Chemometrics and Intelligent Laboratory Systems 155 (2016) 97–108

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

An iterative method for evaluating the inter-comparability between chemical mass balance and multivariate receptor models

George Argyropoulos [a,*], Constantini Samara [a], Evangelia Diapouli [b], Kostas Eleftheriadis [b]

[a] Aristotle University of Thessaloniki, Department of Chemistry, Environmental Pollution Control Laboratory, GR-541 24 Thessaloniki, Greece
[b] National Centre for Scientific Research "Demokritos", Institute of Nuclear & Radiological Sciences & Technology, Energy & Safety, 15310, Ag. Paraskevi, Attiki, Greece

ARTICLE INFO

Article history:
Received 12 December 2015
Received in revised form 17 March 2016
Accepted 29 March 2016
Available online 2 April 2016

Keywords:
Chemical Mass Balance (CMB)
Positive Matrix Factorization (PMF)
Iterative CMB source profiles
Cluster analysis
PMF Factor Mapping

ABSTRACT

A comparative study between Chemical Mass Balance (CMB) and multivariate receptor modeling techniques was conducted. The study involved common application of Robotic Chemical Mass Balance (RCMB) and 2-dimensional Positive Matrix Factorization (PMF). A two-fold methodology was developed for Source Apportionment (SA) of reactive species, in order to address previous limitations of CMB models. The developed methodology (a) uses a detailed set of theoretical source profiles, taking into account secondary reactions that were not considered in CMB modelling until now, and (b) implements a Least Squares (LS) fitting method that iteratively readjusts the values of independent variables in the CMB fit, providing (for the first time, to our knowledge) CMB source profiles as output data, in which secondary transformations may be reflected. A straight-forward computational procedure named Factor Mapping (FM) was developed as well, for intercomparison between RCMB and PMF. The distinctive feature of FM is that similarity measures are used not just to compare the results of the two models, as in previous intercomparison exercises, but, moreover, to actively assist in the physical interpretation of PMF factors, thus minimizing user interference. The intercomparison between RCMB and PMF also involved an independent evaluation of each model's performance in reproducing ambient concentrations of particulate matter (PM) and associated chemical constituents. Overall, it was shown that the employed series of computational steps substantially improve qualitative as well as quantitative agreement between the two models.

1. Introduction

The statistical analyses, which apportion observed concentrations of aerosol components to sources, have been termed Receptor Models (RMs). The major working hypothesis of all RMs is a rather simple mass conservation argument, assuming that there is no chemical interaction between emissions from different aerosol sources. Under this concept, it is possible to express the ambient concentrations of Particulate Matter (PM) and associated chemical species as linear sums of contributions from individual sources, according to the following equation [13,14,17]:

$$C_{ij} = \sum_{k=1}^{n} g_{ik} f_{kj} + e_{ij} \Leftrightarrow C = G \cdot F + E \qquad (1)$$

where C_{ij} is the measured concentration of species j in sample i ($\mu g\, m^{-3}$), g_{ik} is the contribution of source k to sample i ($\mu g\, m^{-3}$), f_{kj} is the mass concentration of species j in source k ($\mu g\, \mu g^{-1}$), e_{ij} is the error residual for each sample/species, and C, G, F, and E are the corresponding matrices.

There are two main categories of RMs, characterized by the degree of "a priori" knowledge that is required for pollutant sources affecting the study area. A common scenario is that the major sources are known and their emissions have been chemically characterized, i.e. the values of matrix F are provided as input data. In that case, Eq. (1) defines a Multiple Linear Regression (MLR) system for each ambient sample i, known as Chemical Mass Balance (CMB), which can be fitted by minimizing a Least Squares (LS) likelihood function (χ_i).

On the other hand, if there is little information about the chemical characteristics of source emissions, then Eq. (1) can be considered as a bilinear model, to which Factor Analysis (FA) may be applied, in order to produce a physically reasonable set of values for both matrices G and F. Depending on the constraints applied to G and F, there are various decomposition algorithms, such as Principal Component Analysis (PCA), UNMIX, Positive Matrix Factorization (PMF), as well as other factor analytical models [13,14,16,17,25–27].

Both categories of RMs are known to have their own strengths and weaknesses. One major disadvantage of CMB models arises from the assumption that chemical species do not react with each other, i.e. they add linearly in Eq. (1). This cannot be the case in the real world, where primary source emissions may undergo intense chemical transformations in the atmosphere to form secondary aerosols, before reaching the receptor site.

---

[*] Corresponding author.
E-mail address: geoarg@chem.auth.gr (G. Argyropoulos).

http://dx.doi.org/10.1016/j.chemolab.2016.03.032
0169-7439/© 2016 Elsevier B.V. All rights reserved.

In order to overcome this problem, the current practice, followed by CMB modelers, is to apportion the primary material that has not changed between source-receptor, and then, the remaining quantities of reactive species are apportioned using theoretical source profiles that correspond to secondary chemical compounds rather than directly to sources [39]. However, the particular methodology has been limited, so far, to a few chemical compounds (mainly ammonium sulphate and nitrate) that are formed by gas-to-particle conversions, neglecting other heterogeneous reactions that occur vastly in the atmosphere, too [12].

Multivariate RMs, on the other hand, can indeed apportion factors that may be linked to various secondary aerosols, since these models are based on information gathered at the receptor site, alone. Nevertheless, this commonly accepted advantage of multivariate RMs comes at a price, as there is often some inherent rotational ambiguity, present in any solution of a bilinear model [21,24].

Another challenging task regarding both categories of RMs is to establish an appropriate methodology for expressing model reliability and quantitative uncertainty levels, since, in the real world, one cannot check the model outputs against the actual values of source contributions [20]. So far, comparing the results of different RMs with each other, and/or against the ones of dispersion models, seems to be the most plausible way of determining their overall uncertainty, for this reason, it has become advisable for every source apportionment study to validate results by cross-checking with independent methods [11,41].

Previous intercomparison exercises have shown that CMB and multivariate RMs can both be successful in reproducing the total PM mass accurately, with good agreement regarding source identification [37]. On the other hand, there is often less comparability between the input source profiles that are employed for CMB modeling and the output source profiles that are correspondingly resolved by multivariate RMs [18]. In some cases, large differences have also been obtained, regarding the quantification of source contributions. For example, Yin et al. [41] found the estimates of the two model groups to differ by a factor of around 1.5 to 2, while even larger differences (up to a factor of 4) were reported by Viana et al. [37]. Further research is clearly needed to access whether RMs have actually become robust enough for integration into policy-making tools [11,12,18,33,37,41].

In view of the aforementioned need, this work presents a comparative study utilizing a series of computational steps, which substantially improve qualitative as well as quantitative agreement between a CMB and a multivariate receptor model. The study involved common application of Robotic Chemical Mass Balance (CMB), and 2-dimensional Positive Matrix Factorization (PFM2).

A two-fold methodology was developed for Source Apportionment (SA) of reactive species, in order to address previous limitations of CMB models. The developed methodology (a) uses a detailed set of theoretical source profiles, taking into account secondary reactions that were not considered in CMB modelling until now, and (b) implements a LS fitting method that iteratively readjusts the values of matrix *F*, providing (for the first time, to our knowledge) CMB source profiles as output data, in which secondary transformations may be reflected.

A straight-forward computational procedure named Factor Mapping (FM) was developed as well, for intercomparison between RCMB and PMF. FM involves an iteration loop, which explores (a) the strength of correlation between source compositions and (b) the degree of convergence between source contributions estimated by the two models. The distinctive feature of FM is that similarity measures are used not just to compare the results of the two models, as in previous intercomparison exercises, but, moreover, to actively assist in the physical interpretation of PMF factors. RCMB and PMF were also subjected to an independent evaluation of their performance in reproducing ambient concentrations of particulate matter (PM) and associated chemical constituents. The proposed approach for intercomparison between the two RMs is user independent, so it is expected to be useful in other exercises, as well.

2. Materials and methods

2.1. The ambient data set

The ambient data set was acquired from a PM_{10} sampling campaign that was carried out in Thessaloniki (40°38′E, 22°56′N), Greece, during the cold and the warm period of year 2007. The sampling site was located in the center of the city, at a place influenced by moderate traffic (~11,000 vehicles day^{-1}) and surrounded by commercial shops and residential buildings. Sampling equipment was situated on the roof (~3.0 m above ground level) of an air-quality monitoring station.

Sixty three 24-hour PM_{10} samples were collected, according to EN-12341, using two identical Low Volume PM_{10} Samplers (flow rate 2.3 m^3/h) operating in parallel. PM_{10} was collected on high purity quartz filters (Tissuquartz^{TM}, Pall) prefired at 500 °C for 4 h, and on Teflon filters (Zefluor, Pall 2 μm). The PM_{10} mass was gravimetrically determined on quartz filters, according to EN-12341.

Chemical analysis of PM_{10} has been described elsewhere [34]. Briefly, OC and EC were determined on punches of quartz filters by the Thermal Optical Transmission method. Ions (Na^+, NH_4^+, K^+, Mg^{2+}, Ca^{2+}, Cl^-, NO_3^-, SO_4^{2-}) were measured using ion chromatography following extraction of quartz filter sections with an ultra-pure water-isopropanol mixture (9:1 v/v). Elemental constituents (Mg, Al, Si, S, Cl, K, Ca, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Se, Br, Sr, Cd, Sn, Te, Ba, and Pb) were determined on Teflon filters by Energy Dispersive X-Ray Fluorescence (ED-XRF) analysis. There were not any missing values, while concentration values below the Limit of Detection (*LOD*) were substituted by *LOD*/2. More detailed information about the ambient data set is provided in the Supplementary Material (SM).

2.2. CMB modeling

2.2.1. Robotic vs. conventional CMB modeling

Robotic Chemical Mass Balance (RCMB) was developed and validated in Argyropoulos and Samara [5] and it has been used, ever since, for SA of PM in various Greek locations [3,4]. Briefly, RCMB is realized by applying a LS fitting method to each and every one of the possible combinations that can be made from the source profiles included to a given set of input data. Any converging applications of the employed LS fitting method are automatically ranked, during run-time, according to standard performance measures of previous CMB models, such as the ones of the US EPA CMB 8.2 model, which are given in Table 1. After the LS fitting method has been applied to all possible combinations of source profiles, if there are any converging applications meeting the individual performance measures, they are ranked according to an overall fitting index.

The explicit advantage of RCMB is that the best-fit combination of source profiles derives automatically from the maximization of an overall fitting index, providing a mathematically unique solution, which cannot be questioned readily, unless additional information becomes available for the study area [5]. It should, however, be noted that there is still no sufficient guarantee that the source contribution estimates will be realistic, if the employed source profiles have been unrepresentative of emissions at the receptor site.

2.2.2. CMB methodology for allowing the source profiles to vary

The standard method for fitting CMB models has so far been a modification of Britt and Luecke's algorithm [7], known as the Effective Variance Weighted Least Squares (EFWLS) solution [40], which takes into account both ambient (σ_{Cij}) and source (σf_{kj}) measurement uncertainties. In its full form, the algorithm of Britt and Luecke