Source apportionment for contaminated soils using multivariate statistical methods

Sonnia Parra [a,*], Manuel A. Bravo [a,1], Waldo Quiroz [a,1], Teresa Moreno [b], Angeliki Karanasiou [b], Oriol Font [b], Víctor Vidal [c], Francisco Cereceda-Balic [c]

[a] Laboratorio de Química Analítica y Ambiental, Pontificia Universidad Católica de Valparaíso, Avenida Brasil, 2950 Valparaíso, Chile
[b] Institute of Environmental Assessment and Water Studies "IDAEA", CSIC, C/Jordi Girona 18–26, 08034 Barcelona, Spain
[c] Laboratory of Environmental Chemistry, Center for Environmental Technology, Universidad Técnica Federico Santa María, Valparaíso, Chile

ARTICLE INFO

ABSTRACT

The application of statistical techniques for the recognition and identification of contamination sources has become an increasingly important tool. The chemical compositions of soil samples collected in the Puchuncaví Valley (Chile) provide a dataset suitable for the application of source apportionment techniques such as positive matrix factorization (PMF) and principal component analysis (PCA) with varimax rotation. PMF allowed the identification of the chemical profile and the relative contribution of three interpretable factors related to three contamination sources. Combining these results with a PCA analysis successfully showed that the main source of pollution emits Cu, Zn, As, Se, Mo, Sn, Sb and Pb. Therefore, the use of source profiles for contaminated soils shows much promise both for incorporating well-established knowledge about pollution sources and as a tool for incremental, exploratory data analysis.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

Receptor models are useful tools to identify sources of a specific pollutant and to estimate the quantitative contribution of each source based on environmental data [1]. In particular, receptor models are mathematical procedures for identifying and quantifying the sources, primarily on the basis of concentration measurements at the receptor site and, generally, without the need for emission inventories and meteorological data [2].

The multivariate statistical model principal component analysis (PCA) has been widely used to identify pollutant emission sources, generating satisfactory results [3–8]. PCA allows the numerical adjustment of a linear model to describe the main relationships among process variables [9]. However, the application of PCA is subject to limitations including the exclusion of uncertainties during matrix decomposition and the possible presence of negative factor loadings, which are difficult to interpret in terms of physical parameters such as concentration and mass [10,11].

The positive matrix factorization (PMF) model is a multivariate factor analysis tool developed by Paatero and Tapper [10,12,13]. This receptor model largely overcomes the limitations of PCA by using experimental uncertainties in the data matrix analysis and constraining the solutions to non-negative values [10,14]. PCA has been used extensively in studies of atmospheric pollution [5,15,8,16–18], soil contamination [19–21] and other environmental matrices.

The main objective of this study was to establish the individual contribution of the different sources that contribute to the levels of trace elements in the soil of Puchuncaví Valley (Chile). In contrast to previous studies based only on an analysis of the relationship between individual concentration and distance from the source, this study utilized the multivariate receptor models PCA and PMF to identify emission sources along with their relative contribution to soil pollution.

2. Materials and methods

2.1. Chemical analyses

The concentrations of major and trace elements in each sample were determined by inductively coupled plasma-atomic emission spectroscopy (ICP-AES, Thermo Jarrell-Ash), and inductively coupled plasma-mass spectroscopy (ICP-MS, Thermo Electron Corporation), respectively. A weight of ~0.1 g of each sample was digested using pure HNO_3 (65%) and HF (40%) for 4 h at 90 °C. Solutions were then evaporated after the addition of 60% pure $HClO_4$. The resulting solid was dissolved by adding HNO_3 and water (MQ) to obtain a 5% (v/v) HNO_3 solution for analysis. Finally, the samples were diluted to 25 ml with MQ distilled water.

The nitrate and sulfate contents were evaluated by high-performance liquid chromatography and conductivity detection (Waters). Samples for mercury analysis were pyrolyzed in a combustion tube at approximately

750 °C under an oxygen-rich atmosphere and collection on a gold amalgamator. Mercury was then determined directly on the solid with atomic absorption spectroscopy using gold amalgam (LECO AMA-254). All chemicals used were of analytical grade or equivalent.

Finally, to validate the analytical methodologies, two certified reference materials (certified in all elements considered in this study) were analyzed using the above-described methodology, and the obtained values were compared with the certified contents. The materials selected were NIST 1633b (bituminous Coal Fly Ash) and soil sample SO-2 (Podzolic soil, CANMET). The results obtained for both materials were statistically similar to the certified values ($p < 0.05$) (For details, please review the supplementary information).

2.2. Data analysis

2.2.1. Principal component analysis

The main objective of PCA is to reduce a large number of variables to a smaller set of factors while retaining most of the information from the original dataset [22]. PCA assumes that the variables are linearly related to a number of factors, p, so that the reduced concentration of each element is composed of the sum of the elemental contributions of each source of pollution in the respective site, as in Eq. (2):

$$Z_{ij} = \sum_{k=1}^{p} a_{ik} \cdot c_{kj} \tag{2}$$

Table 1
Concentrations of major elements in soils of Puchuncaví Valley (2007–2009).

Sampling sites	(mg kg^{-1} × 10^3)	Al	Ca	Fe	K	Mg	Na	Ti
2007								
La Greda	Average	39.5	164.5	20.1	7.6	3.8	10.7	2.4
	Max.	47.0	209.6	23.4	9.8	4.0	14.5	3.0
	Min.	26.1	135.1	14.4	5.1	3.3	7.1	1.5
Los Maitenes	Average	69.5	15.1	40.8	16.6	4.5	23.0	5.1
	Max.	72.8	15.7	41.9	17.1	4.6	23.7	5.5
	Min.	66.1	14.3	39.3	16.0	4.4	21.8	4.8
Valle Alegre	Average	69.8	17.8	47.7	12.9	4.8	19.7	4.3
	Max.	75.6	27.2	74.3	13.6	9.3	23.4	6.1
	Min.	66.5	13.9	36.3	12.0	2.6	18.1	2.6
Puchuncaví	Average	65.1	15.2	32.1	13.0	3.0	20.6	4.1
	Max.	69.6	17.0	33.1	13.7	3.5	21.8	5.8
	Min.	61.3	13.5	31.3	12.1	2.3	19.6	1.1
2008								
La Greda	Average	50.7	133.5	25.2	10.2	4.1	15.0	3.1
	Max.	79.1	273.3	37.2	16.5	4.6	25.3	4.3
	Min.	16.6	27.8	9.7	3.3	3.6	5.1	1.5
Los Maitenes	Average	66.9	14.5	42.3	15.2	4.5	21.6	5.4
	Max.	74.6	17.4	46.6	16.6	5.0	24.0	6.1
	Min.	51.9	12.6	33.3	13.1	3.5	18.9	4.1
Valle Alegre	Average	78.1	17.5	51.5	17.5	7.4	26.0	5.9
	Max.	83.4	18.7	57.3	17.6	9.6	26.6	7.1
	Min.	75.7	16.3	48.9	17.4	6.0	25.6	3.4
Puchuncaví	Average	75.2	14.7	39.7	12.8	5.6	21.5	3.2
	Max.	85.7	18.7	60.0	13.0	14.2	23.2	5.9
	Min.	68.9	13.0	32.1	12.4	2.5	20.4	1.3
2009								
La Greda	Average	86.9	21.8	39.9	15.3	5.1	23.5	4.4
	Max.	93.5	41.0	43.3	17.3	5.2	26.4	4.9
	Min.	83.0	13.4	35.0	11.9	4.8	18.5	3.8
Los Maitenes	Average	81.7	17.1	49.9	16.9	4.9	24.7	4.7
	Max.	83.3	18.7	53.5	17.5	5.3	25.5	7.2
	Min.	80.2	15.4	43.3	16.3	4.5	24.1	2.6
Valle Alegre	Average	80.1	18.1	53.7	16.9	7.2	25.8	7.2
	Max.	81.2	18.5	55.2	17.9	8.3	26.2	7.6
	Min.	79.4	17.7	51.2	16.5	5.9	24.8	6.9
Puchuncaví	Average	74.7	13.4	34.7	13.1	3.3	21.9	4.8
	Max.	76.6	15.4	38.2	13.8	4.0	23.6	6.7
	Min.	71.7	12.2	31.4	11.9	2.9	19.3	1.7
Maitencillo	Average	85.2	24.2	58.2	14.9	8.5	25.1	4.1

where a_{ik} is the loading factor of element i for each component k (source) and C_{kj} is the score factor of each component k (source) for each j sample.

The orthogonal transformation (normalized varimax) is applied to maximize the number of loading factors without changing the total variance or the variance of each element in the model. Once the rotation has been completed, the elements originating from the same source correspond to a single component with a high weight (loading factor), and this component is associated with a specific source. To interpret the respective factors, only loads above 0.70 are considered; this value generally allows the definition of the least number of factors with the highest weights [23].

2.2.2. Positive matrix factorization

The investigation of sources responsible for soil contamination in the Puchuncaví Valley was carried out by PMF.

PMF is a least-squares factor analysis based on the principle of mass conservation to assist in the identification of sources and their contributions to observed pollutant loadings [10,12]. The objective of PMF is to factorize a data matrix X (nxm), where n and m are the number of species and samples, respectively, into two separate matrices G (nxp) and F (pxm), in accordance with the following equation:

$$X_{ij} = \sum_{k=1}^{p} g_{ik} \cdot f_{kj} + e_{ij} \tag{3}$$

where X_{ij} is the jth elemental concentration measured in the ith sample, g_{ik} is the concentration contributed by source k to sample i (the contribution of each of the sources), f_{kj} is the mass fraction (mg kg^{-1}) of species j in source k (the profile of each of the sources) and e_{ij} is the part of the measurement that is not accounted for by the model, or the so-called residual [15].

PMF estimates values by minimizing the sum of the square of the residuals, which is expressed by the following equation [10,12,14,24]:

$$Q = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{e_{ij}}{s_{ij}} \right)^2 \tag{4}$$

where,

$$e_{ij} = X_{ij} - \sum_{k=1}^{p} f_{ik} \cdot g_{kj} \tag{5}$$

and S_{ij} is the uncertainty in the concentration of X_{ij}.

PMF analysis was performed using a software (PMF 3.0) developed by the US Environmental Protection Agency (EPA). The principal difference between PCA and PMF is that the non-negativity of factors (both loadings and scores, where $g_{ik} > 0$ and $f_{kj} > 0$) is built into the PMF model.

Another advantage of PMF is the calculation of the uncertainty in the concentration X_{ij}, which is determined by the following equation

$$S_{ij} = \left(l_j^2 + (d_j X_{ij})^2 \right)^{1/2} \tag{6}$$

where l_j is the detection limit for component j, which is the control sample plus three times the standard deviation of the control [25], and d_j represents the relative uncertainty in the matrix values X_{ij}, for high values of the measured parameters.

In similar studies, this uncertainty has been adjusted in the range 0.1–0.3 in order to minimize the Q values [15,25,26]. In this case, several approaches were evaluated and the best results were obtained using a combination of standard practices and estimation prior to be testing on the model.

In this work, several approaches were evaluated and the best results were obtained using a combination of standard practices and estimation