Unsupervised pattern-recognition techniques to investigate metal pollution in estuaries

A. Gredilla, S. Fdez-Ortiz de Vallejuelo, J.M. Amigo, A. de Diego, J.M. Madariaga

There has been a significant increase in the application of unsupervised pattern-recognition techniques to the analysis of long datasets emerging from the monitoring of metal pollution in estuaries. In this work, we thoroughly review the most important articles published on this topic in recent years.

Keywords: Artificial Neural Networks (ANN); Chemometrics; Cluster Analysis (CA); Estuary; Factor model; Metal pollution; Multivariate data; Pattern recognition; Principal Component Analysis (PCA); Unsupervised technique

A. Gredilla*,
S. Fdez-Ortiz de Vallejuelo,
A. de Diego,
J.M. Madariaga
Department of Analytical Chemistry, University of the Basque Country UPV/EHU, P.O. Box 644, 48080 Bilbao, Basque Country, Spain

J.M. Amigo
Department of Food Science, Quality and Technology, Faculty of Life Sciences, University of Copenhagen, DK-1958 Frederiksberg C, Denmark

*Corresponding author.
Tel.: +34 94 6015551;
E-mail: ainara.gredi@ehu.es

1. Introduction

In recent decades, chemometrics has grown significantly as a result of many advances in the field of analytical instrumentation and computer sciences, becoming a discipline, well-recognized within the field of chemistry. Since Svante Wold and Bruce R. Kowalski introduced the concept of chemometrics at the start of the 1970s, a number of definitions have been proposed for this field of analytical chemistry. It has been defined as the art or ability to extract relevant and meaningful information from data obtained through chemical analysis [1]. Most definitions underline the following advantages:

- real and current information can be obtained from data very quickly;
- meaningful, clear and precise information can be achieved from multidimensional datasets; and,
- it is inexpensive.

In short, using chemometrics, a lot of high-quality information is obtained quickly and cheaply.

In the field of environmental analytical chemistry, identifying relations between chemically characterized objects (samples) is a common problem. Pattern-recognition techniques acknowledge that, amongst different samples, there are groups that have similar characteristics and take these similarities as a basis to classify or to group samples [2]. Pattern-recognition techniques were first used in classifying documents, biometrics, financial forecasting and identifying languages [3]. All of these techniques share the same three basic steps:

- the application of mathematical methods to data;
- graphical representation of data; and,
- a decision about the object's classification.

This article looks in detail at pattern-recognition techniques, more specifically at unsupervised pattern-recognition methods applied to pollution monitoring in estuaries.

2. Pattern-recognition techniques

Pattern-recognition techniques are either supervised or unsupervised, depending on the information the analyst wishes to use or is available about samples that constitute the data matrix. In supervised techniques, also known as classification techniques, classes are defined beforehand because the rule or the characteristic used to group samples into a sub-set is known, whereas, in unsupervised techniques, classification is

0165-9936/$ - see front matter © 2013 Elsevier Ltd. All rights reserved. doi:http://dx.doi.org/10.1016/j.trac.2013.01.014

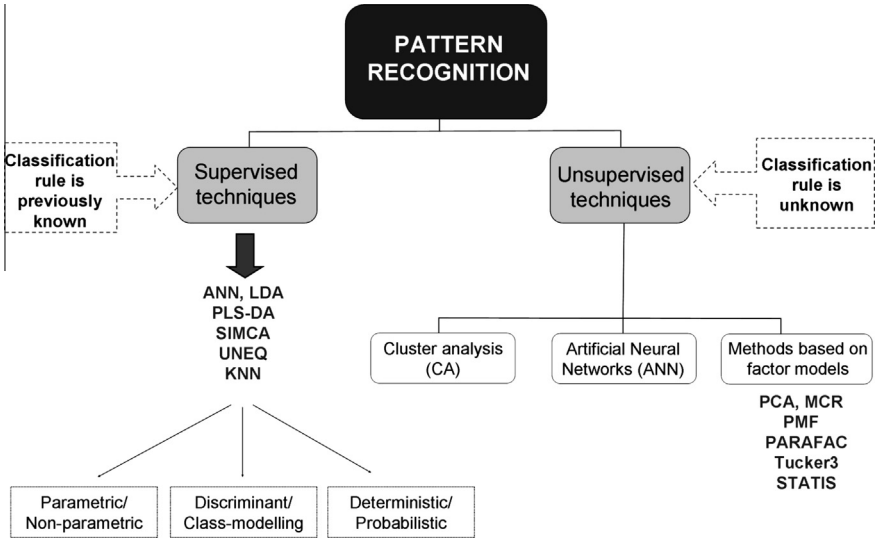

Figure 1. Classification of pattern-recognition techniques.

carried out taking into account only similarities and differences between samples, without using any additional prior information about them (see Table 1).

2.1. Supervised techniques

Supervised pattern-recognition techniques are used in the analysis of chemical data from different sources (e.g., chromatography, spectroscopy and sensor measurements). There are numerous supervised methods and they have been widely applied in the field of analytical chemistry [4]. In each case, the most appropriate technique depends on the problem at hand, as the bases and the criteria of the methods change significantly from one problem to another. As can be observed in Fig. 1, various criteria can be used to sort supervised techniques. Below are some of the most common ones.

2.1.1. Parametric and non-parametric techniques. Parametric techniques use mathematical models that have adjustable parameters to carry out sample classification. Such techniques include LDA (Linear Discriminant Analysis), SIMCA (Soft Independent Modeling of Class Analogy), UNEQ (UNEQual dispersed classes) and PLS-DA (Partial Least Squares Discriminant Analysis).

Non-parametric methods do not use any parameter based on a mathematical model to classify samples. Amongst the most used non-parametric methods are kNN (k-Nearest Neighbors), ANN (Artificial Neural Networks) and CAIMAN (Classification And Influence Matrix ANalysis).

2.1.2. Discriminant and class-modeling analysis. Supervised pattern-recognition techniques divide the hyperspace of variables that characterize samples into different classes. Using discriminant techniques, if a new sample is placed in one of the hyperspace classes, it belongs to that class, but, if it is placed outside, it does not. There is no middle ground or in between. These methods include LDA, kNN, PLS-DA and ANN.

In class-modeling analysis, samples fitting the model are considered part of the class, while objects that do not fit are rejected as non-members. When more than one class is modeled, three different situations can be distinguished (i.e. each sample can be assigned to a single category, to more than one category or to no category at all). SIMCA and UNEQ are amongst the most commonly used class-modeling techniques.

2.1.3. Deterministic/probabilistic methods. When a deterministic system is used to assign a class to each sample, no statement about the reliability of the decision is made. Probabilistic methods, in contrast, estimate the reliability of the classification. Deterministic methods include kNN and CAIMAN, and probabilistic methods include LDA, PLS-DA, SIMCA, UNEQ and ANN.

2.2. Unsupervised techniques

The rule for classifying samples is not often known – neither the number nor the identity of the classes. This is common in studies that carry out monitoring of pollution. In these cases, and some others, unsupervised pattern-recognition techniques are used. The methods usually used can be divided into three main groups (see Fig. 1), as follows.

2.2.1. Cluster Analysis (CA). Until a few years ago, this was the most widely used pattern-recognition method. This technique assigns samples to the same cluster on the basis of the degree of similarity among the variables (properties) that have been used to characterize the

Table 1. Summary of the publications that use unsupervised pattern-recognition techniques to analyze data from metal-pollution monitoring in estuarine sediments

Unsupervised technique used	Data input (sample x variable)	Metals considered and analytical method used	Main objective	Ref.
PCA, MCR-ALS, PARAFAC, Tucker 3	136 × 14 (PCA, MCR-ALS) 8 × 17 × 14 (PARAFC, Tucker3) (8 sites, 17 campaigns)	Al, As, Co, Cu, Cr, Cd, Fe, Mn, Mg, Ni, Pb, Sn, V, Zn (ICP-MS)	Practical comparison between two- and three-way methods. Spatial and temporal distribution of metals	[17]
PCA, MCR-ALS, PARAFAC Tucker3	51 × 11 (PCA, MCR) 17 × 3 × 11 (PARAFAC, Tucker3) (17 sites, 3 compartments)	As, Ba, Cd, Co, Cu, Cr, Fe, Mn, Ni, Pb, Zn (ICP-OES)	Identify metal sources and their spatial distribution	[22]
FA, CA	36 × 22 (12 sites, 3 compartments)	Cu, Zn, Pb, Cd, As, Cr (XRF)	Toxicological and sediment quality assessment	[24]
PCA	21 × 18 (7 sites, 3 compartments)	Ag, Al, As, Cd, Co, Cr, Fe, Mn, Ni, Pb, Zn (ICP-MS)	Toxicological assessment	[25]
PCA	36 × 13 (12 sites, 3 campaigns)	Cr, Cd, Pb, Ni, Zn, Cu, As (ICP-MS)	Identify metal sources	[26]
PCA	29 × 11	Al, Fe, Li, Mn, Ag, Cd, Co, Ni, Pb, Zn (AAS)	Identify metal sources and describe spatial distribution	[27]
PCA	56 × 7 (8 sites, 7 campaigns) 51 × 7 (2 cores)	Pb (ICP-MS)	Study the role of physico-chemical variables in the fate of Pb and identify metal sources	[28]
PCA	20 × 11 (5 sites, 4 seasons)	Al, Fe, Mn, Zn, Cu, Cr, Pb, Ni, Cd (AAS)	Toxicological and sediment quality assessment. Spatial and temporal distribution of metals and their sources	[29]
PCA	33 × 19	Cu, Fe, Pb, Zn, Na, K, Ca, Mg, Al (AAS)	Identify metal sources and their spatial distribution (tide effect). Evaluate anthropogenic impacts	[30]
PCA	32 × 24	Cd, Cr, Cu, Ni, Pb, Zn (AAS)	Identify metal sources and their spatial distribution	[32]
CA	10 × 9 (zone1) 7 × 9 (zone2)	Al, Fe, Ni, As, Cu, Pb, Zn, Cd, Cr (ICP-MS)	Identify metal sources and evaluate anthropogenic impacts	[33]
FA, CA	39 × 21	Cd, Zn, Pb, Cr, Fe, Mn, Al, Ni, Cu (ICP-OES)	Demonstrate the effectiveness of a multivariate analysis technique. Identify metal sources	[34]
CA	300 × 6 (5 cores)	Cd, Pb, Zn (ICP-MS)	Identify metal sources and their spatial distribution. Evaluate anthropogenic impacts	[35]
PCA	15 × 7 (zone1) 15 × 7 (zone2) 9 × 7 (zone3)	Pb, Cu, Co, Cr, Cd, Zn (AAS)	Define background values and identify metal sources	[36]
CA	210 × 10 (15 sites, 14 campaigns)	As (AAS). Al, Fe, Cr, Cu, Ni, Mn, Pb, Zn, (FAAS). Hg (AAS)	Spatial and temporal distribution of metals and evaluate anthropogenic impacts	[37]
HCA	27 × 19 (Filtered) 5 × 8 (Centrifuged) 12 × 10 (Superficial)	Mg, Al, K, Ca, Ti, Mn, Fe (EPXMA)	Identify metal sources	[38]
CA, FA	31 × 9 (CA) 31 x 32 (PCA)	Fe, Mn, Zn, Cu, Co, Cr, Ni, Pb, Cd (AAS)	Sediment-quality assessment and identification of metal sources	[39]
CA, PCA	49 × 11 (14 samples in zone1, 20 in zone2 and 20 samples in zone3)	Cu, Fe, Mn, Zn, Cr, Co, Ni, Pb, Ba (ICP-OES). Al (EDXRF)	Identify metal sources and their spatial distribution. Evaluate anthropogenic impacts	[40]
PCA	16 × 13	Cd (ICP-MS). Co, Cr, Cu, Fe, Hg, Mn, Ni, Pb, Zn (FS-FAAS)	Sediment-quality assessment, spatial distribution of metals and identification of their possible origin	[43]
CA, FA	20 × 13 (Pre-monsoon) 20 × 13 (Post-monsoon)	Fe, Mn, Cr, Cu, Ni, Co, Pb, Zn, Cd (GF-AAS)	Identify metal sources and their spatial distribution	[44]
CA	8 × 10	Cu, Ni, Pb, Co, Zn, Mn, Fe, Al, Ca (NS)	Identify metal sources and their spatial distribution. Evaluate anthropogenic impacts	[45]
PCA	116 × 14 (8 sites, 12 campaigns)	Al, As, Cd, Co, Cr, Cu, Fe, Mg, Mn, Ni, Pb, Sn, V, Zn (ICP-MS)	Spatial and temporal distribution of metals and identification of their sources. Sediment-quality assessment and evaluate anthropogenic impacts	[46]
HCA, PCA	12 × 10 (3 sites, 4 seasons)	Cd, Cr, Cu, Fe, Pb, Mn, Ni, Zn (ETAAS, FAAS)	Toxicological assessment. Identify metal sources and study spatial distribution of metals	[47]

(*continued on next page*)

Table 1 (*continued*)

Unsupervised technique used	Data input (sample x variable)	Metals considered and analytical method used	Main objective	Ref.
PCA	7 × 23	Fe, Mn, Zn, Cu (FAAS). Hg, As (MHS-FIAS). Pb, Cd, Ag, Sn, V, Ni (GF-AAS)	Toxicological assessment and identify metal sources	[48]
PCA	66 × 16	Al, Fe, Ti, Mn, Cu, Pb, Cr, Zn, Co, As, Ni, Cd, Sr (ICP-OES)	Spatial distribution of metals, study their distribution patterns and evaluate anthropogenic impacts	[49]
PCA	117 × 16 (6 cores)	Cd, Cr, Co, Cu, Ni, Pb, Zn, Mn (ICP-OES)	Study spatial distribution of metals in cores	[50]
PCA	8 × 12 (zone1) 6 × 12 (zone2)	Zn, Cd, Pb, Cu, Ni, Co, V (DPASV-HMDE)	Toxicological and sediment-quality assessment, study spatial and temporal distribution of metals and their sources	[51]
HCA	16 × 15 (core1) 16 × 15 (core2)	Pb, Zn, Cu, Cr, V, Mn, Cd, Co, Sb, Sn Ag, Mo (ICP-MS)	Identify metal sources and sediment quality assessment	[52]
HCA, PCA	36 × 6	Mn, Fe, Ni, Cu, Zn, Pb (NS)	Identify metal sources and evaluate anthropogenic impacts	[53]
HCA, PCA	59 × 6 (23 samples in zone1 and 36 in zone2)	Pb, Cd, Cu, Zn, Ni (AAS). Hg (MHS-FIAS)	Sediment-quality assessment and spatial distribution of metals	[54]
FA, HCA	60 × 17 (30 samples, 2 seasons)	Pb, Ni, Cu, Co, Cd, Cr, Fe (AAS)	Identify metal sources and their spatial distribution. Evaluate anthropogenic impacts	[55]
PMF	34 × 18	Al, As, Cd, Co, Cr, Cu, Fe, Mg, Mn, Ni, Pb, Sn, V, Zn. (ICP-OES)	Identify metal sources	[57]
PCA	24 × 6 (October) 24 × 6 (February)	Fe, Cu, Zn, Cd, Pb, Cr (AAS)	Spatial distribution of metals	[58]
PCA, CA	235 × 15 (7 sites, 4 cores)	Li, Mg, Al, Cr, Mn, Fe, Ni, Cu, Zn, (FAAS). Na, K, Ca (AES). Cd (FAAS). Hg (MHS-FIAS)	Spatial distribution of metals	[60]
CA	60 × 16 (6 cores)	Fe, Mn, Cr, Cu, Ni, Pb, Cd, Mo, Ag, As, Ba (ICP-MS)	Sediment quality assessment, evaluation of anthropogenic impacts and the mobility of metals	[62]
PCA, k-NN, SIMCA, LDA, ANNs	95 × 14 (8 sites, 12 campaigns)	Al, As, Cd, Co, Cr, Cu, Fe, Mg, Mn, Ni, Pb, Sn, V, Zn (ICP-MS)	Spatial and temporal distribution of metals	[63]
PARAFAC	7 × 5 × 4	Cd, Cr, Cu, Pb, Zn (ICP-OES)	Spatial distribution of metals	[64]
PCA	15 × 20	Ag, Cd, Cu, Cr, Ni, Pb, Zn, As, Se (FAAS). Cu, Cr, Ni, Pb, Zn, (GF-AAS). Ag, Cd, (HG-AAS). Hg (CV-AAS)	Toxicological and sediment quality assessment.	[65]
HCA	116 × 14 (8 sites, 12 campaigns)	Al, As, Cd, Co, Cr, Cu, Fe, Mg, Mn, Ni, Pb, Sn, V, Zn (ICP-MS)	Clustering of samples prior to PLS modeling	[67]
PCA PLS-DA LDA QDA CVA ECVA CART CP-ANNS	3048 × 16	As, Cd, Cr, Cu, Pb, Hg, Ni, Ag, Zn (NS)	Exploratory analysis prior to build predictive models	[69]
HCA	116 × 14 (8 sites, 12 campaigns)	Al, As, Cd, Co, Cr, Cu, Fe, Mg, Mn, Ni, Pb, Sn, V, Zn (ICP-MS)	Clustering of samples prior to PLS modeling	[70]
SOM	27 × 48 (9 sites)	Cu, Mn, Ni, Cr, Pb, Zn (AAS)	Identify associations between metal content and sediment phases	[71]
PARAFAC	7 × 5 × 7	Cd, Cr, Cu, Pb, Zn (ICP-OES)	Identify metal sources and their spatial distribution	[72]
CA, PCA	20 × 27 (PCA) 20 × 15 (CA)	Be, Sc, V, Cr, Co, Ni, Cu, Zn, As, Y, Zr, Mo, Ag, Ba, Pb (ICP-MS)	Define background values	[74]

AAS, Atomic absorption spectroscopy; AES, Atomic emission spectroscopy; CV-AAS, Cold-vapor atomic absorption spectroscopy; DPASV, Differential pulse anodic stripping voltammetry; EDXRF, Energy-dispersive X-ray fluorescence; EPXMA, Electron-probe X-ray microanalysis; ETAAS, Electrothermal atomic absorption spectroscopy; FAAS, Flame atomic absorption spectroscopy; FS-FAAS, Fast sequential flame atomic absorption spectrometry; GF-AAS, Graphite-furnace atomic absorption spectroscopy; HG-AAS, Hydride-generation atomic absorption spectroscopy; HMDE, Hanging mercury dropping electrode; ICP-MS, Inductively-coupled plasma mass spectroscopy; ICP-OES, Inductively-coupled plasma optical emission spectroscopy; MHS-FIAS, Mercury-hydride system flow-injection atomic spectroscopy; XRF, X-ray fluorescence; NS, Not specified.Abbreviation of analytical methods that do not appear in the text;
CVA, Canonical variate analysis; CART, Classification and regression trees; CP-ANNS, Counter-propagation artificial neural networks; ECVA, Extended canonical variate analysis; QDA, Quadratic discriminant analysis.

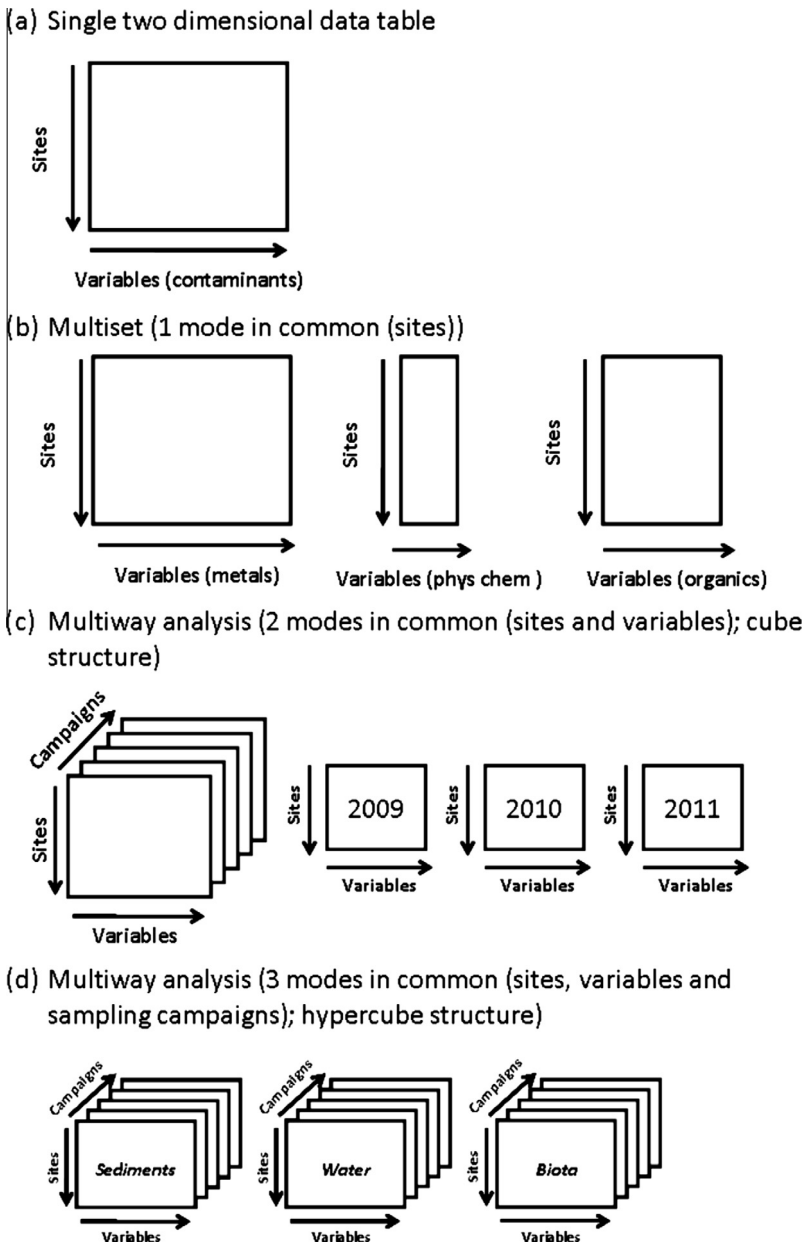

Figure 2. Different ways to organize tables of environmental data. In this example, several compartments (sediment, water, and biota) have been sampled in different campaigns (2009, 2010 and 2011) at different points (sites), and several variables have been measured in samples (concentration of metals, organics and physico-chemical parameters).

objects, and, at the same time, assigns samples that are not similar to different clusters. It is normally used to develop a new classification of the samples under study, but it can also be applied to confirm an already known grouping. Massart and Kaufman [5] published a comprehensive monograph on cluster analysis of analytical chemical data.

2.2.2. Artificial Neural Networks (ANN). ANN are mathematical techniques that simulate the nervous system in human beings, creating models for pattern recognition. They are usually very effective at overcoming the difficulties frequently encountered in the classification process. ANN starts from a so-called data-training set, which usually has properties (e.g., concentration levels or spectra) measured in different samples, to calculate the probability of samples being a member of a class (output variables). ANN can be used in either supervised or unsupervised pattern recognition, but, as their application is less simple than CA, their uses are still limited. Further information on the application of ANN in chemistry can be found elsewhere [6].

2.2.3. Methods based on factor models. The objective of these methods is to limit *n*-dimensional information about objects to a reduced, more representative dimension. In this manner, each sample can be graphically depicted in a two- or three-dimensional (2D or 3D) space, making it easier to identify the main tendencies. Data can be structured in different ways (see Fig. 2). If samples are organized in one direction and variables in another, a single 2D data table is obtained (e.g., the variables might be spectral or concentration levels of sample components).

Methods that handle data organized in this manner are known as two-way methods. Data can also be organized according to a multi-set or multi-way arrangement. In both cases, a simultaneous analysis of several tables of 2D data is performed.

In the multi-set approach, the data tables to be handled simultaneously have one mode in common and an irregular structure and/or meaning [7].

In multi-way analysis, several 2D data tables with several modes in common and a cube or hypercube structure are considered.

PCA (Principal Component Analysis), MCR (Multivariate Curve Resolution) and Positive Matrix Factorization (PMF) are, amongst others, the most popular techniques to deal with 2D data tables and multi-set approaches. As a matter of fact, the multi-set approach has been adopted by Environmental Protection Agency (EPA) for adequate air-quality management [8].

The papers by Wold et al. [9], Rutan et al. [10] and Paatero et al. [11] illustrate the basics and the applications of these three techniques.

The analysis of multi-way data has been most usually attempted by PARAFAC (PARallelFACtor Analysis), Tucker3 or models based on STATIS (*Structuration des Tableaux A Trois Indices de la Statistique*). Acar and coworkers published a literature survey on these techniques [12].

3. Application of unsupervised pattern-recognition techniques to the study of metal pollution in estuaries

Since the United Nations Conference on the Human Environment in Stockholm in 1972, there has been a clear recognition of the importance of dealing with environmental issues that has resulted in a growing number of environmental institutions, researchers and strict regulations [13]. Consequently, the use of unsupervised pattern-recognition techniques in the field of environmental analysis has also significantly increased. Specifically, environmental quality assessment has been a topic of great concern [8]. Certainly, pattern recognition has been more widely applied to soil and/or sediment-related problems. A good summary of the work done concerning soil pollution was recently published by Mostert [14].

There are many studies that investigate the geographical distribution of contaminants in soils and sediments using pattern-recognition techniques [15–17]. Water and air pollution has received less attention [18–20], followed at a distance by works related to biota [21,22]. Regarding chemicals, both organic and inorganic compounds have been considered independently or simultaneously [14].

In the case of water bodies, pattern-recognition techniques have been used increasingly since the introduction of the European Water Framework Directive (WFD, 2000/60/EC) in 2000. The monitoring carried out in order to reach the objectives set out in the WFD generates a large amount of data containing numerous variables that need to be interpreted.

In monitoring studies, many sampling campaigns, sample types and areas are taken into account at the same time, generating datasets of great complexity and variability. Since they contain correlated information, it makes no sense to use univariate models to analyze these datasets, as a lot of information would be lost in the process. This is particularly true in estuaries. These transitional areas are considered to be one of the most productive ecosystems. They are usually densely-populated zones with important industrial, agricultural and recreational activities, susceptible to produce large amounts of pollutants [23]. Among pollutants, metals are of especial interest, due to their high toxicity and persistence in the environment.

Although the study of different estuarine compartments at the same time is becoming more common [22,24–26], most of the studies are still focused on the analysis of a single matrix. Considering that water samples on their own merely give current information about a specific space or moment in time, it is preferable to study sediments. Their capacity to act as accumulators of metals and to become potential sources of pollution contributes in this sense. The following is a summary of the most important works that use unsupervised pattern-recognition techniques to investigate metal pollution in estuarine sediments (see Table 1).

First, we carried out a bibliometric analysis to investigate the evolution of the number of publications concerning unsupervised pattern-recognition techniques applied to the study of metal pollution in estuaries between 1980 and 2011. The on-line version of SciFinder was used in this survey. This research tool integrates the databases included in CAplus and MEDLINE. In a first step, the databases were searched with the concepts ''clustering'', ''unsupervised pattern recognition'', ''unsupervised classification'', ''cluster analysis'', ''PCA'', ''MCR'', ''PARAFAC'' and ''Tucker3'' to retrieve all the documents (books, reviews and articles) related to ''pollution'' studies. Nearly 7200 documents were obtained.

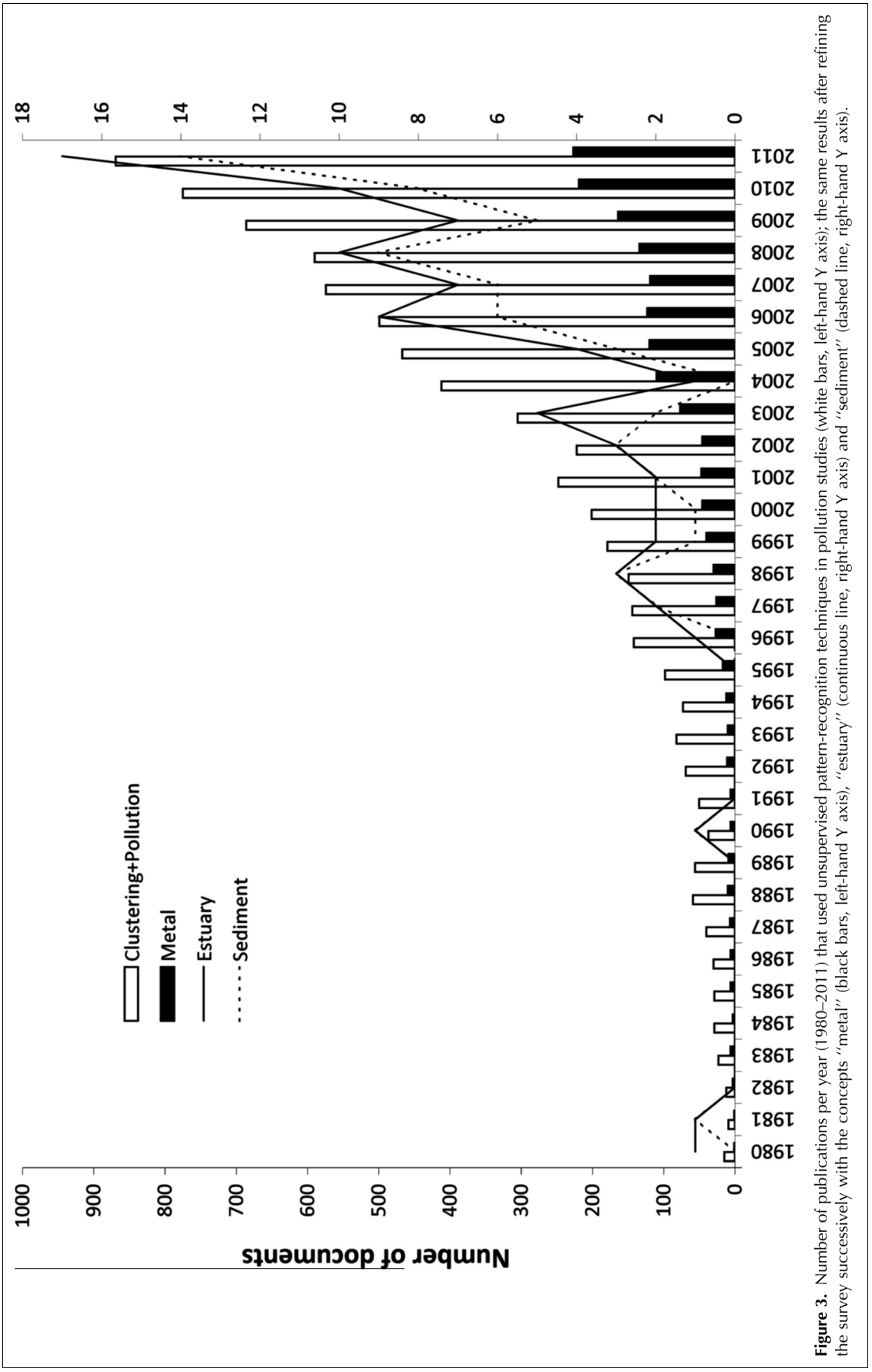

Figure 3. Number of publications per year (1980–2011) that used unsupervised pattern-recognition techniques in pollution studies (white bars, left-hand Y axis); the same results after refining the survey successively with the concepts ''metal'' (black bars, left-hand Y axis), ''estuary'' (continuous line, right-hand Y axis) and ''sediment'' (dashed line, right-hand Y axis).

This first result was refined using successively the keywords ''metal'' (nearly 1700 documents), ''estuary'' (88 documents) and ''sediment'' (67 documents). Only 57 of those 67 final references were written in English. The evolution (1980–2000) in the number of documents published per year in the above-mentioned four categories is depicted in Fig. 3.

Identification of pollution sources is one of the objectives most frequently pursued. In this line of work, PCA and CA predominate [26–51]. Nevertheless, both Hierarchical Cluster Analysis (HCA) [38,52–55] and Factor Analysis (FA) [34,39,44,55] have also been used. Pollution episodes of anthropogenic [31,54] or geogenic origin [32,56] are clearly distinguished. In most of the cases, agricultural and industrial activities are responsible for pollution of anthropogenic origin [26,27].

CA has also been used to identify the origin of pollution in a wide area by the simultaneous analysis of data obtained from sediments of different estuaries located in it [35]. The information obtained in this case was useful to develop effective management strategies to control pollution in the vast estuarine area investigated.

Recently, PMF was used with a similar objective [57]. In this work, the results obtained by PMF were compared with those obtained by FA.

In other work, analytical data obtained after sequential extraction were analyzed by FA and CA not only to define the origin of pollution, but also to connect this information with the evaluation of risk assessment using sediment-quality guidelines [39].

Classification of the sampling sites according to metal content in sediments is another important research area. The aim in this case is to obtain more information about the spatial distribution of chemicals along the estuary. PCA and CA [27,29,30,32,35,37,40–47,49–52,54,58–63] are again the methods preferred to carry on this task, although HCA [47,54,55] and PARAFAC [64] have also been used. In this last work, the PARAFAC model was also useful to visualize and to interpret the distribution of the five heavy metals (Cd, Cr, Cu, Pb and Zn) in different sediment fractions and sampling points of an estuary in Argentina. In most of the cases, the sampling sites were simply classified according to their metal content [29,32,64,65].

The influence of a specific contamination source [37] and the tidal effect [30] on the distribution of metals have been investigated in detail by CA and PCA, respectively. Chatterjee et al. used PCA and CA to study concentration gradients in sediment cores, and to identify geochemical factors responsible of the spatial distribution of trace elements in the sediment profiles [40].

The study of spatial distribution of metals in specific estuaries has been a very active research line. In traditional monitoring studies, both spatial and temporal variations are considered simultaneously. However, when unsupervised pattern-recognition techniques are applied to analyze data, only the spatial distribution of pollution is usually investigated [59,66]. Simultaneous consideration of both trends is scarce [17,26,28,37,46,63,67], and only in a few cases is the seasonal variability the issue of interest [44,47,55,58].

Unsupervised techniques have been also frequently used in combination with certain parameters [e.g., enrichment factor (EF), ecological risk index (E_{RI}), pollution-load index (PLI) or the geoaccumulation index (I_{geo})] to evaluate the magnitude of anthropogenic impacts in selected areas [30,33,35,37,40,41,45,46,53,55,62,65,66]. The efficiency of a newly developed pollution index (I_{poll}) was studied by CA and other chemometric tools [42]. Sediment-quality guidelines (SQG) are frequently used together with pattern-recognition techniques to estimate the toxicity of sediments according to their metal content [24,29,39,43,46,51,54,62,65]. However, studies based in *in-vivo* toxicological tests have become more popular in recent years. For example, PCA was used to identify the ranges in which chemical concentrations may result in adverse effects on local living organisms [48]. In this case, particularly significant correlations between chemical concentrations in sediment and biological effects were identified. The relation between certain physico-chemical parameters of mangrove sediments (including metal concentration) and the abundance and the diversity of ammonia-oxidizing archaea and ammonia-oxidizing beta-protobacteria has been studied using unsupervised pattern-recognition techniques [68]. Moreover, the changes observed in microbial activity under different degrees of metal contamination have been investigated by multidimensional scaling (MDS), HCA and PCA [47]. Data from physical, chemical and biological sources have also been combined to estimate sediment quality after data analysis by different unsupervised techniques [24]. Wepener et al. [25] reported the results of a multivariate statistical analysis (including unsupervised techniques) of chemical data (metal concentrations) from water, sediment and mussels collected in the Scheldt Estuary and biomarker responses in resident mollusc population. A similar approach was described by Mucha et al. concerning macrobenthic communities and sediments from the Douro Estuary [29].

PCA and CA have also been used to classify sediments before using supervised techniques [e.g., SIMCA, ANN, LDA, QDA (quadratic discriminant analysis), or PLS-DA] to define classification models based on metal concentration [63,69]. In a similar approach, clustering by HCA was the first step in a process to define PLS models able to predict metal concentrations in sediments from XRF [67] and NIR/MIR [70] spectral measurements. A new Kohonen unsupervised ANN approach, the so-called Self Organizing Maps (SOM), has also been applied to assess metal contamination in dredged sediments using sequential extraction [71]. In this work, the

authors concluded that certain metals (e.g., Cu, Pb and Zn) are of special interest due to their capacity to be remobilized and to associate with organic matter.

In addition, multi-way analysis of data from sequential extractions of sediments has been carried out by PARAFAC to investigate the mobility and the availability of metals (e.g., Cd, Cr, Cu, Pb and Zn) [64,72]. The pretreatment technique applied to a dataset clearly influences the result obtained by PCA, as evidenced by Reid and Spencer [73]. Furthermore, the same dataset (concentrations of 14 elements in more than 136 sediment samples) was analyzed by PCA, MCR, PARAFAC and Tucker3, and the results were critically compared [17]. Although the results observed after the use of the different chemometric methods varied slightly, the main conclusions extracted from them were similar.

The evaluation of geochemical processes that control the presence and the spatial distribution of metals in sediments was also carried out by FA and HCA [55]. Specifically, the loadings of FA demonstrated that the variability in metal concentration in the Hugli Estuary (India) was mainly governed by sediment properties. It was therefore concluded that physico-chemical properties (e.g., texture and organic matter) play a critical role in the sorption and the complexation of transition metals.

With a similar objective, the physico-chemical properties of sediments (e.g., total organic carbon and carbonate content) were included in the statistical analysis of the metal-concentration data of sediments collected in different estuaries [43,44,62].

Analogously, geochemical processes affecting the mobility of metals from sediments to water and *vice versa* have been evaluated by CA and correlation analysis [42].

In addition, different multivariate-analysis techniques (i.e. FA, CA and Canonical Discriminant Analysis) were simultaneously considered to evaluate spatial variations and identify pollution sources, connecting physico-chemical properties with natural biogeochemical processes [34].

Finally, several works estimated the background concentrations of metals in estuarine areas using analysis of data by PCA [36] and PCA and CA [74].

4. Conclusions

It is obvious from Fig. 2 that there is increasing interest in using sediments in the application of unsupervised pattern-recognition techniques to the analysis of metal pollution in estuaries. It may be thought that the increase in the number of articles published on this topic during the past few years is concomitant with the general trend observed in most branches of science. However, the publication rate in other related, but more general, fields has reached a plateau (''chemometrics'') or even shown a slight decrease (''environmental analysis'') in the past few years. This is evidence that unsupervised pattern-recognition techniques applied to pollution studies is certainly a hot topic in both absolute and relative terms.

However, it is also true that the average number of documents published in the area per year is still low. The use of chemometrics by researchers involved in environmental chemistry issues is limited, probably due to

- in some cases, ignorance of even the existence of these techniques and, of course, of their potential;
- a kind of inertia that involves doing things as they have been done in the past (i.e. simple univariate and graphical analysis of data); and,
- to some extent, a veiled fear of the mathematics involved in multivariate-analysis techniques.

However, the potential advantages are important, as the only correct way to approach multivariate problems is indeed using multivariate techniques. Multivariate techniques are necessary because the whole pollution scenario needs to be described by several variables and requires tools that take into account the relations among them. Popularization of these techniques among researchers involved in environmental issues is certainly a challenge for the future.

In any case, the works published so far clearly indicate that unsupervised pattern-recognition techniques constitute an appropriate tool for pattern recognition in environmental data. Undoubtedly, the most popular techniques to analyze single tables of data, whatever the objective of the study, are PCA and CA. The current predominance of these techniques may be due to different aspects:

- the application of multi-way and multi-set methods may turn out more difficult due to the complexity of the structure of data and to the interpretation of the results obtained being harder (as is also applicable to ANN);
- PCA and CA are the techniques that were traditionally and systematically selected for unsupervised pattern recognition, so that researchers suffer again from a kind of inertia in this field.

In most cases, PCA and/or CA are applied to the analysis of single tables of 2D data. However, the analysis of augmented data in the form of multi-set or multi-way structures has gained popularity in recent years. Although simultaneous handling of several single tables of data may lead to formal and conceptual difficulties, this approach provides us with a powerful key to interpret the system under study. Moreover, the augmentation of the data matrix, including sampling at different seasons and/or in different compartments, informs us about not only the past and current situation of the estuary, but also its most probable immediate future. Thus, when the problem in hand deals with several data

sets that refer to variables with different units, and we aim to quantify to what extent the structural relations between variables are the same across the data sets, the use of the multi-set or multi-way approach is the most effective. Separate analysis of each individual data set by PCA or similar and further comparison of the obtained loadings may also be attempted, but this approach will be indeed less straightforward.

It is extremely important to define the objective of the study first and to structure data according to that objective. The selection of the most appropriate technique(s) in each case relies to a great extent on a correct definition of the objective and a wise organization of data. Despite a single technique often giving useful information about metal pollution in estuaries, the combined use of two or more techniques may be helpful to interpret the results more adequately or to confirm findings obtained with other techniques. Moreover, complementary information is usually obtained when different techniques are considered simultaneously.

Finally, it should be underlined that the application of unsupervised pattern-recognition techniques has been a turning point in the assessment of ecological quality in estuaries, which is a key issue in development, implementation and enforcement of any regulation in the framework of aquatic ecosystems.

Acknowledgements

This work has been financially supported by the SUDOE Interreg IV B programme through the ORQUE SUDOE (Ref. SOE3/P2/F591/5) project and the UFI program of the UPV/EHU through the Global Change and Heritage project (Ref. UFI11/26). A. Gredilla is grateful to the UPV/EHU for her post-doctoral fellowship.

References

[1] S. Wold, Chemom. Intell. Lab. Syst. 30 (1995) 109.
[2] L. Devroye, L. Gyorfi, G. Lugosi, A Problabilistic Theory of Pattern Recognition, Springer-Verlag, New York, USA, 1996.
[3] A.K. Jain, R.P.W. Duin, J. Mao, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 4.
[4] B.K. Lavine, W.S. Rayens, Compr. Chemom. 3 (2009) 507.
[5] D.L. Massart, L. Kaufman, The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, John Wiley & Sons, New York, USA, 1983.
[6] K.L. Peterson, Rev. Comp. Chem. 16 (2000) 53.
[7] A. Smilde, R. Tauler, Chemom. Intell. Lab. Syst. 106 (2011) 1.
[8] S. Mas, A. de Juan, R. Tauler, A.C. Olivieri, G.M. Escandar, Talanta 80 (2010) 1052.
[9] S. Wold, K. Esbensen, P. Geladi, Chemom. Intell. Lab. Syst. 2 (1987) 37.
[10] S.C. Rutan, A. de Juan, R. Tauler, Comp. Chem. 2 (2009) 249.
[11] P. Paatero, P.K. Hopke, B.A. Begum, S.K. Biswas, Atmos. Environ. 39 (2005) 193.
[12] E. Acar, B. Yener, IEEE Trans. Knowl. Data Eng. 21 (2009) 6.
[13] J. Jabbour, F. Keita-Ouane, C. Hunsberger, R. Sánchez-Rodríguez, P. Gilruth, N. Patel, A. Singh, M.A. Levy, S. Schwarzer, Environ. Dev. 3 (2012) 5.
[14] M.M.R. Mostert, G.A. Ayoko, S. Kokot, Trends Anal. Chem. 29 (2010) 430.
[15] O. Abollino, M. Aceto, M. Malandrino, E. Mentasti, C. Sarzanini, R. Barberis, Environ. Pollut. 119 (2002) 177.
[16] B. Škrbic, N. Đurišić-Mladenovic, Chemosphere 80 (2010) 1360.
[17] A. Gredilla, J.M. Amigo, S. Fdez-Ortiz de Vallejuelo, A. de Diego, R. Bro, J.M. Madariaga, Anal. Methods 4 (2012) 676.
[18] K.P. Singh, A. Malik, V.K. Singh, D. Mohan, S. Sinha, Anal. Chim. Acta 550 (2005) 82.
[19] T. Yli-Tuomi, P.K. Hopke, P. Paatero, M.S. Basunia, S. Landsberger, Y. Viisanen, J. Paatero, Atmos. Environ. 37 (2003) 4381.
[20] M. Terrado, M.P. Lavigne, S. Tremblay, S. Duchesne, J.P. Villeneuve, A.N. Rousseau, D. Barceló, R. Tauler, J. Hydrol. 369 (2009) 416.
[21] L. Balbinot, P. Smichowski, S. Farias, M.A.Z. Arruda, C. Vodopivez, R.J. Poppi, Spectrochim. Acta Part B 60 (2005) 725.
[22] E. Peré-Trepat, A. Ginebreda, R. Tauler, Chemom. Intell. Lab. Syst. 88 (2007) 69.
[23] D.S. McLusky, M. Elliott, The Estuarine Ecosystem-Ecology, Oxford University Press, Oxford, UK, Threats and Management, 2004.
[24] N. Fernández, J. Bellas, J. Lorenzo, R. Beiras, Water Air Soil Pollut. 189 (2008) 163.
[25] V. Wepener, L. Bervoets, V. Mubiana, R. Blust, Mar. Pollut. Bull. 57 (2008) 624.
[26] R.A. Doong, S.H. Lee, C.C. Lee, Y.C. Sun, S.C. Wu, Mar. Pollut. Bull. 57 (2008) 846.
[27] A.C. Ruiz-Fernandez, F. Paez-Osuna, C. Hillaire-Marcel, M. Soto-Jimenez, B. Ghaleb, Bull. Environ. Contam. Toxicol. 67 (2001) 741.
[28] A. Helland, G. Aberg, J. Skei, Mar. Chem. 78 (2002) 149.
[29] A.P. Mucha, M.T.S.D. Vasconcelos, A.A. Bordalo, Mar. Environ. Res. 60 (2005) 531.
[30] S.M. Praveena, A. Ahmed, M. Radojevic, M.H. Abdullah, A.Z. Aris, Bull. Environ. Contam. Toxicol. 81 (2008) 52.
[31] E.H.P. Van Hees, E.I.B. Chopin, T.M. Sebastian, G.D. Washington, L.M. Germer, P. Domanski, D. Martz, L. Schweitzer, J. Great Lakes Res. 36 (2010) 606.
[32] E.d.A. Passos, J.C. Alves, I.S. dos Santos, J.d.P.H. Alves, C.A.B. Garcia, A.C. Spinola Costa, Microchem. J. 96 (2010) 50.
[33] W. Zhang, X. Liu, H. Cheng, E.Y. Zeng, Y. Hu, Mar. Pollut. Bull. 64 (2012) 712.
[34] C.Y. Chung, J.J. Chen, C.G. Lee, C.Y. Chiu, W.L. Lai, S.W. Liao, Environ. Monit. Assess. 173 (2011) 499.
[35] W. Tang, B. Shan, H. Zhang, Z. Mao, J. Hazard. Mater. 176 (2010) 945.
[36] F. Lorenzo, A. Alonso, M. Pellicer, J. Pagés, M. Pérez-Arlucea, Environ. Geol. 52 (2007) 789.
[37] A. Sainz, F. Ruiz, Chemosphere 62 (2006) 1612.
[38] W. Jambers, A. Smekens, R. Van Grieken, V. Shevchenko, V. Gordeev, Colloids Surf. A 120 (1997) 61.
[39] S.K. Sundaray, B.B. Nayak, S. Lin, D. Bhatta, J. Hazard. Mater. 186 (2011) 1837.
[40] M. Chatterjee, E.V. Silva Filho, S.K. Sarkar, S.M. Sella, A. Bhattacharya, K.K. Satpathy, M.V. Prasad, S. Chakraborty, B.D. Bhattacharya, Environ. Int. 33 (2007) 346.
[41] A.S. Maest, D.A. Crerar, R.F. Stallard, J.N. Ryan, Chem. Geol. 81 (1990) 133.
[42] T. Nasrabadi, G.N. Bidhendi, A. Karbassi, N. Mehrdadi, Environ. Monit. Assess. 171 (2010) 395.
[43] I.C.A.C. Bordon, J.E.S. Sarkis, G.M. Gobbato, M.A. Hortellani, C.M. Peixoto, J. Braz. Chem. Soc. 22 (2011) 1858.
[44] M. Jayaprakash, M.P. Jonathan, S. Srinivasalu, S. Muthuraj, V. Ram-Mohan, N. Rajeshwara-Rao, Estuar. Coast. Shelf Sci. 76 (2008) 692.

[45] A.R. Karbassi, I. Bayati, F. Moattar, Environ. Sci. Technol. 3 (2006) 35.
[46] S. Fdez-Ortiz de Vallejuelo, G. Arana, A. de Diego, J.M. Madariaga, J. Hazard. Mater. 181 (2010) 565.
[47] A. Machado, C. Magalhaes, A.P. Mucha, C.M.R. Almeida, A.A. Bordalo, Estuar. Coast. Shelf Sci. 99 (2012) 145.
[48] T.A. DelValls, J.M. Forja, A. Gómez-Parra, Ciencias Marinas 24 (1998) 127.
[49] B. Rubio, M.A. Nombela, F. Vilas, Mar. Pollut. Bull. 40 (2000) 968.
[50] Z.L. He, M. Zhang, P.J. Stoffella, X.E. Yang, Environ. Geol. 50 (2006) 250.
[51] A. Cesar, R.B. Choueri, I. Riba, C. Morales-Caselles, C.D. Pereira, A.R. Santos, D.M. Abessa, T.A. DelValls, Environ. Int. 33 (2007) 429.
[52] N. Tue, T. Quy, A. Amano, H. Hamaoka, S. Tanabe, M. Nhuan, K. Omori, Water Air Soil Pollut. 223 (2012) 1315.
[53] A.M. Idris, Microchem. J. 90 (2008) 159.
[54] C. Quelle, V. Besada, J.M. Andrade, N. Gutiérrez, F. Schultze, J. Gago, J.J. González, Talanta 87 (2011) 197.
[55] D.P. Mukherjee, B. Kumar, Arch. Appl. Sci. Res. 4 (2012) 1155.
[56] K.K. Balachandran, C.M. Laluraj, G.D. Martin, K. Srinivas, P. Venugopal, Environ. Forensics 7 (2006) 345.
[57] H. Pekey, G. Doğan, Microchem. J. 106 (2013) 233.
[58] E. Daka, M. Moslen, C. Ekeh, I. Ekweozor, Bull. Environ. Contam. Toxicol. 78 (2007) 515.
[59] J.A. González-Pérez, J.R. De Andrés, L. Clemente, J.A. Martín, F.J. González-Vila, Environ. Chem. Lett. 6 (2008) 41.
[60] R.H.C. Emmerson, S.B. O'Reilly-Wiese, C.L. Macleod, J.N. Lester, Mar. Pollut. Bull. 34 (1997) 960.
[61] G.E. Millward, I. Herbert, Environ. Pollut. Bull. 2 (1981) 265.
[62] M.P. Jonathan, S.K. Sarkar, P.D. Roy, M.A. Alam, M. Chatterjee, B.D. Bhattacharya, A. Bhattacharya, K.K. Satpathy, Ecotoxicology 19 (2010) 405.
[63] S. Fdez-Ortiz de Vallejuelo, G. Arana, A. de Diego, J.M. Madariaga, Chemosphere 85 (2011) 1347.
[64] M.B. Álvarez, M. Garrido, A.G. Lista, B.S. Fernández Band, Anal. Chim. Acta 620 (2008) 34.
[65] R.B. Choueri, A. Cesar, R.J. Torres, D.M.S. Abessa, R.D. Morais, C.D.S. Pereira, M.R.L. Nascimento, A.A. Mozeto, I. Riba, T.A. DelValls, Ecotoxicol. Environ. Saf. 72 (2009) 1824.
[66] U.C. Panda, P. Ratha, K.C. Sahu, S. Majumdar, S.K. Sundaray, Asian J. Water Environ. Pollut. 3 (2006) 85.
[67] J. Moros, A. Gredilla, S. Fdez-Ortiz de Vallejuelo, A. de Diego, J.M. Madariaga, S. Garrigues, M. de la Guardia, Talanta 82 (2010) 1254.
[68] H. Cao, M. Li, Y. Hong, J.D. Gu, Syst. Appl. Microbiol. 34 (2011) 513.
[69] M. Alvarez-Guerra, D. Ballabio, J.M. Amigo, J.R. Viguria, R. Bro, J. Chemometr. 24 (2009) 379.
[70] J. Moros, S. Fdez-Ortiz de Vallejuelo, A. Gredilla, A. de Diego, J.M. Madariaga, S. Garrigues, M. de la Guardia, Environ. Sci. Technol. 43 (2009) 9314.
[71] R. Arias, A. Barona, G. Ibarra-Berastegi, I. Aranguiz, A. Elías, J. Hazard. Mater. 151 (2008) 78.
[72] M.B. Alvarez, C.E. Domini, M. Garrido, A.G. Lista, B.S. Fernandez-Band, J. Soils Sediments 11 (2011) 657.
[73] M.K. Reid, K.L. Spencer, Environ. Pollut. 157 (2009) 2275.
[74] F. Ruiz, M.L. González-Regalado, J. Borrego, J.A. Morales, J.G. Pendón, J.M. Muñoz, Environ. Geol. 34 (1998) 270.