Fresenius J Anal Chem (2001) 370 : 719–722

ORIGINAL PAPER

V. Simeonov · I. Stanimirova · S. Tsakovski

Multivariate statistical interpretation of coastal sediment monitoring data

Received: 30 January 2001 / Revised: 20 March 2001 / Accepted: 23 March 2001

Abstract Multivariate statistical analysis of sediment data (input matrix 122×15) collected from 122 sampling sites from the western coastline of the USA and analyzed for 15 analytes indicates that the data structure could be explained by four latent factors. These factors are conditionally named "anthropogenic", "organic", "natural", and "hot spots". They explain over 85% of the total variance of the data system, which is an acceptable value for the PCA model. The receptor models obtained after regression of the mass on the absolute principal components scores ensures reliable estimation of the contribution of each possible natural or anthropogenic source to the mass of each chemical component. It can be concluded that the region of interest reveals a different pattern of pollution compared with the eastern coastline treated statistically in a previous study.

Introduction

There is serious concern throughout the world about the extent of sediment contamination in coastal regions. The factors affecting the composition of the marine sediments are, in principle, naturally occurring or anthropogenic. Thus, the significance of sediment monitoring as a tool for indication both of natural diagenetic processes and of pollution events of the aquatic environment is constantly increasing. Most contaminants released into coastal waters rapidly become associated with marine particulate matter and incorporated into sediments. In this way the coastal sediments become a very important and, to some extent, specific bioindicator of the nature of the pollution events. Although traditional natural processes modify and redistribute contaminants between solid and water phases, immobilization by sedimentation plays a dominant role for most of the typical pollutants. According to Martin and Whitfield [1] the accumulation of metal contaminants in coastal sediments gives a picture of the spatial and temporal history of pollution.

Large-scale studies in the USA in the last decade have monitored major and trace elements in a wide coastal region including sampling sites on the eastern and western coasts of the USA. The studies have been organized and performed by large institutions such as the National Oceanic and Atmospheric Administration (NOAA) by programs such as National Status and Trends (NS & T) Program for Marine Environment [2–4]. These programs make it possible to compare contaminant metal levels over large scales of distance and time periods.

The idea of separating contaminant and natural components, and their role in the formation of the sediments is very interesting and important. Correlation and regression analysis are usually used to estimate the contribution of the different types of factor and source (either natural or anthropogenic) to sediment formation and contaminant accumulation [5–7]. These approaches are usually performed after appropriate selection of major components, related to the morphology of the sediments, as tracers of the solid sediment phase. No strict rules are introduced about the choice of tracers but generally iron, aluminium, total organic carbon (TOC) content or sediment grain size are preferred for data interpretation. The reasons for these choices are that iron and aluminium are present at constant and high concentrations, and are not affected by anthropogenic activities, TOC content reveals important complexation processes in the sediment phase with participation of anthropogenic effects, and grain size is related to the processes of adsorption of heavy metals.

The traditional univariate approach (linear regression to some tracer component [5]) is widely applied including for samples collected by the NS & T program. The baseline model suggested enables simple estimation of pollution processes (with iron as tracer). It is assumed that sampling sites with population less than 10,000 located 20 km from

Dedicated to Professor Dr. Klaus Danzer on the occasion of his 65th birthday

V. Simeonov (✉) · I. Stanimirova · S. Tsakovski
Chair of Analytical Chemistry, Faculty of Chemistry,
University of Sofia "St. Kl. Okhridski", 1164 Sofia,
J. Bourchier Blvd. 1, Bulgaria
e-mail: vsimeonov@chem.uni-sofia.bg

the coastline are anthropogenically unaffected. The slope of the regression line (iron concentration as a function of heavy metal concentration) is compared with the ratio of the concentrations in the sediment formation rock. If both values are statistically comparable, the baseline model enables calculation of the natural heavy metal concentration in the sediment. Thus, each sampling site could be characterized in respect of the heavy metal pollution (polluted sites are those where there is a statistically significant difference between the amount of heavy metal predicted and real concentration measured).

As was pointed out in a recently published study, this model is seriously restricted [8]. In addition, the advantages of multivariate statistical data treatment were demonstrated as applied to a large data-set, provided by NOAA, from sites on the Atlantic coastline of the USA (Eastern coast and Mexican Gulf). In the study cited four latent factors responsible for the data structure, conditionally named "inorganic natural", "inorganic anthropogenic", "bioorganic", and "bioorganic anthropogenic", were extracted. Further, a source-apportioning model was proposed for estimating the contribution of each possible source of emission to sediment formation. This multivariate modeling approach using principal components analysis (PCA) and multiple regression on principal components (PC/MR) was developed further with partial least-squares modeling (PLS) [9] to evaluate the role of a group of tracers in the determination of pollution by heavy metals and, vice versa, to predict the type of sediment if pollution data is available. In both instances the advantage of multivariate statistical modeling was discussed.

In this study multivariate statistical modeling has been conducted on a sediment monitoring data-set from the western coast of the USA collected by the same NS & T program of the NOAA. Comparison of the factors determining the data structure and the apportioning procedure enable comparison of the role of natural and anthropogenic influences in the sediment formation along the two coastlines.

Experimental

Sampling and sample analysis. Since 1984 NOAA of the USA has regulated the NS & T program, which consists of two major projects – The National Benthic Surveillance and the Mussel Watch. The aim of the program is to evaluate pollution along the USA coastline and to locate sources of emissions by monitoring coastal and estuary systems [1–3].

Sampling for the data used in this study was performed along the western part of the USA coastline. A total of 122 sampling sites was involved. Three samples were obtained from each site; each of these samples is a mean from three sampling sessions. During sediment monitoring three of the samples were analyzed for organic material and the other three for inorganic components and grain size.

In the surveillance projects of NOAA [10] sediment samples were obtained with a specially constructed box corer or a standard Smith–McIntyre bottom grab. Sediment analysis for each site consisted of the analytical steps mentioned above (organic analysis, inorganic analysis and, grain-size measurement). The elements measured in the samples were Al, As, Cd, Cr, Cu, Fe, Mn, Pb, Hg, Ni, Ag, Sn, and Zn; grain-size and total organic carbon (TOC) were also measured.

Dry sediment (0.10 to 0.45 g) was placed in a closed Teflon vial inside a Teflon-lined stainless-steel bomb, treated with mixtures such as HNO_3–HF, HNO_3–KCl–HF, or HNO_3–$HClO_4$–HF, and heated either conventionally or by use of microwaves. The digestion vessel was designed to withstand sufficient temperature and pressure to affect complete sample dissolution while minimizing contamination and preventing loss of volatile elements such as Hg. Aqua regia was used to oxidize organic matter partially and to maintain oxidizing conditions to stabilize elements such as Hg and As. A solution of H_3BO_3 was added to dissolve the insoluble fluorides in the sediment. The final solution was analyzed by ETAAS (Al, Fe, Mn, Zn), ICP–AES, AAS (cold vapor method for Hg), and XRF (Ag, As, Cd, Cr, Cu, Ni, Pb, Sn). A more detailed description of sampling, sample preparation, and analytical signal measurement (including TOC and grain size) can be found elsewhere [5, 11].

Certified reference materials (MESS-1, BCSS-1, and NBS 1646) were analyzed with each batch of samples. A measured value for the reference material within the published tolerance limits was the criterion for bias control and checking data quality. Precision for Al, Fe, Mn, and Zn was ≤ 5% (relative standard deviation); for As, Cd, Cu, Cr, Hg, Ni, and Pb it was ≤ 10%; for Ag and Sn it was ≤ 15%. The methods have been evaluated in four intercomparison sessions [10].

Chemometric methods for multivariate analysis. Principal components analysis (PCA) and receptor modeling were applied in this study. As a chemometric approach PCA is very important method that enables reduction of the space dimension of the variables and the extraction of new variables called "latent factors" which are linear combinations of the old variables [12, 13]. The substitution of many correlated variables by a smaller number of independent factors (principal components) is very convenient for data interpretation and determination of the data structure. PCA is so widely used that a more detailed description is not needed.

Receptor modeling (multiple regression on the absolute principal components scores – APCS modeling [14]) is widely used in source-apportioning models for aerosol data. It enables determination of the contribution of each possible source to the mass of each chemical variable. In this way it is much easier to understand the quantitative contribution of each source (identified as a latent factor) to the pollution.

In Table 1 summarizes the statistics of the monitoring data for all sampling sites (122) and for all chemical analytes (15).

Results and discussion

The input data matrix was treated by means of the software package Statistica 5.0. PCA in its orthogonal rotation mode was used; the factor loadings of the four principal components are given in Table 2. The principal components found describe over 85% of the total variance of the system.

The first latent factor could be conditionally named "anthropogenic", the second – "organic", the third – "natural" and the fourth – "hot spots". These names reflect the natural composition, (PC3), on the one hand, and, on the other hand, anthropogenic influences (PC1, PC2, PC4) on the sedimentation, for example various industrial activities in the region of interest.

A question that often arises in multivariate data modeling is how many meaningful eigenvectors should be retained, especially when the objective is to reduce the dimensionality of the data. It is assumed that, initially, eigenvectors contribute only structural information, which is also referred to as systematic information. With increasing number of eigenvectors, however, noise progressively

Table 1 Statistics for the chemical concentrations in the sediments (mg kg^{-1} dry weight; number of samples included 122)

Variable	Mean	Confidence –95%	Confidence 95%	Minimum	Maximum	Standard deviation	Standard error
TOC	8513	7612	9563	88	50600	7211	455
Al	60408	60005	62511	128	131072	17418	1036
As	7	6.5	7.3	0.001	25	4.2	0.25
Cd	0.5	0.4	0.5	0.001	6.4	0.5	0.03
Cr	105	98	120	20	697	125	7
Cu	38	34	45	0.001	219	41	3
Fe	36161	30371	42798	1540	928421	47994	3210
Pb	24	22	28	0.001	133	22	1.4
Mn	533	507	568	49	1933	239	14
Hg	0.2	0.15	0.21	0.0001	2	0.2	0.01
Ni	33	27	35	0.001	147	29	2
Ag	0.4	0.3	0.5	0.001	8.2	0.7	0.03
Sn	1.8	1.5	2	0.001	14.6	1.3	0.1
Zn	103	102	112	5.6	389	62	4
Grain size	0.4	0.4	0.5	0.01	1	0.3	0.02

Table 2 Factor loadings (PCA)

Variable	Factor 1	Factor 2	Factor 3	Factor 4
TOC	0.47	<u>0.78</u>	0.06	0.12
Al	0.22	0.06	<u>0.75</u>	0.09
As	0.51	0.18	0.06	<u>0.70</u>
Cd	<u>0.70</u>	0.11	0.23	0.31
Cr	0.37	0.03	0.41	<u>0.88</u>
Cu	<u>0.81</u>	0.11	0.31	0.36
Fe	0.41	0.03	<u>0.74</u>	0.23
Pb	<u>0.79</u>	0.02	0.18	0.38
Mn	0.18	0.12	<u>0.89</u>	0.28
Hg	<u>0.71</u>	0.07	0.04	0.34
Ni	0.42	0.15	0.12	<u>0.79</u>
Ag	<u>0.69</u>	0.17	0.09	0.27
Sn	<u>0.77</u>	0.08	0.23	0.44
Zn	<u>0.82</u>	0.10	0.36	0.39
Grain size	0.47	<u>0.72</u>	0.37	0.04
Explained variance %	32.2	24.4	15.8	12.6

Note: statistically significant loadings are underlined

Table 3 Residual variance values as function of number of latent factors

Residual variance (%)	Number
67.8	1
43.4	2
27.6	3
15.0	4
13.8	5
12.4	6
11.6	7

contaminates the eigenvectors and eventually only pure noise may be carried by the high-order eigenvectors. The problem then is to define the number of eigenvectors which account for a maximum of structure while carrying minimum noise. There are different means of validating the PCA model [15]. Some well-known and widely used methods of validation are, for instance, the empirical test based on the scree plot, which represents the residual variance as a function of the number of eigenvectors that have been extracted; Malinowski's F-test statistically designed to compare variance contributed by a structural eigenvector with that contributed by the error eigenvectors; cross-validation method based on internal validation, which means that one predicts each element in the data-set from the results of an analysis of the remaining elements. In our study we relied on scree-plot considerations (Table 3).

In scree-plot validation it is assumed that structural eigenvectors explain successively less variance in the data. The error eigenvectors, however, when they account for random errors in the data, should be equal. It is readily apparent when the structural information in the data is nearly exhausted. This situation determines the number of structural eigenvectors (four in our work). The first four principal components indicate real structural eigenvectors, because the next are typical error vectors (explanation of the variances within the range of the noise).

There are some differences between the structures of the results data obtained from the eastern coast and the Mexican Gulf. For the latter the four latent factors were more easily interpretable – two types of anthropogenic effect (inorganic and organic) probably corresponding to industrial and indoor activities could be defined, as could two other types of natural influence related to the macrocomponent structure of, and organic interactions in, the sediments [8, 9]. In this case (western coastline) separation between the anthropogenic effects is achieved relative to the existing "hot-spot" emitters in the region (smelters and metallurgical plants as typical sources of As, Cr, and Ni). Because PC3 is regarded as a "natural" factor containing the major matrix components manganese, iron, and aluminum, it is of significant importance to understand if there is also an anthropogenic contribution of local sources of

Table 4 Contribution of the emitting sources (latent factors) to the total mass of the chemical content of the sediments (%, for 122 sites)

Variable	Intercept	"Anthropogenic"	"Organic"	"Natural"	"Hot spots"	Exp. found mass	Calculated mass	R^2
TOC	3.1	31.5	56.2	–	9.2	8512	8322	0.82
Al	11.6	9.8	–	78.6	–	60987	60121	0.79
As	–	32.3	5.1	–	62.6	7.1	6.7	0.84
Cd	–	65.5	5.2	10.8	18.5	0.41	0.37	0.67
Cr	11.3	18.1	–	22.7	47.9	118	114	0.69
Cu	–	50.5	–	24.4	25.1	38	37	0.86
Fe	–	32.1	–	57.6	10.3	36220	36054	0.88
Pb	11.2	51.5	–	12.6	24.7	26	25	0.79
Mn	6.6	7.3	6.1	61.3	18.7	515	517	0.81
Hg	9.1	62.2	–	–	28.7	0.2	0.2	0.83
Ni	6.5	22.4	9.1	8.6	53.4	33	32	0.87
Ag	18.3	49.9	10.4	–	21.6	0.4	0.4	0.69
Sn	7.2	53.2	–	20.5	19.1	2.1	2.2	0.78
Zn	–	57.2	–	22.3	20.5	109	108	0.83

these components to sediment formation. Similar "hot spots" could not be identified in the eastern region.

By use of the results from PCA and multiple regression of the absolute principal components scores (APCS), a receptor modeling procedure was performed. The total mass of sediments for each sampling site was calculated (dependent variable MASS). The independent variables were the calculated by a simple procedure of regression on APCS [14] corresponding to the four principal components found by PCA. The regression model enables determination of the contribution of each latent factor (identical in interpretation with source of emission type – natural or anthropogenic). The apportioning results are presented in Table 4.

In the section "intercept" of the table the percentage of the unexplained mass is indicated. It is apparent that the statistical significance of the receptor models obtained is quite satisfactory, as indicated by the value of the correlation coefficient r^2 for comparison of the experimentally determined mass with that calculated by use of the models.

The latent factors in the western coastal region determine a slightly different apportioning pattern compared with the results from the eastern coast and the Mexican Gulf [8, 9]. Again, the anthropogenic factors contain most of the chemical elements, but to different extents. It is apparent that the major sediment matrix components comprising the natural factor (Al, Mn, and Fe) have some anthropogenic origin. The organic pollution reflected in the "organic" latent factor is probably because of the influence of typical organic components but also because of a variety of metal–organic compounds. The grain size is also influenced by the sedimentation of organic matter.

In this way a relatively well-defined picture is obtained of the data structure of the sediment from the region of interest.

Conclusion

This multivariate statistical modeling reveals a picture for the western coastline of the USA similar to that obtained by monitoring the data structure of sediments from the eastern coast and the Mexican Gulf regions. The same number of structural factors is extracted from the data-set if each sampling site is considered as the average of the chemical concentrations obtained for samples collected in several consecutive years.

References

1. Martin JM, Whitfield M (1983) The Significance of the River Input of Chemical Elements to the Ocean, Plenum Press, New York
2. Cantillo AY, O'Connor TP (1992) Chem Ecol 7:31
3. Hanson PJ, Evans DW, Colby DR (1993) Marine Environ Res 36:237
4. Daskalakis K, O'Connor TD (1995) Marine Environ Res 40: 389
5. Daskalakis K, O'Connor TD (1979) Environ Sci Technol 13: 470
6. Goldberg ED, Griffin JJ, Hodge V, Koide M, Windom H (1979) Environ Sci Technol 13:588
7. Calder FD, Ryan JD, Smith RG (1989) Environ Sci Technol 23:47
8. Stanimirova I, Tsakovski S, Simeonov V (1999) Fresenius J Anal Chem 365:489
9. Simeonov V, Tsakovski S, Massart DL (1999) Toxicol Environ Chem 72:81
10. Second Summary of data on Chemical Contaminants in Sediments from the NS&TP, NOAA, Technical Memorandum 59, NOS OMA, Rockville, MD, 1991
11. National Status and Trends Program: Monitoring Site Descriptions (1984–1990) for the National Mussel Watch and Benthic Surveillance Projects. NOAA Tech Memo NOS OMA 40, NOAA Office of Oceanography and Marine Assessment, Rockville, MD, 1988
12. Massart DL, Kaufman L (1983) The Interpretation of Analytical Chemical Data by the Use of Cluster Analysis, J. Wiley, New York
13. Esbensen K, Schoenkopf S, Midtgaard T (1994) Multivariate Analysis in Practice, CAMO AS, Trodheim
14. Thurston G, Sprengler J (1985) Atmos Environ 19:9
15. Vandenginste B, Massart DL, Buydens L, De Jong S, Lewi P, Smeyers–Verbeke J (1998) Handbook of Chemometrics and Qualimetrics, Elsevier, Amsterdam