. Tauler, Multivariate exploratory data analysis of the organic micropollutants found in the Llobregat River (Catalonia, Spain), Int. J. Environ. Anal. Chem. 81 (2001) 295-313.

19 I. M. Fharnham, A. K. Singh, K. J. Stetzenbach, K. H. Lohannesson, Treatment of nondetects in multivariate analysis of groundwater geochemistry data, Chemom. Intell. Lab. Syst. 60 (2002) 265-281.

20 A. Smilde, R. Bro, P. Geladi, Multi-way Analysis with Applications in the Chemical Sciences, John Wiley & Sons Ltd., New York, 2004.

21 D. Comas, E. Ruiz, Fundamentos de los Sistemas de Información Geográfica, Arial Geografía, Barcelona, 1993.

22 J. Gutiérrez-Puebla, M. Gould, Sistemas de Información Geográfica, Síntesis, Madrid, 1994.

23 F. Calvete, J. Carrera, Geoestadística: Aplicaciones a la hidrología subterránea, Centro Internacional de Métodos Numéricos en Ingeniería, UPC, Barcelona, 1990.

24 N. Cressie, Statistics for Spatial Data, JohnWiley & Sons Inc., New York, 1993.

25 P. Goovaerts, Geostatistics for Natural Resources Evaluation, Oxford University Press, 1997.

26 J. Grimalt, L. Canton, J. Olive, Source input elucidation in polluted coastal systems by factor-analysis of sedimentary hydrocarbon data, Chemom. Intell. Lab. Syst. 18 (1993) 93-109.

27 V. Zitko, Principal component analysis in the evaluation of environmental data, Mar. Pollut. Bull. 28 (1994) 718-722.

28 R. Tauler, A. Smilde, B. Kowalski, Selectivity, local rank, 3-way data analysis and ambiguity in multivariate curve resolution, J. Chemometrics 9 (1995) 31-58.

29 R. Tauler, Multivariate Curve Resolution Applied to Second Order Data, Chemom. Intell. Lab. Syst. 30 (1995) 133-???.

30 R. Tauler, I. Marques, E. Casassas, Multivariate curve resolution applied to three-way trilinear data: study of a spectrofluorimetric acid-base titration of salicylic acid at three excitation wavelengths, J. Chemometrics 12 (1998) 55-75.

31 A. De Juan, R. Tauler, Comparison of three-way resolution methods for non-trilinear chemical data sets, J. Chemometrics 15 (2001) 749-???.

32 A. Smilde, R. Bro, P. Geladi, Multi-way Analysis, JohnWiley & Sons Ltd., Chichester, England, 2004.

33 D. S. Burdick, An introduction to tensor-products with applications to multiway data-analysis, Chemom. Intell. Lab. Syst. 28 (1995) 229-237.

34 H. A. L. Kiers, Towards a standardized notation and terminology in multiway analysis, J. Chemometrics 14 (2000) 105-122.

35 E.Teixidó, L.Olivella, M.Figueras, A.Ginebreda and R.Tauler. Multivariate exploratory data analysis of the organic micropollutants found in the Llobregat River (Catalonia, Spain), Intern.J. Environ. Anal. Chem. Vol 81 (2001) pp 295-313.

36 U.S. Environmental Protection Agency. 1984. Method 624- Purgeables. 40 CFR Part 136, 43373; Federal Register 49, No. 209.

37 U.S. Environmental Protection Agency. 1989. Method 503.1 - Revision 2.0. Volatile aromatic and unsaturated organic compounds in water by purge and trap gas chromatography. T.A. Bellar.

38 U.S. environmental Protection Agency Method 625. Guidelines Establishing Test Procedures for the Analysis of Pollutants Under the Clean Water Act: Final Rule and Interim Final Rule and Proposed Rule. Federal Register Vol 49, (209,Page 153-174).October 26,1984.

39 K. Abrahamsson and T. M. Xie. Direct determination of trace amounts of chlorophenols in fresh water, waste water and sea water. Journal of Chromatography, 279 (1983) 199-208.

40 H. Lee, L. Weng and A.S. Chau. Chemical derivatization analysis of pesticides residues. VIII. Analysis of 15 chlorophenols in natural water by in situ acetylation. J. Assoc. Off. Anal. Chem., Vol. 67, No. 4, 1984.

Table 1: Measured compounds and their descriptive statistics

Compound	Min	max	mean	median	Std. Dev.	% data
Ethylbenzene	0.25	9.4	0.3	0.25	0.6	2.6
m,p-Xylene	0.25	33.2	0	0.25	2	5.3
Toluene	0.3	644	0	0.3	40	5.6
Chlorpyrifos	5	1452	10	5	90	13.5
Diazinon	5	3894	0	16	200	62.4
Fenitrothion	5	201	0	5	20	7.3
Malathion	5	338	0	5	30	2.0
Acenaphthene	2	919	0	2	50	17.2
Acenaphtylene	2	255	0	2	20	8.9
Anthracene	2	34	2	2	2	3.0
Phenanthrene	2	245	10	6	30	64.0
Fluoranthene	2	16	3	2	2	23.1
Fluorene	2	201	0	2	20	33.3
Pyrene	2	52	4	2	5	35.3
PCL-Phenol	0.01	1.45	0.04	0.035	0.08	97.4
4,4'-DDD	0.5	141	1	0.5	8	4.0
4,4'-DDE	0.1	61.2	0	0.1	4	2.3
4,4'-DDT	0.5	152	2	0.5	10	3.3
α-Hexachlorocyclohexane	0.1	16675	0	0.1	1000	14.5
β-Hexachlorocyclohexane	0.5	2706	0	0.5	200	5.3
δ-Hexachlorocyclohexane	0.1	1679	0	0.1	100	3.0
Endosulfan I	0.1	544.7	0	0.1	30	12.5
Endosulfan II	0.5	273	0	0.5	20	9.2
Endosulfan sulfate	0.5	465	0	0.5	40	12.9
Hexachlorobenzene	0.1	74	1	0.1	6	8.3
Lindane (γ-hexachlorocyclohexane)	0.1	15308	0	5.8	1000	90.1
1,1,1-Trichloroethane	0.025	1.2	0	0.025	0.1	17.8
1,2-Dichloropropane	3	93	4	3.5	8	99.7
1,2,4-Trichlorobenzene	0.1	4.2	0.1	0.1	0.3	2.6
1,2-Ddichlorobenzene	0.25	12.3	0.3	0.25	0.7	1.3
Bromodichloromethane	0.025	3.12	0.1	0.025	0.2	34.3
Bromoform	0.05	6.71	0.1	0.05	0.4	21.8
Chloroform	0.025	8.6	0	0.025	1	41.6
Dibromochloromethane	0.025	6.39	0.1	0.025	0.4	34.3
Tetrachloroethylene	0.025	21.3	0	0.06	2	57.4
Carbon tetrachloride	0.025	0.94	0	0.025	0.1	19.1
Trichloroethylene	0.025	20	0	0.025	2	45.2

Figure captions

Figure 1: Map of Catalonia, Spain, showing the sampling locations and studied rivers.

Figure 2: PCA Loadings with MinMax of the log of the whole data set

Figure 3: MCR-ALS Loadings with MinMax of the log of the whole data set

Figure 4: Geographical location, superimposed on a map of Catalonia, Spain, of the spatial distribution of loadings of first MCR-ALS component with MinMax of the log of the whole data set. The scale of the contour lines is such that red corresponds to the maximum and blue to the minimum.

Figure 5: Same as Figure 3, corresponding to the second MCR-ALS component.

Figure 6: Same as Figure 3, corresponding to the third MCR-ALS component.

Figure 7: MCR-ALS results for the decomposition of the data matrix by imposing non-negativity and trilinearity restrictions. The colored bars indicate the relative intensities of the three identified components, and their distributions by sampling sites (red), variables or chemical compounds (blue) and campaigns (green).

Figure 1

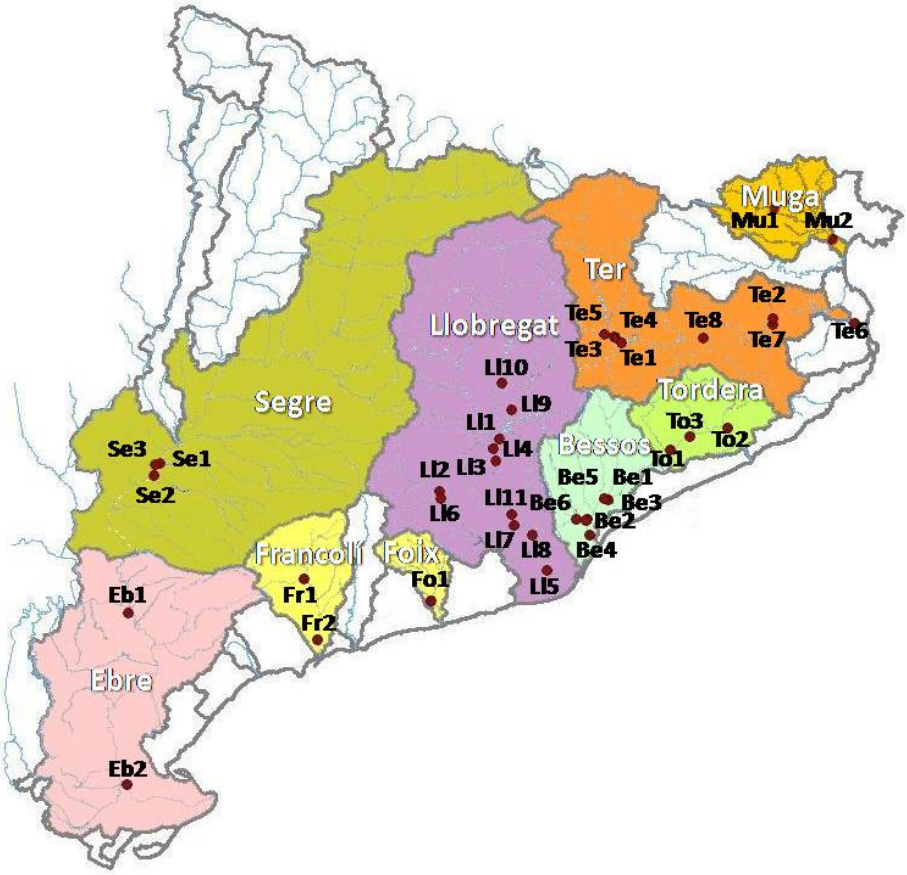

Muga
Mu1
Mu2
Ter
Te2
Te5
Te4
Te8
Te7
Te6
Te3
Te1
Llobregat
Ll10
Ll9
Segre
Tordera
Ll1
To3
To2
Ll4
Bessos
To1
Se3
Se1
Ll2
Ll3
Be5
Be1
Se2
Ll11
Be6
Be3
Ll6
Be2
Ll7
Ll8
Be4
Francolí
Foix
Fo1
Ll5
Eb1
Fr1
Fr2
Ebre
Eb2

Figure 2

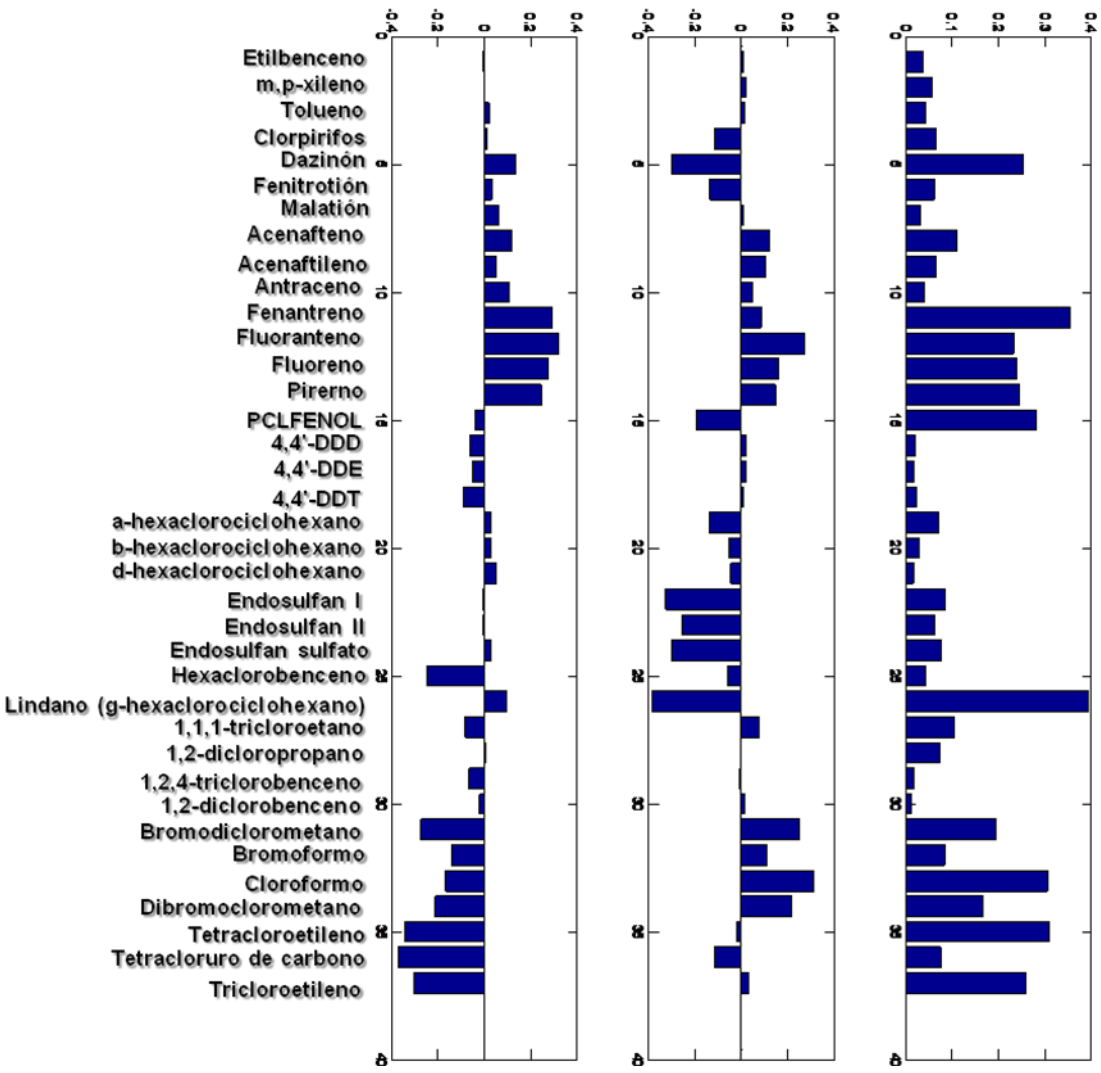

Etilbenceno
m,p-xileno
Tolueno
Clorpirifos
Dazinón
Fenitrotión
Malatión
Acenafteno
Acenaftileno
Antraceno
Fenantreno
Fluoranteno
Fluoreno
Pirerno
PCLFENOL
4,4'-DDD
4,4'-DDE
4,4'-DDT
a-hexaclorociclohexano
b-hexaclorociclohexano
d-hexaclorociclohexano
Endosulfan I
Endosulfan II
Endosulfan sulfato
Hexaclorobenceno
Lindano (g-hexaclorociclohexano)
1,1,1-tricloroetano
1,2-dicloropropano
1,2,4-triclorobenceno
1,2-diclorobenceno
Bromodiclorometano
Bromoformo
Cloroformo
Dibromoclorometano
Tetracloroetileno
Tetracloruro de carbono
Tricloroetileno

Figure 3

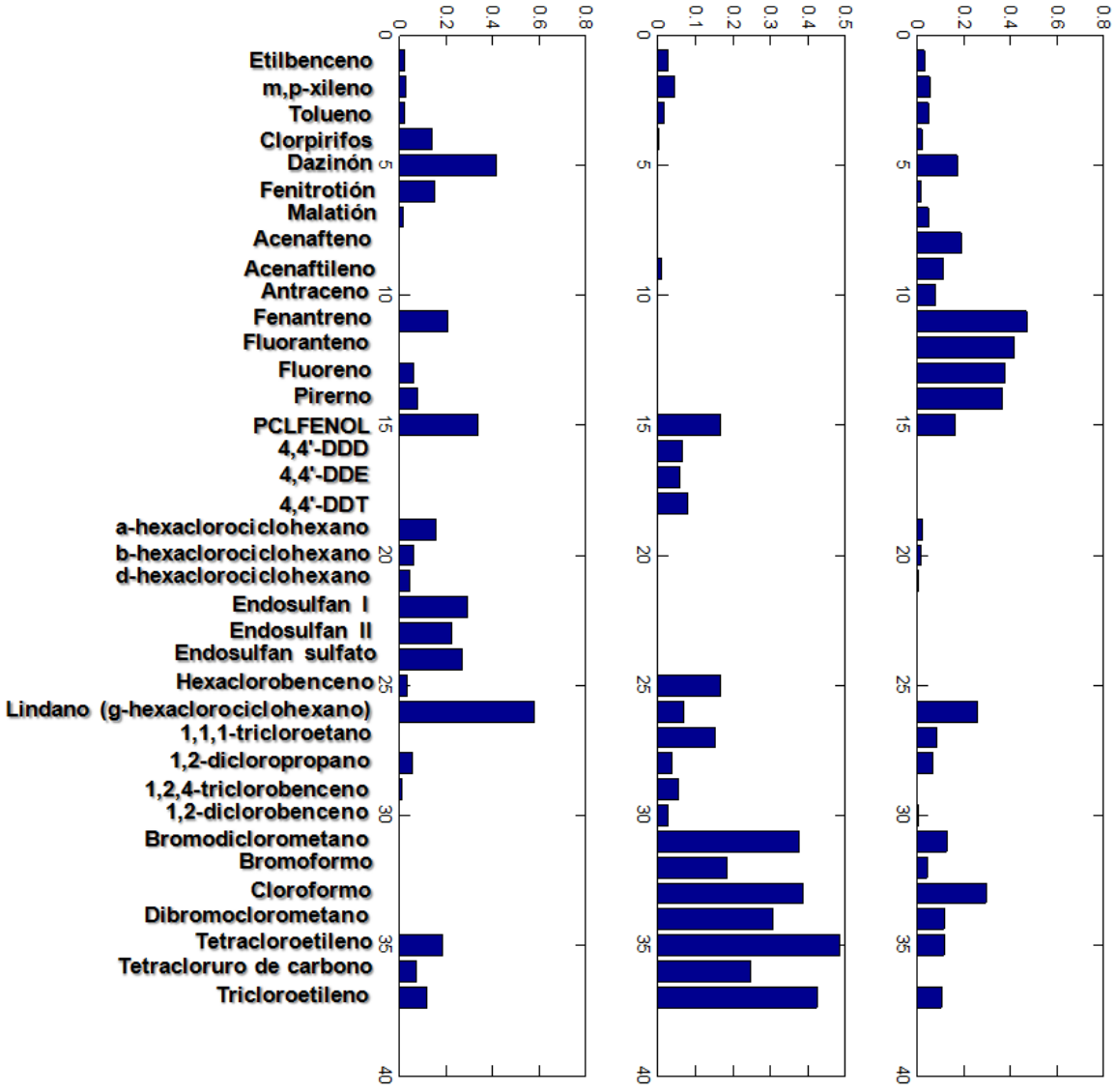

Etilbenceno
m,p-xileno
Tolueno
Clorpirifos
Dazinón
Fenitrotión
Malatión
Acenafteno
Acenaftileno
Antraceno
Fenantreno
Fluoranteno
Fluoreno
Pirerno
PCLFENOL
4,4'-DDD
4,4'-DDE
4,4'-DDT
a-hexaclorociclohexano
b-hexaclorociclohexano
d-hexaclorociclohexano
Endosulfan I
Endosulfan II
Endosulfan sulfato
Hexaclorobenceno
Lindano (g-hexaclorociclohexano)
1,1,1-tricloroetano
1,2-dicloropropano
1,2,4-triclorobenceno
1,2-diclorobenceno
Bromodiclorometano
Bromoformo
Cloroformo
Dibromoclorometano
Tetracloroetileno
Tetracloruro de carbono
Tricloroetileno

Figure 4

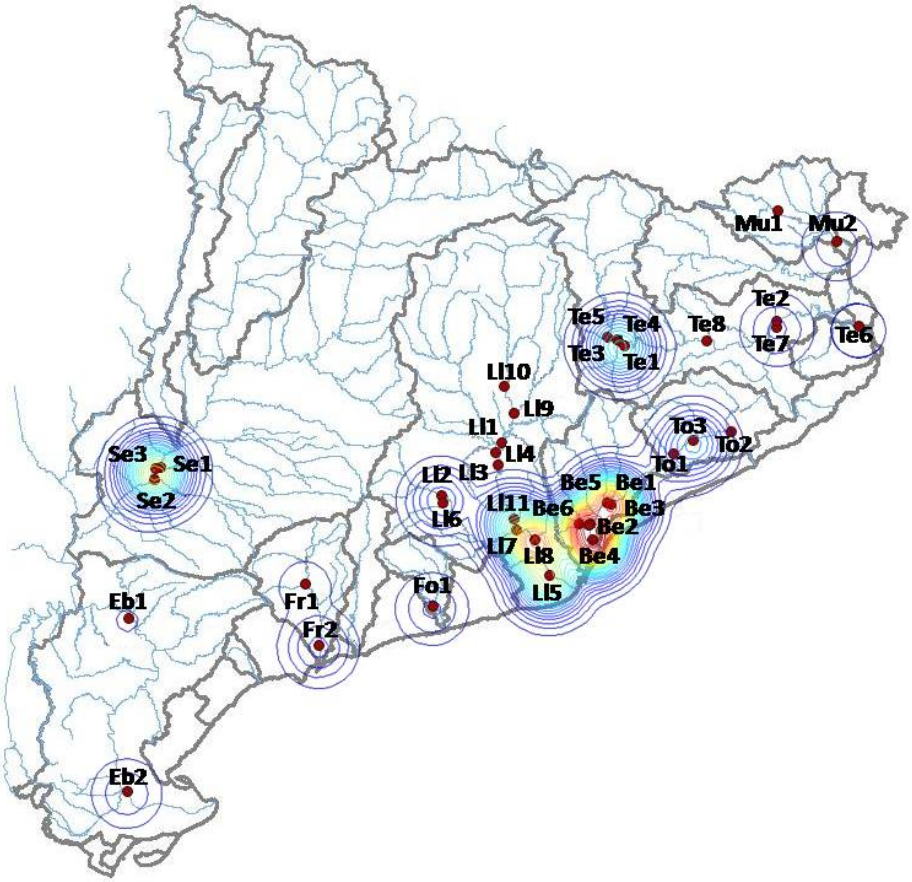

Mu1
Mu2
Te2
Te5
Te4
Te8
Te7
Te6
Te3
Te1
Ll10
Ll9
Ll1
To3
To2
Ll4
Se3
Se1
Ll2
Ll3
To1
Be5
Be1
Se2
Ll11
Be6
Be3
Ll6
Be2
Ll7
Ll8
Be4
Fo1
Eb1
Fr1
Ll5
Fr2
Eb2

Figure 5

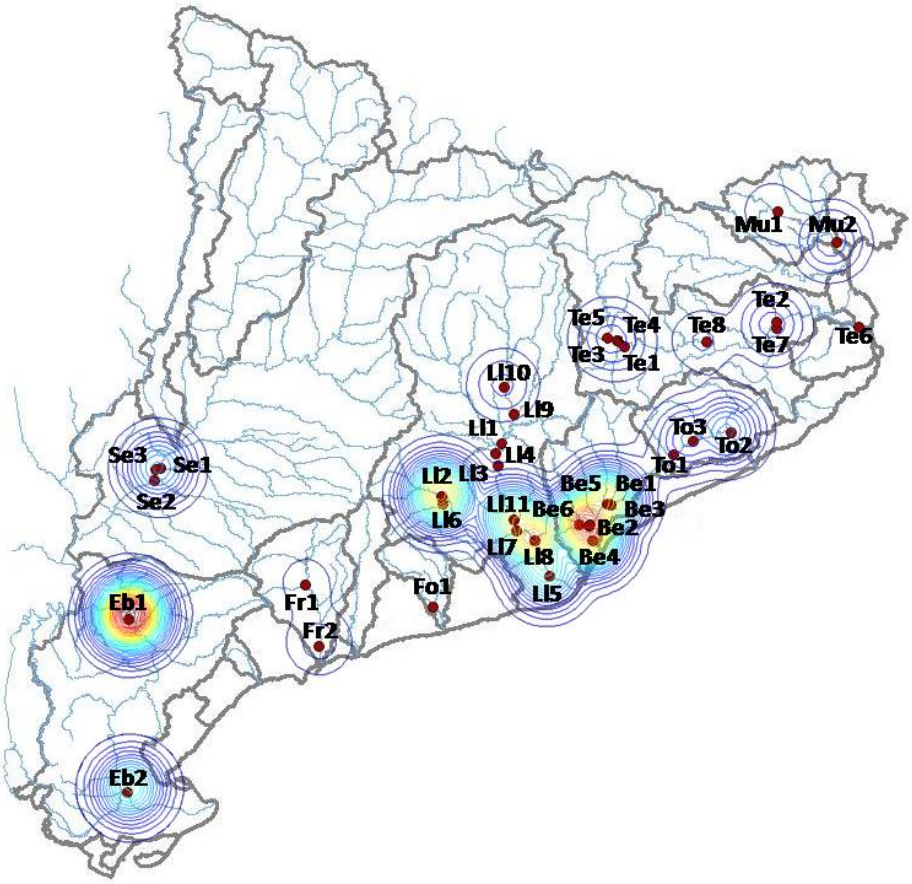

Mu1
Mu2
Te2
Te5
Te4
Te8
Te7
Te6
Te3
Te1
Ll10
Ll9
Ll1
Ll4
To3
To2
To1
Se3
Se1
Se2
Ll2
Ll3
Be5
Be1
Ll11
Be6
Be3
Ll6
Be2
Ll7
Ll8
Be4
Fo1
Ll5
Eb1
Fr1
Fr2
Eb2

Figure 6

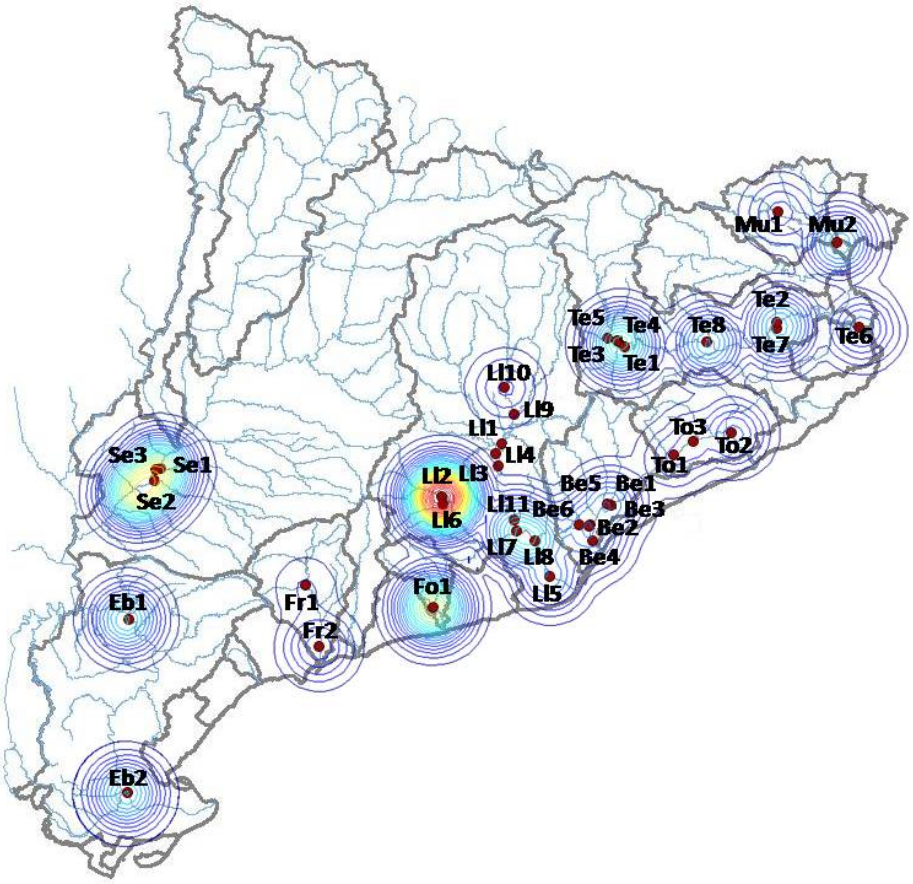
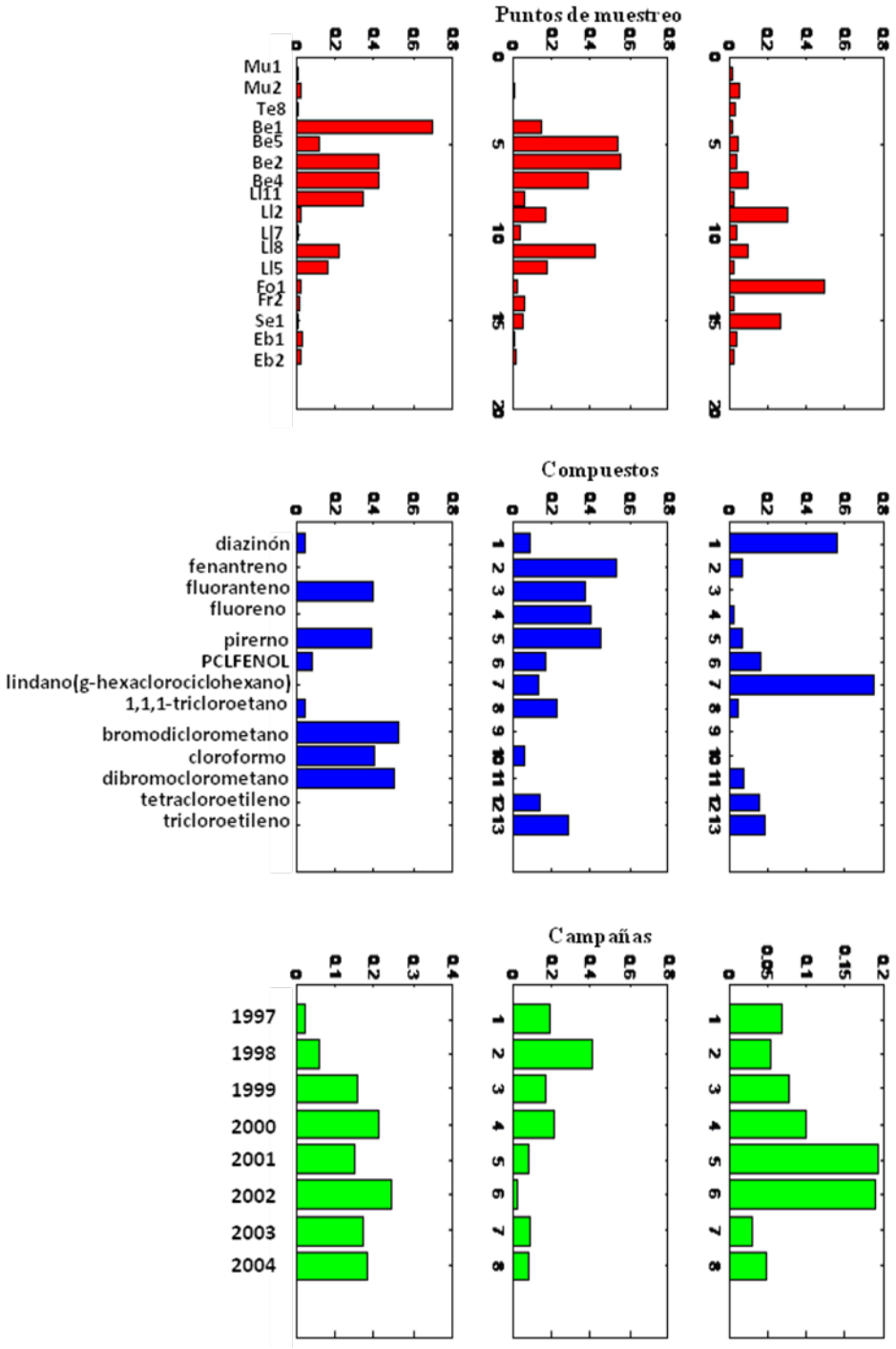

Mu1
Mu2
Te2
Te5
Te4
Te8
Te7
Te6
Te3
Te1
Ll10
Ll9
Ll1
To3
To2
Ll4
Se3
Se1
Ll2
Ll3
To1
Be5
Be1
Se2
Ll11
Be6
Be3
Ll6
Be2
Ll7
Ll8
Be4
Fo1
Eb1
Fr1
Ll5
Fr2
Eb2

Figure 7


Puntos de muestreo
Mu1
Mu2
Te8
Be1
Be5
Be2
Be4
Ll11
Ll2
Ll7
Ll8
Ll5
Fo1
Fr2
Se1
Eb1
Eb2
Compuestos
diazinón
fenantreno
fluoranteno
fluoreno
pirerno
PCLFENOL
lindano(g-hexaclorociclohexano)
1,1,1-tricloroetano
bromodiclorometano
cloroformo
dibromoclorometano
tetracloroetileno
tricloroetileno
Campañas
1997
1998
1999
2000
2001
2002
2003
2004