Atmospheric Environment 123 (2015) 240–250

Contents lists available at ScienceDirect

Atmospheric Environment

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

A new methodology to assess the performance and uncertainty of source apportionment models II: The results of two European intercomparison exercises

C.A. Belis [a, *], F. Karagulian [a], F. Amato [b], M. Almeida [c], P. Artaxo [d], D.C.S. Beddows [e], V. Bernardoni [f], M.C. Bove [g], S. Carbone [h], D. Cesari [i], D. Contini [i], E. Cuccia [g], E. Diapouli [j], K. Eleftheriadis [j], O. Favez [k], I. El Haddad [l], R.M. Harrison [e, m], S. Hellebust [n], J. Hovorka [o], E. Jang [e], H. Jorquera [p], T. Kammermeier [q], M. Karl [r], F. Lucarelli [s], D. Mooibroek [t], S. Nava [s], J.K. Nøjgaard [u], P. Paatero [v], M. Pandolfi [b], M.G. Perrone [w], J.E. Petit [k, z], A. Pietrodangelo [x], P. Pokorná [o], P. Prati [g, h], A.S.H. Prevot [l, m], U. Quass [q], X. Querol [b], D. Saraga [y], J. Sciare [z], A. Sfetsos [y], G. Valli [f, g], R. Vecchi [f, g], M. Vestenius [h, i], E. Yubero [aa], P.K. Hopke [ab]

[a] European Commission, Joint Research Centre, Institute for Environment and Sustainability, Via Enrico Fermi 2749, Ispra (VA) 21027, Italy
[b] Institute of Environmental Assessment and Water Research, Spanish Research Council (IDÆA-CSIC), c/Jordi Girona 18-26, 08034 Barcelona, Spain
[c] C2TN, Instituto Superior Técnico, Universidade de Lisboa, Estrada Nacional 10 km 139.7, 2695-066 Bobadela LRS, Portugal
[d] Instituto de Fisica, Universidade de Sao Paulo, Rua do Matao, Traversa R, 187 05508-900 Sao Paulo, Brazil
[e] Division of Environmental Health and Risk Management, School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom
[f] Dept. of Physics, Università degli Studi di Milano & INFN-Milan, via Celoria 16, Milan 20133, Italy
[g] University of Genoa, Dept. of Physics and INFN, via Dodecaneso 33, 14146 Genova, Italy
[h] Finnish Meteorological Institute, Atmospheric Composition Research, PO Box 503, FI-00101 Helsinki, Finland
[i] Istituto di Scienze dell'Atmosfera e del Clima, ISAC-CNR Str., Prv. Lecce-Monteroni km 1.2, 73100 Lecce, Italy
[j] Institute of Nuclear and Radiological Science & Technology, Energy & Safety, N.C.S.R. "Demokritos", 15341 Athens, Greece
[k] Institut National de l'Environnement Industriel et des Risques (INERIS), Verneuil-en-Halatte, France
[l] Laboratory of Atmospheric Chemistry (LAC), Paul Scherrer Institut, Villigen, Switzerland
[m] Department of Environmental Sciences/Center of Excellence in Environmental Studies, King Abdulaziz University, PO Box 80203, Jeddah 21589, Saudi Arabia
[n] Centre for Research into Atmospheric Chemistry, Dept. Chemistry, University College, Cork, Ireland
[o] Institute for Environmental Studies, Charles University in Prague, Albertov 6, 128 43 Prague 2, Czech Republic
[p] Departamento de Ingeniería Química y Bioprocesos, Pontificia Universidad Católica de Chile, Avda. Vicuña Mackenna 4860, Santiago 6904411, Chile
[q] IUTA e.V., Bereich Luftreinhaltung & Nachhaltige Nanotechnologie, Institut für Energie- und Umwelttechnik e.V., Bliersheimer Strasse 60, D-47229 Duisburg, Germany
[r] Urban Environment and Industry, Norwegian Institute for Air Research (NILU), PO Box 100, NO-2027 Kjeller, Norway
[s] Department of Physics and Astronomy and INFN, Firenze, Italy
[t] National Institute of Public Health and the Environment, Centre for Environmental Quality (MIL), Department for Air and Noise Analysis (ILG), PO Box 1, 3720 BA Bilthoven, The Netherlands
[u] Department for Environmental Science, Aarhus University, Frederiksborgvej 399, PO Box 358, DK-4000 Roskilde, Denmark
[v] Department of Physics, University of Helsinki, Rikalantie 6, FI-00970 Helsinki, Finland
[w] Department of Earth and Environmental Sciences, University of Milano-Bicocca, P.zza della Scienza 1, 20126 Milan, Italy
[x] C.N.R., Institute of Atmospheric Pollution Research, Area della Ricerca di Roma 1, Via Salaria Km 29,300, Monterotondo (RM) 00015, Italy
[y] IN.RA.S.T.E.S., NCSR Demokritos, P. Grigoriou and Neapoleos Str, 153 10 Agia Paraskevi, Greece
[z] CNRS LSCE, France
[aa] Laboratory of Atmospheric Pollution (LCA), Miguel Hernández University, Av. de la Universidad s/n, Edif. Alcudia, 03202 Elche, Spain
[ab] Center for Air Resources Engineering and Science, Clarkson University, Box 5708, Potsdam, NY 13699-5708, USA

* Corresponding author.
E-mail address: claudio.belis@jrc.ec.europa.eu (C.A. Belis).

http://dx.doi.org/10.1016/j.atmosenv.2015.10.068

HIGHLIGHTS

- Intercomparisons were carried out to test the performance and uncertainty of receptor models.
- More than 85% of the reported sources met the model quality objectives.
- Two thirds of the output uncertainties were coherent with those in the input data.
- PMF v2, v3 and CMB 8.2 estimated the source contributions satisfactorily.
- The accuracy of receptor models is in line with the needs of air quality management.

ARTICLE INFO

Article history:
Received 24 February 2015
Received in revised form
14 September 2015
Accepted 24 October 2015
Available online 3 November 2015

Keywords:
Source apportionment
Receptor models
Intercomparison exercise
Model performance indicators
Model uncertainty
Particulate matter

ABSTRACT

The performance and the uncertainty of receptor models (RMs) were assessed in intercomparison exercises employing real-world and synthetic input datasets. To that end, the results obtained by different practitioners using ten different RMs were compared with a reference. In order to explain the differences in the performances and uncertainties of the different approaches, the apportioned mass, the number of sources, the chemical profiles, the contribution-to-species and the time trends of the sources were all evaluated using the methodology described in Belis et al. (2015).

In this study, 87% of the 344 source contribution estimates (SCEs) reported by participants in 47 different source apportionment model results met the 50% standard uncertainty quality objective established for the performance test. In addition, 68% of the SCE uncertainties reported in the results were coherent with the analytical uncertainties in the input data.

The most used models, EPA-PMF v.3, PMF2 and EPA-CMB 8.2, presented quite satisfactory performances in the estimation of SCEs while unconstrained models, that do not account for the uncertainty in the input data (e.g. APCS and FA-MLRA), showed below average performance. Sources with well-defined chemical profiles and seasonal time trends, that make appreciable contributions (>10%), were those better quantified by the models while those with contributions to the PM mass close to 1% represented a challenge.

The results of the assessment indicate that RMs are capable of estimating the contribution of the major pollution source categories over a given time window with a level of accuracy that is in line with the needs of air quality management.

1. Introduction

Source Apportionment (SA) is the practice of deriving information about the pollution sources and the amount they contribute to measured concentrations. Receptor models (RMs) apportion the measured mass of pollutants to its emission sources by using multivariate analysis to solve a mass balance equation (Friedlander, 1973; Schauer et al., 1996; Thurston and Spengler, 1985). RMs derive information from measurements including estimations of their uncertainty and have been extensively used in Europe to estimate the contribution of emission sources to atmospheric pollution at a given site or area (Belis et al., 2013; Viana et al., 2008a). In the Chemical Mass Balance (CMB) approach, both chemical concentrations of pollutants, including their uncertainties, and chemical fingerprints of the sources (source profiles) are used as input. In the multivariate factor analytical approach (MFA), only environmental concentrations and uncertainties of pollutants are used as input data and the model computes the factor profiles and the mass contributed by the factors. The CMB approach is sensitive to the selection of sources, their stability and the collinearity among them. Differences between the methods used to analyse the source and ambient samples may also impact the results. On the other hand, MFA models identify factors that have to be attributed to emission sources. For a more thorough discussion about the pros and cons of the two approaches see Hopke (2010), Watson et al. (2008) and Belis et al. (2013).

Previous studies provided first estimates of the output variability by comparing the results of different RMs on the same dataset (Hopke et al., 2006; Larsen et al., 2008; Favez et al., 2010; Viana et al., 2008b; Pandolfi et al., 2008). In the present work, carried out in the frame of FAIRMODE (Forum for Air Quality Modelling), intercomparison exercises aimed at quantitatively assessing the performance and the uncertainty of RMs by comparing the results reported from different practitioners on the same dataset using different RM techniques.

2. Methodology

The methodology adopted in this research to assess the model results evaluates all the aspects of a source apportionment study, including the variability due to the influence of different practitioners using the same model on the same data (Belis et al., 2015). The procedure includes: complementary, preliminary and performance tests.

The “complementary tests” aim at providing ancillary information about the performance of the solutions in terms of apportioned mass and number of source categories. The “preliminary tests” are targeted at establishing whether the entities identified in the results, either a factor or a source (hereon, factor/source), are attributable to a given source category. In addition to the correlation coefficient (hereafter, Pearson), the standardized identity distance (SID), that prevents the distortions caused by source profiles with dominant species, is used (more details in Belis et al., 2015). The “ff tests” are the comparison among factor/sources attributed by participants to the same source category in all the solutions while “fr tests” refer to the comparison between reported factor/sources and a reference value. The objective of the “performance tests” is to evaluate whether the source contribution estimates (SCEs) are coherent with a 50% standard uncertainty target value using the z-score performance indicator complemented by the z′-score and zeta-score indicators (Thomson et al., 2006; ISO 13528, 2005). In this study, SCE denotes the mass attributed to a source

or factor in the results obtained with either CMB or MFA approaches. The methodology is fully described in the companion paper by Belis et al. (2015) and was implemented using the open source software R (and R-studio). Source categories with less than five factors/sources were not evaluated and profiles attributed by participants to more than one category were tested in each of the proposed categories.

Considering that source apportionment studies are mostly targeted at identifying and quantifying the typical sources in the studied area, the performance tests were conducted on the average SCE over the whole time window represented in every dataset. Moreover, the SCE time series were evaluated using the root mean square error normalised by the standard deviation/uncertainty of the reference value ($RMSE_u$), as discussed in Belis et al. (2015).

The intercomparison exercises were structured in two rounds involving 16 and 21 organizations respectively. In the first round, 22 results were reported and 25 were provided in the second one. A real-world $PM_{2.5}$ dataset collected in Saint Louis (USA) was used in Round 1 (Table 1). The dataset used for the intercomparison was developed by merging two datasets: one of inorganic species collected every day (Lee et al., 2006) and one of organic species collected every sixth day over the same time window (Jaeckels et al., 2007). In the final dataset, the structure of the uncertainties of the different species was heterogeneous with differences between species deriving from the data treatment in the original datasets and variability within single species due to the different analytical batches that were necessary to cover the whole monitoring campaign. In addition, the uncertainty of organic tracers was complex to quantify due to the possible influence of atmospheric chemistry and radiation on the degradation of these compounds (Galarneau, 2008; Hennigan et al., 2010).

The site and time window in which the real-world dataset was collected was not revealed to the intercomparison participants. The dataset containing the concentrations of 44 species in 180 samples with their analytical uncertainties was distributed to participants together with the analytical parameters (uncertainty of the method and minimum detection limits) and the emission inventory of the study area.

In Round 1, the following preliminary tests were performed: Pearson and SID between factor/source profiles, Pearson between log-transformed factor/source profiles, and Pearson between factor/source time trends. Only ff tests were accomplished in this round because of the absence of independent unbiased reference values.

In the performance tests of Round 1, the SCE reference value for each source category was the average of the results reported by the participants. The reference values were obtained by calculating the robust average (Analytical Methods Committee, 1989) using only the SCEs of source/factors that passed the preliminary tests (Table 2).

In the second round, a synthetic dataset with known reference values that were unbiased and independent from the results reported by participants was used (Supplementary Material S1). The chemical species included in the synthetic dataset (Round 2) are reported in Table 1 and the procedure followed to generate it is given in Belis et al. (2015).

Since the site was not disclosed to participants, the emission inventory of the study area and a set of 23 local source profiles (more than one for every source category) were distributed to them in order to: a) provide the necessary information to create the input files for CMB models, and b) support the interpretation of the models' output.

In addition to the preliminary tests performed in the previous round, the Pearson between the factor/source contribution-to-species of the Round 2 results was also computed. All of the preliminary tests were performed by comparing factor/sources reported by participants with the reference source for the considered source category (fr tests).

The model abbreviations used in this document are: CMB8.2, Chemical Mass Balance v. 8.2 by U.S. EPA; ME, Multilinear Engine; PCA, Principal Component Analysis; APCS, Absolute Principal Component Score; FA-MLRA, Factor Analysis-Multilinear Regression; COPREM, constrained physical receptor model and PMF, Positive Matrix Factorization. The code "PMF2″ denotes the program PMF2 described by Paatero (1997). The codes "EPAPMF3, EPAPMF4, and EPAPMF5″ denote the respective releases of the U.S. EPA program "EPA PMF".

3. Results and discussion

3.1. Complementary tests

3.1.1. Mass apportionment

The sample-wise comparison between the sum of the SCEs in every result and the gravimetric mass are summarised using normalised target diagrams (Fig. 1). More than 70% of the solutions in Round 1 rank in the area of acceptance (outer circle). Most scores rank in the lower quadrants indicating a tendency to underestimate the observed mass (the distance to the horizontal axis is proportional to the $PM_{2.5}$ mass that was not apportioned). On the contrary, the evident overestimation of the mass observed in two solutions is likely due to problems in the conversion of normalised data to concentration values rather than to errors in the apportionment of the mass. In Round 2, the majority of solutions (ca. 90%) rank in the

Table 1
Outline of the datasets used in every round of the intercomparison exercises.

	Round 1	Round 2
Type of data	Real-world dataset	Synthetic dataset
Site	Saint Louis (USA)	Milan (Italy)
Time window	June 2001–May 2003	January–December 2005
Pollutant	$PM_{2.5}$	$PM_{2.5}$
Number of samples	178, 24 h samples	364, 24 h samples
Number of chemical species	44	38
Carbonaceous species	OC/EC (steps)	OC/EC (total)
Ionic species	sulphate, nitrate, ammonium	sulphate, nitrate, ammonium, chloride
Elements	Al, As, Ca, Cr, Cu, Fe, K, Mn, Ni, Pb, Rb, Si, Sr, Ti, V, Zn[a]	
	Ba, Co, Hg, P, Se, Zr	Sb, Sn, Na, Mo, Cd, Mg
Organic species	indeno(cd)pyrene, benzo(ghi)perylene, benzo(a)pyrene, coronene, benzo(e)pyrene, dibenz[a,h]anthracene, levoglucosan[a]	
	benz(a)anthracene, fluoranthene, pyrene, benzo(b,k)fluoranthene, benzo(j) fluoranthene	chrysene, benzo(b)fluoranthene, benzo(k) fluoranthene

[a] The species in this line are common to both datasets.

Table 2
Source categories, codes and reference values used in every round of the intercomparison.

Round 1			Round 2		
Code	Source category	Reference SCE (μg/m^3)	Code	Source category	Reference SCE (μg/m^3)
BioB	Biomass burning/wood burning	1.59	BioB	Biomass burning/wood burning	4.33
BRA	Road dust/brake abrasion	0.83	SO4	Ammonium sulphate	7.12
COPPER	Copper production	0.57	NO3	Ammonium nitrate	12.69
DIE	Diesel vehicles	0.42	DUST	Soil dust/crustal	4.01
DUST	Soil dust/crustal	0.74	ROAD	Road dust	2.68
GAS	Gasoline vehicles	0.59	SALT	Sea salt/road salting	0.52
INDU	Industrial emissions/combustion	1.07	TRA	Exhaust emission from vehicles	6.63
LEAD	Lead smelter	0.42	INDU	Industrial emissions/point sources	5.11
NO3	Ammonium nitrate	2.98			
SEC	Secondary aerosol	6.36			
SHIP	Ship emissions	1.63			
SO4	Ammonium sulphate	5.99			
STEEL	Steel processing	1.57			
TRA	Traffic exhaust	2.44			
ZINC	Zinc smelter	0.58			

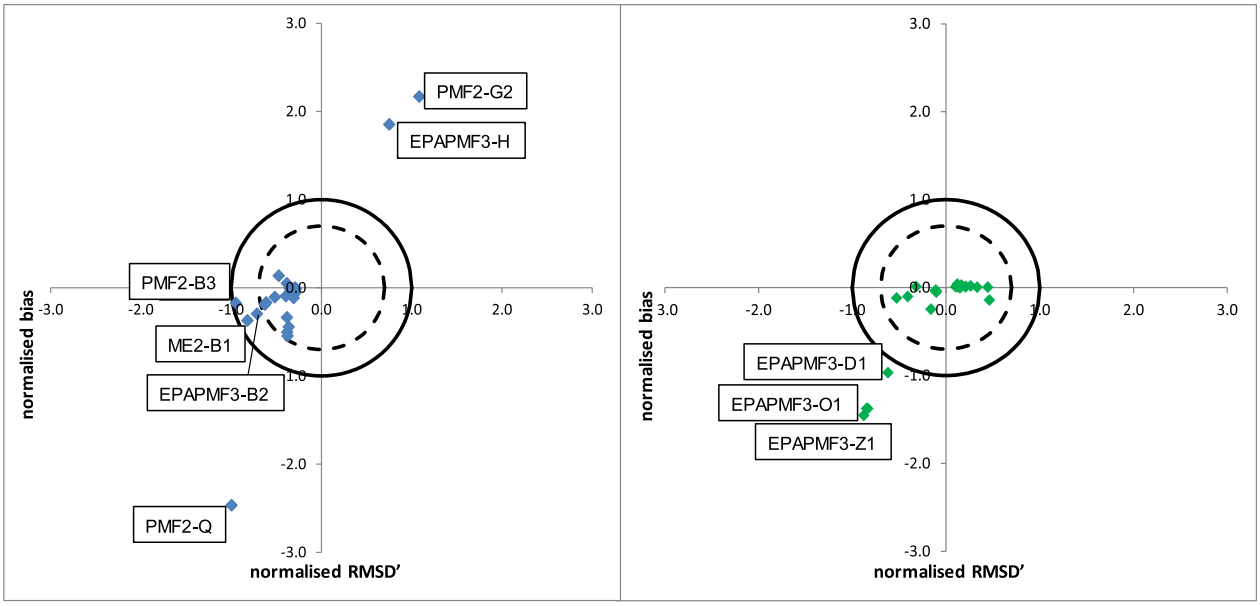

Fig. 1. Target diagrams summarizing the mass apportionment in the first (left) and second (right) rounds. The outer circle delimits the acceptance area and the inner circle represents the boundary of scores with Pearson equal to 0.7. Only scores outside the inner circle are labelled with the model abbreviation and solution code. RMSD': unbiased root mean square difference (Jolliff et al., 2009).

area of acceptance and show little bias indicating that many solutions achieved a quite satisfactory apportionment of the gravimetric mass to its sources. In these tests, no clear relation between the type of model used and the performance is observed.

3.1.2. Number of factor/sources

There are different techniques to determine the number of sources (e.g. Henry et al., 1984). The procedures followed by participants to determine the number of sources were based on multicriteria, the most common of which were: a) the impact of the number of factors on the model diagnostics, b) the stability of factor profiles across different models set up, and c) the physical meaning of the factor profiles and their comparability with source profiles from the literature.

In Round 1, nine factor/sources per solution are reported on the average (Table 3). One half of the solutions identifies between six and ten factor/sources while six solutions report more than 10. An approximation of the expected number of factor/sources for this round is derived from the original solution of the inorganic dataset obtained using PMF (Lee et al., 2006), which identified 10 different source categories. In this round, the estimations of PMF and CMB are relatively close. In Round 2, more than half of the solutions report the exact number of factor/sources used to design the dataset (8) and all the solutions, except one, report between six and nine factor/sources.

The tests suggest that the reliability of the performance diagnostics influence the ability of the tools to establish the most suitable number of factor/sources. Often, unconstrained MFA tools rank far from the average. The higher number of factor/sources in COPREM is likely due to the attempt to apportion the secondary organic aerosols (not present in the synthetic dataset) and the split of ammonium sulphate into $(NH_4)_2SO_4$ and $(NH_4)HSO_4$.

No relevant differences in the number of factor/sources are observed between CMB8.2 and the different versions of PMF.

3.2. Identity and uncertainty of the factor/sources

3.2.1. Factor/source identity

3.2.1.1. Chemical profiles. Fig. 2 shows the distribution of the Pearson and SID values used for comparing the chemical profile of

Table 3
Average number of reported factor/sources by model.

Model	Round average	CMB8.2	PMF2	EPA PMF3	EPA PMF4	EPA PMF5	ME-2	COPREM	PCA	APCS	FA-MLRA	Reference
Round 1	9	8	9	9	–	–	6	13	7	11	–	10[a]
Round 2	9	8	8	7	7	8	8	13	–	–	6	8

[a] Indicative reference.

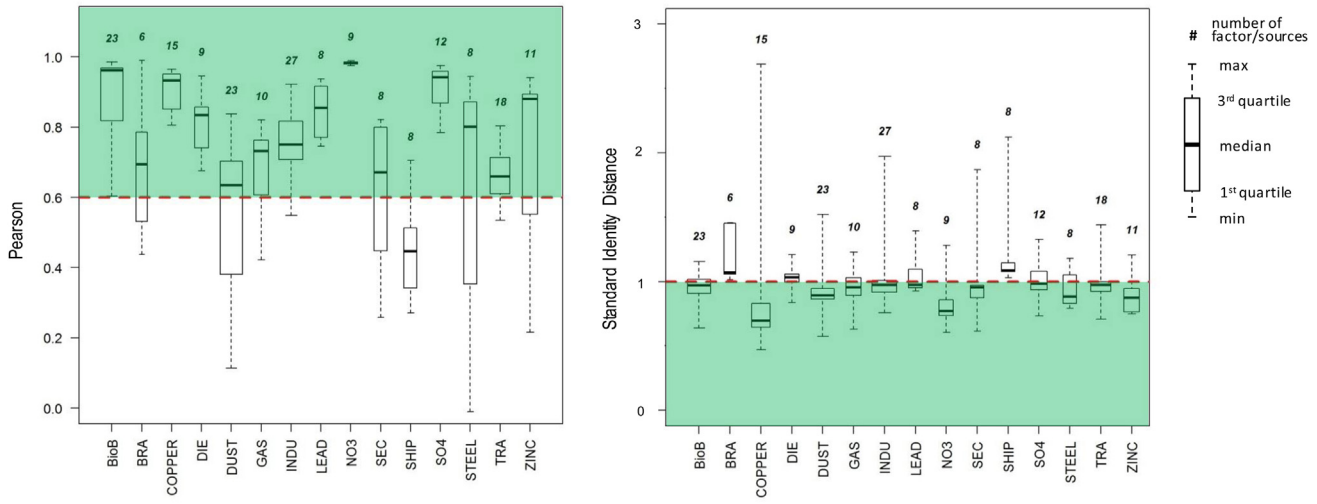

Fig. 2. Similarity of factor/source chemical profiles in each source category (ff tests) in Round 1 calculated using Pearson (left) and SID (right). Pearson: values above the broken line rank in the area of acceptance. SID: accepted values are those below the broken line. The number of tested factor/sources is reported on top of each bar.

each factor/source to all of the others attributed by practitioners to the same source category (ff tests) in Round 1. More than 75% of the Pearson values are above the limit of acceptance (broken line), indicating that the majority of the source categories present relatively comparable chemical compositions. The most heterogeneous categories (SHIP, BRA, DUST, SEC, STEEL and ZINC) show between 25% and 75% of factor/sources in the rejection area.

In this step, the number of factor/sources passing the SID test is, in the majority of cases, lower than those passing the Pearson. Therefore, there are more categories with profiles in the rejection area (e.g. DIE and LEAD).

Considering the two indicators, SHIP and BRA are amongst the most heterogeneous categories. The dissimilarities observed within SHIP are likely due to the variety of chemical profiles allocated to this source category in the reported solutions. Due to similar fuel and combustion conditions, SHIP source profiles may be difficult to distinguish from stationary sources such as energy plants, oil refineries and other industrial processes (Viana et al., 2014). Only six profiles were attributed to the heterogeneous category BRA. Some of them, obtained with unconstrained factor analysis (APCS), are of difficult interpretation due to the extremely high concentration of Ca or the absence of Ba.

In Round 2, Pearson and SID tests point out SALT and TRA as categories where a discrete number of chemical profiles diverge from the reference (Fig. 3; see discussion in sections 3.2.1.2 and 3.2.1.3). In addition, Pearson test highlights also factor/sources in INDU as poorly comparable to their reference source chemical profile. This source category is, by definition, quite heterogeneous considering that it includes factor/sources attributed to different types of industries, combustion processes, without excluding regional (secondary) aerosol. Because of their simple chemical composition, SO4 and NO3 are the source categories in which factor/source profiles resemble more the reference profile in the Pearson tests. Nevertheless, these source categories are much less homogeneous when tested using SID, which gives more weight to minor components in the factor/source profiles. This may indicate there are different sources of precursors associated to these secondary compounds.

The very limited changes observed in the Pearson values with log-transformed data in the two steps suggest that this kind of transformation is not solving efficiently the problem of dominant species in the profiles. For a more detailed discussion about the indicators of similarity see the companion paper by Belis et al. (2015).

The correlation (Pearson) between factor/sources identified in Round 1, on the basis of their time series, is summarized in Fig. 4 (left). The time series of BioB, COPPER, LEAD, NO3 and ZINC are quite comparable among the different reported results. For the industrial sources, the time correlation is attributed to the effect of the intermittent pattern determined by the changes in wind direction and the time windows in which the emitting facility was in operation. Other sources, such as BioB and NO3, are synchronous due to common seasonal patterns determined by the trends in the emission rates and in atmospheric variables (e.g. air temperature, thermal inversion).

Factor/sources in the categories BRA, DIE, INDU, SEC, SHIP, and TRA display different temporal patterns. Most of these sources show also medium to poor correlation among the different chemical profiles (Fig. 2). The poor time correlations in factor/sources of the categories TRA, DIE and GAS may, at least in part, be connected with the time resolution of the data used for Round 1. One sample every sixth day may not be optimal to capture a sufficient number of weekends to show the week day/weekend patterns.

In Round 2, the time trends of the factor/sources are quite comparable with the reference for the majority of the source categories.

Despite the good correlations among the reported chemical profiles, likely determined by the presence of a combination of

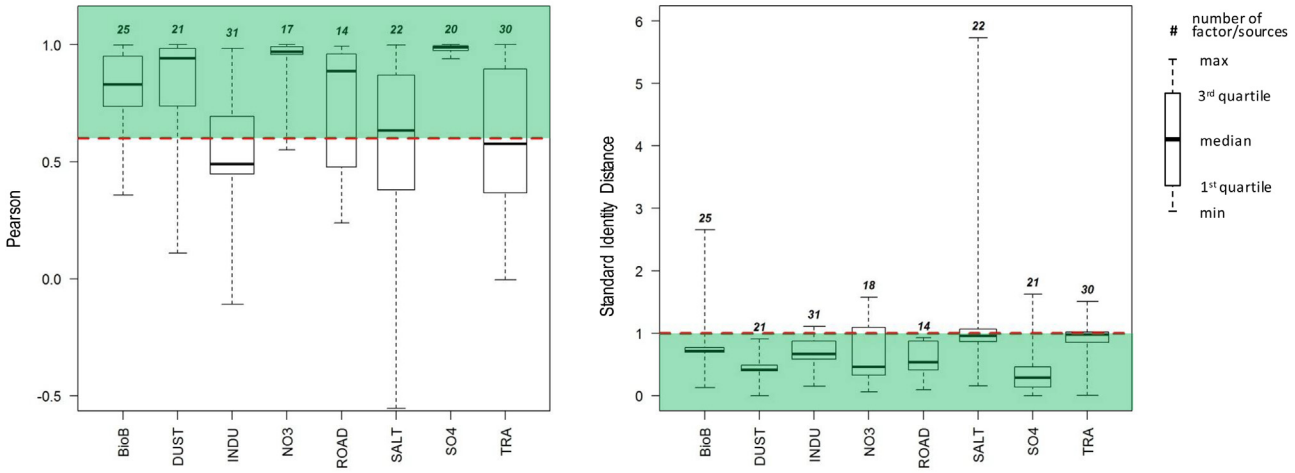

Fig. 3. Comparison of factor/source chemical profiles with the reference profile for every source category (fr tests) in Round 2 calculated using Pearson (left) and SID (right). Pearson: values above the broken line rank in the area of acceptance. SID: accepted values are those below the broken line. The number of tested factor/sources is reported on top of each bar.

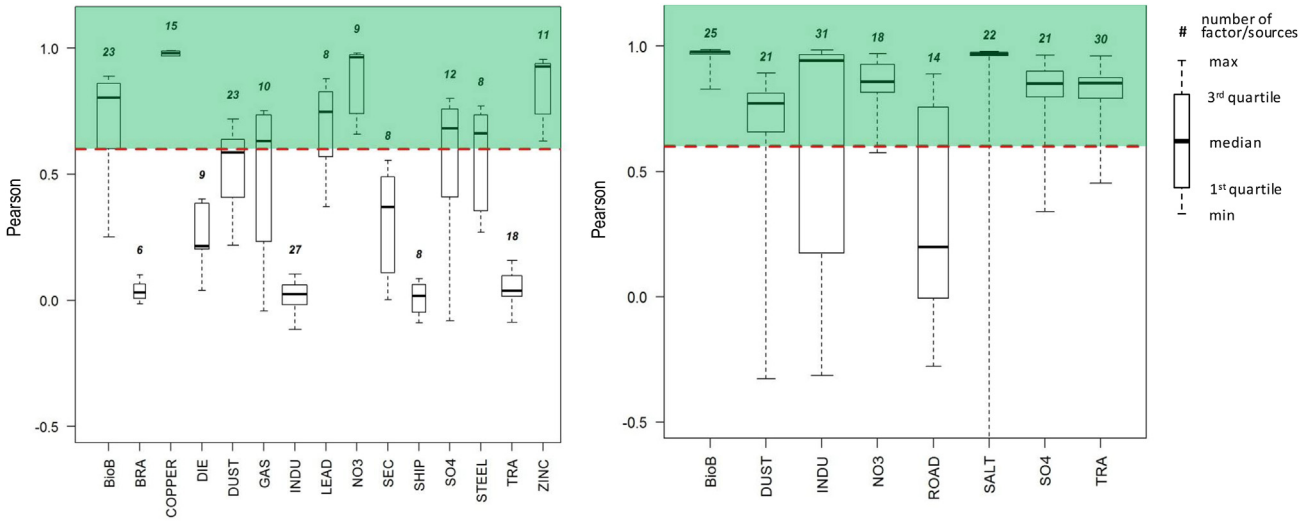

Fig. 4. Comparison of factor/source time series in Round1 (ff tests, left) and in Round 2 (fr tests, right) using Pearson. Values above the broken line rank in the area of acceptance. The number of tested factor/sources is reported on top of each bar.

organic carbon and characteristic trace elements (e.g. Cu, Sn and Cr), ROAD is the source category with the lowest correlation between the reported time trends and the reference. This has been interpreted as the influence, to varying extents in each solution, of elements like Si, Al, and Mg that are also typical of DUST profiles and that may blur the boundary between these two categories. Also INDU shows quite variable results in this test and the considerations made for Round 1 are valid also in this case.

Source categories with inhomogeneous chemical profiles, such as INDU, often present poorly correlated time trends suggesting that an imperfect separation and identification of the sources leads to a poor fit in both the chemical composition and the temporal pattern. Nevertheless, this general rule is not always valid. For instance, the time trends of SALT in Round 2 are quite comparable (Fig. 4) even though the chemical profiles of the factor/sources attributed to it are not homogeneous (Fig. 3). This apparent contradiction is explained by the high variance between the SALT time trends in the different reported results that is not detected by the Pearson test because the oscillations are synchronous.

3.2.1.2. Contribution-to-species. The contributions of sources to the mass of every single species in the dataset expressed as percentage (contribution-to-species) were reported only in Round 2 (Fig. 5). The results reported in the different solutions are quite comparable among each other and with the reference source. As already observed in the tests for chemical profiles, INDU and ROAD show a number of records in the action area. Also the factor/sources in NO3, that are comparable with the reference in terms of time trend, show a non-negligible share of scores in the action area. In this category, the lower scores observed in the contribution-to-species may be attributed to the lower influence of dominating species, like ammonium nitrate, and higher influence of minor species such as Ca, As, Mo, Rb, Cl and PAHs.

On the other hand, factor/sources in the SALT category, which show poor correlation with the concentrations in the reference profile, are well correlated with the reference in terms of contribution-to-species. In the SALT chemical profiles, Cl and Na represent on average 81% and 49% of the source mass, respectively, and their relationship is close to the stoichiometric ratio in sodium

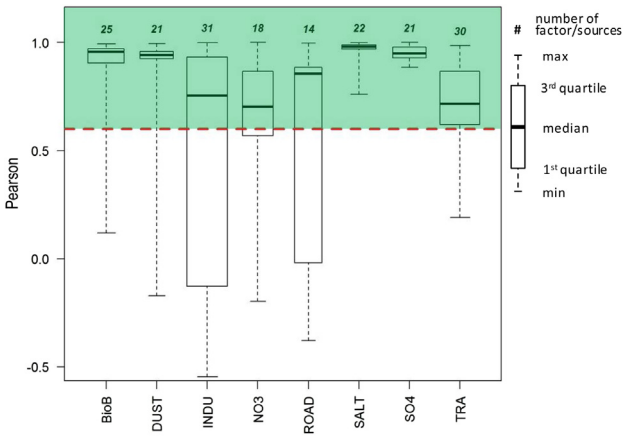

Fig. 5. Comparison of factor/source contribution-to-species with the reference profile for every source category (fr tests) in Round 2. Values above the broken line rank in the area of acceptance. The number of tested factor/sources is reported on top of each bar.

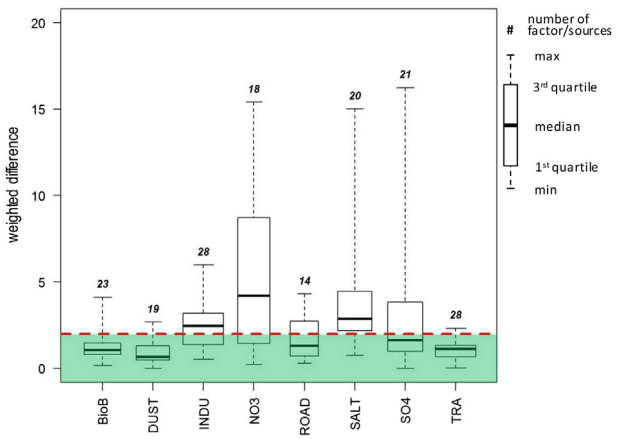

Fig. 6. Evaluation of chemical profiles uncertainties, using the weighted difference (WD) indicator in Round 2 (fr tests). Values below the broken line rank in the area of acceptance. The number of tested factor/sources is reported on top of each bar.

chloride. As for the contribution-to-species, the ratio between the two elements (39% and 58% of the SALT mass, respectively) indicates that the share of Cl in SALT is lower than the one it would have been if the only source consisted of NaCl. This mismatch indicates the contribution of additional sources to this element other than sea and road salt (e.g. INDU).

3.2.2. Chemical profile uncertainty

In order to assess the uncertainty of the factor/source profiles, the weighted differences (WD, Karagulian and Belis, 2012) between the source profiles reported by participants and the corresponding reference profiles were computed.

The interpretation of WD scores depends on the relevance of the reference value for the factor/sources being tested. If a factor/source has been attributed to the wrong source category, the reference is not appropriate to evaluate that factor/source. For that reason, WD are interpreted by taking into account the results of the chemical profile tests (see section 3.2.1.1).

In Round 1, the fr tests were carried out using external reference profiles available in the literature and are, therefore, used only for informative purposes (not reported).

The WD test shows that, in Round 2, SALT is the category with the highest proportion of scores outside the area of acceptance (above the broken line) followed by NO3, INDU, SO4 and ROAD (Fig. 6). The analysis of the chemical profile's uncertainty using the WD indicator shows that, in this round, 65% of factor/sources present acceptable WD scores. In addition, the joint evaluation with the chemical profile test suggests that only 18% of the factor/source profiles, which allocation to source categories was confirmed, underestimated their uncertainty.

3.3. Performance tests

In this section the results of the tests aiming at evaluating the SCEs, the most important output of a source apportionment study, are presented. The assessment of the SCE time trends is discussed in the companion paper by Belis et al. (2015).

3.3.1. Reported source contribution estimates

The distributions of the SCEs reported by participants in Round 1 and 2 are shown in Supplementary Material S2. The coefficients of variation (CVs) of the SCE reported by participants for every source category are, on average, 0.77 and 0.45 in the first and second round, respectively. NO3 and SO4 are the source categories with the lowest CV (between 0.26 and 0.48). In Round 1, CVs higher than the unity are observed in DUST, SHIP, INDU and ZINC while GAS, DIESEL and BRA show values in the range 0.80–1.00. In Round 2, the SCEs are higher, because of the higher PM levels, and their relative variability within source categories is lower than in Round 1. The highest CV is the one of SALT (0.70) followed by DUST and INDU (0.60 and 0.55. respectively). As in Round 1, the lowest CVs are those in SO4 and NO3 (0.28 and 0.31, respectively).

3.3.2. Z-scores

Fig. 7 summarises the z-scores assigned to each factor/source reported by participants in Round 1. The z-scores are in the acceptance area 85% of the time, 3% in the warning area, and 12% in the action area. The majority of solutions, 19 out of 22, present at least 75% of the scores in the acceptance area. Only solution G2 presents the majority of scores in the action area. Such performance is likely due to the problems in mass quantification highlighted in the complementary tests (section 3.1.1).

DUST is the source category with the highest variability and the highest number of scores in the action area due to overestimation (6 scores) while SHIP and BRA are the ones with the highest number of scores in the action area due to underestimation (4 and 2 scores, respectively). Source categories DIE, GAS, BIOB, INDU and ZINC present three or less profiles with scores in the upper action area each. Inaccuracy in the SCE estimation of DUST, SHIP and BRA have been associated with the lack of homogeneity in the chemical profiles of the source factors attributed to them, as pointed out in the preliminary tests. Alternatively, those factors/sources with poor scores in DIE and GAS are likely connected to results affected by the limited number of weekend days included in the dataset, as indicated by the preliminary test on time trends. The few z-scores of INDU ranking in the action area may be associated with divergences in both time trends and chemical profiles.

In Round 1, about 80% of the reported factor/sources were obtained either with EPAPMF3, PMF2 or CMB8.2. In each of these models, more than 80% of the z-scores are placed in the area of acceptance. Interpretation of the results of the other models should be made with caution due to the limited number of reported solutions obtained with them.

An 89% of the z-scores assigned to factor/sources reported by participants in Round 2 are in the acceptance area, while 2% and 9% are in the warning and action areas, respectively (Fig. 8). The majority of solutions, 21 out of 25, had more than 75% of the scores in the acceptance area.

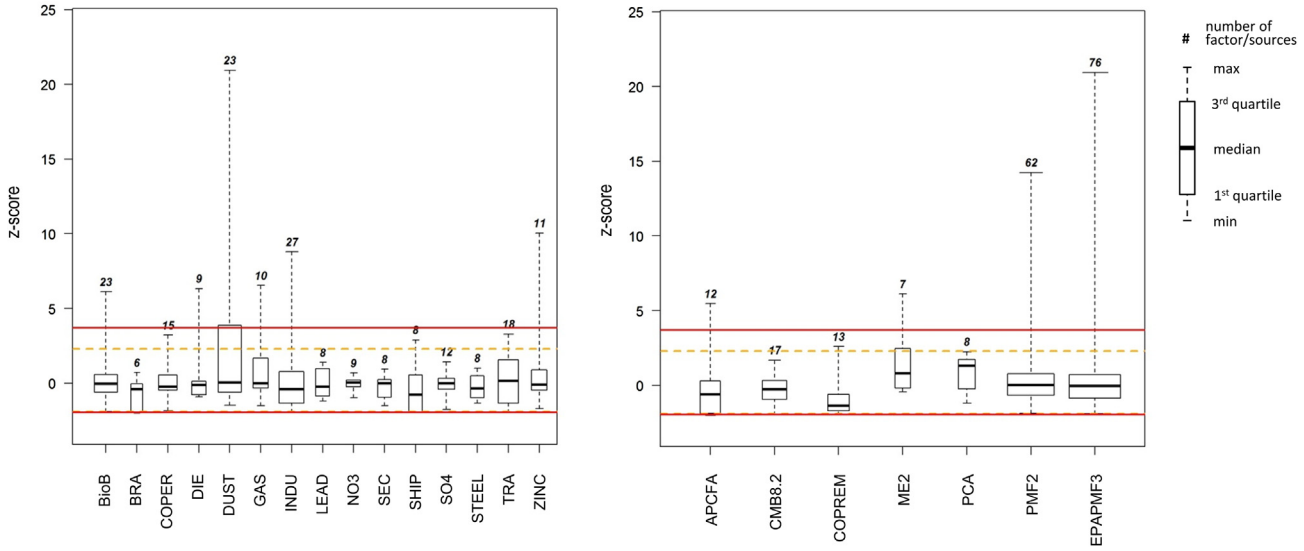

Fig. 7. Z-scores attributed to the factor/profiles in Round 1 arranged by source category (left) and by model (right). Scores outside the zone between continuous lines rank in the action area, those in the space between the continuous and the broken lines rank in the warning area and those in the zone within the broken lines rank in the acceptance area. The number of tested factor/sources is reported on top of each bar.

SALT is the only source category with more than half of the scores in the action area. The overestimation of the SALT SCEs in the majority of solutions is likely due to the small contribution of this source category, which represents only 1% of the total PM mass. These low-contributing factors are likely to be severely affected by the remaining ambiguity derived from scaling indeterminacy. Their contributions and composition could be underestimated/overestimated by a large unknown coefficient (Amato et al., 2009). The negative SCE reported in a result obtained with FA-MLRA also contributed to the poor performance in this source category and further highlights the limitations of fully unconstrained factor analytical methods. A common drawback of tools without non-negativity constraints is the attribution of negative SCEs to minor sources to compensate the excess of mass attributed to others.

As in Round 1, INDU shows some z-scores ranking either in the warning or in the action areas. The performance of this source category in the two rounds is likely caused by the poor match in the chemical composition and time trends between the factor/sources reported in the solutions and the reference values. A limited degree of overestimation is also observed in ROAD, as shown by one of the scores in the action area. As discussed in Section 3.2.1.2, this can be attributed to the interference of DUST, especially during windy days, that may also lead to inaccuracies in the time trends. A propensity to underestimate source categories with high SCEs such as NO3 and to a lesser extent SO4 (29% and 17% of the PM mass, respectively) is present in many solutions. Nevertheless, the bias is too small to give rise to poor scores.

In Round 2, about 75% of the reported SCEs derive from solutions obtained with EPAPMF3, PMF2 and CMB8.2 and their performances are comparable to those observed in Round 1. Although a limited

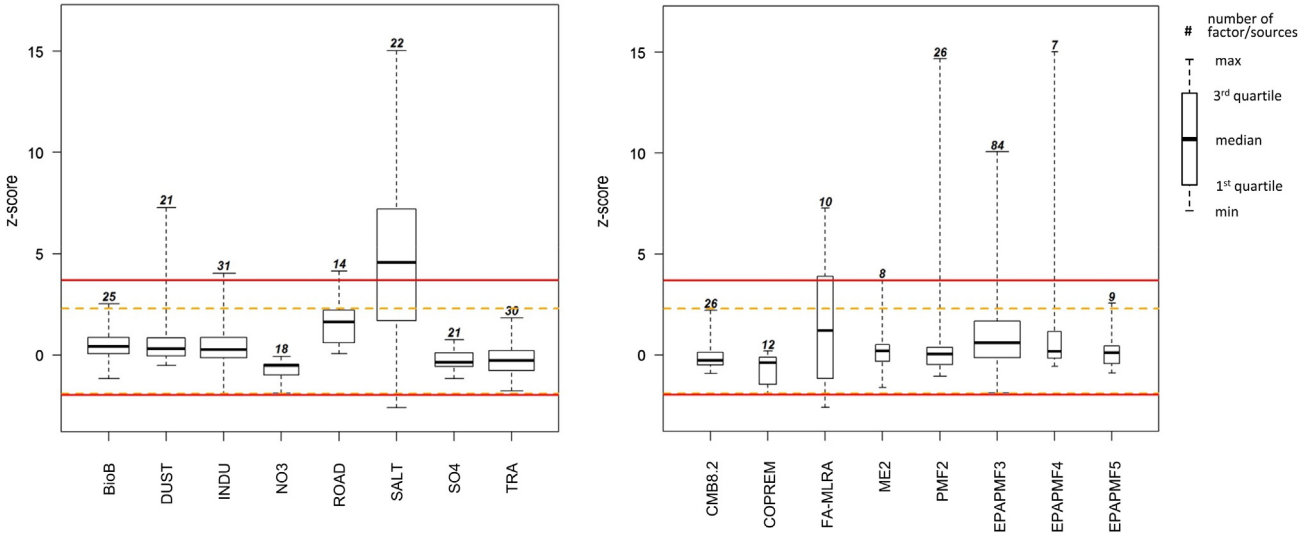

Fig. 8. Z-scores attributed to the factor/sources in Round 2 arranged by source category (left) and by model (right). Scores outside the zone between continuous lines rank in the action area, those in the space between the continuous and the broken lines rank in the warning area and those in the zone within the broken lines rank in the acceptance area. The number of tested factor/sources is reported on top of each bar.

number of solutions are available for the other models, it is worth mentioning the good performances of COPREM, EPAPMF5 and ME-2. FA-MLRA is the only model with 50% of the scores either in the warning or action areas.

The z′-score indicator was used in Round 2 to assess the difference between solutions and the reference value taking into account the reference's uncertainty. No substantial differences were observed between z-scores and z′-scores indicating that the uncertainty of the reference had no impact on the evaluation of participant's performance.

3.3.3. The uncertainty of the source contribution estimates

In source apportionment modelling, there are different sources of error: random error, modelling error (bias), and rotational ambiguity (Paatero et al., 2013). One important source of random error is the one present in the input data and is commonly approximated from their analytical uncertainty. Modelling error arises in situations in which the RM assumptions (Belis et al., 2013) are seriously infringed. It may derive from wrong number of sources or variation of sources in time and is mostly contributing to the bias kind of error. Also atmospheric composition and meteorology acting selectively on the degradation of organic tracers (Galarneau, 2008) are a component of the bias error.

Many RM tools supply the output uncertainty. In EPA PMF versions, the uncertainty of the output profiles is estimated using resampling and more recently also with displacement methods while the CMB EPA 8.2 model performs a propagation of the input analytical uncertainty. Many practitioners using non-US EPA tools compute the output uncertainty with resampling and error propagation techniques in post-processing. The rotational ambiguity is not discussed in this section because only one of the used of tools (EPA PMF v5) was designed to estimate this kind of uncertainty. More discussion about the uncertainty test can be found in the companion paper by Belis et al. (2015).

The tests described in the previous sections were mostly oriented to assess: a) the bias by comparison with a reference value and b) the reproducibility intended as the range of results that can be obtained from a single dataset (with a given degree of noise) by different practitioners using the same or different tools. In the following, the analysis will focus on the assessment of the SCEs uncertainty estimation accomplished by RMs by comparing them with the one of the reference. Considering that unbiased reference values are available only for the synthetic dataset, in this section are discussed only the results of Round 2.

The mean of the reported relative standard uncertainties for the SCE of the whole time window in Round 2 is 13%. The lowest values are those in NO3 source categories and the highest are those in INDU. As for the models, the lowest uncertainties are those reported in ME-2 and CMB8.2 solutions and the highest are those of COPREM solutions. No uncertainty was reported for the SCEs obtained with FA-MLRA. The uncertainty attributed to the reference was equivalent to the noise introduced in the synthetic dataset (20% standard deviation) that was derived from the analytical uncertainty in the input dataset (Belis et al., 2015). The zeta-score test indicates that a 68% of the declared factor/source SCE uncertainties are coherent with the one of the reference while a 19%, ranking in the action area, are likely underestimated (Fig. 9).

SALT is the only source category with the majority of the zeta-scores in the action area (75%). Likely, models do not allow for the higher relative uncertainty due the very low SCEs in this source category compared to the others. Uncertainty underestimation is observed also in ROAD, which shows a 60% of the scores either in the warning or in the action areas.

A considerable proportion of factor/sources obtained with EPAPMF4 and EPAPMF3 show underestimated uncertainties (29% and 24% of scores in the action area, respectively). COPREM showed uncertainties higher than the reference in a 31% of the factor/sources. The satisfactory performance of CMB8.2 (more than 90% successful scores) suggests that propagating the uncertainty of the source profiles can provide a satisfactory estimation of the SCEs uncertainty.

3.3.4. The impact of the operator

The variability between solutions obtained by different practitioners using the same tool and the same input data are an indicator of the maximum impact of the operator subjectivity on the reproducibility. The tools with the highest number of reported solutions: EPAPMF3, PMF2, and CMB8.2 present a high consistency among solutions obtained by different practitioners using the same tool. The standard deviations of the SCE mean in each of these models ranges between 0.2 and 0.3 μg/m^3 and 1.4–1.7 μg/m^3, in the first and second rounds, respectively. These values are, in addition, close to the standard deviation of the overall mean (0.2 μg/m^3 and 1.7 μg/m^3, in the first and second rounds, respectively). These results suggest a limited impact of the practitioners' subjectivity, on average. However, “outliers” were often associated with less experienced practitioners in terms of both years of use of the tool and number of studies performed.

4. Key findings of the intercomparison

The tests on chemical profiles confirmed, in the majority of cases (83%), the attribution of factors/profiles to source categories in the reported results and the majority of the SCEs (87%) reported by participants met the 50% standard uncertainty quality objective established for the performance test. A high share of the tested solutions (70%–80%) apportioned a considerable amount of the $PM_{2.5}$ mass to its pollution sources and many solutions estimated a number of sources close to the expected value.

In this study, the estimation of source contribution was most critical for SALT, DUST, SHIP and categories associated with mobile sources. The majority of the solutions overestimated the SCE of SALT, a source category with a contribution of about 1% of the PM mass. Such relative contribution may be considered a first approximation of the lower limit that the tested methodologies are able to quantify. Poor scores attributed to some DUST and ROAD SCEs were ascribed to the similarities in the chemical composition between road dust and crustal material that may have interfered with the allocation of mass between these sources. The uncorrelated time trends and, in some cases, the heterogeneous chemical profiles observed in INDU and SHIP were attributed to the lack of a common definition of these categories. Sources with appreciable contributions and chemical profiles dominated by few species, such as NO3 and SO4, were more efficiently recognised by the models even though there was a tendency to slightly underestimate their SCEs.

The most commonly used models, EPAPMF3, PMF2, and CMB8.2 showed quite satisfactory performance with successful z-scores ranging between 80% and 100%. The good agreement between CMB and PMF may be partially due to the main RM assumptions being substantially respected in the used datasets: limited alteration of the species between source and receptor and relatively stable source profiles. In addition, both types of tools account for the uncertainties in the input data, have built-in performance indicators and have been available long enough to allow a wide number of practitioners be familiar with them. For those models used in a limited number of solutions, only preliminary conclusions can be drawn at this stage. In general, fully unconstrained models which do not account for the input data uncertainty (e.g. FA-MLRA and APCFA) showed performances below the average. This result is

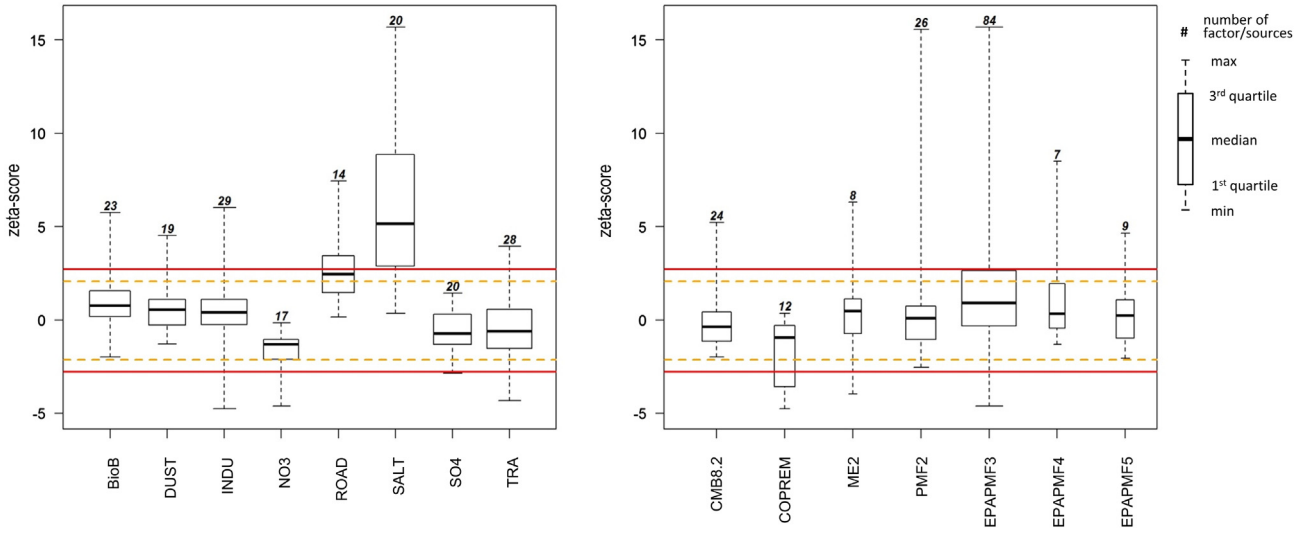

Fig. 9. Zeta-scores attributed to the factor/sources in Round 2 arranged by source category (left) and by model (right). Scores ranking above or below the continuous lines are in the action area, those in the space between the continuous and the broken lines are in the warning area and those in the zone within the broken lines are in the acceptance area. The number of tested factor/sources is reported on top of each bar.

likely because in these tools, the noise deriving from the uncertainty structure of the datasets is incorporated into the factor/sources (Paatero and Hopke, 2003).

The tests used to assess the SCE uncertainty reported in the solutions confirmed that the RMs output uncertainty estimation is coherent with the analytical/random uncertainty of the input data. Other components of the uncertainty could be evaluated in specially designed intercomparisons where RMs are either compared with other types of models or synthetic datasets with known perturbing factors are used. Processes altering the factor/source chemical profiles could be detected in the preliminary tests by comparison with the reference source profiles. In addition, diagnostic ratios could be used to detect long-range transport or photochemical age of aerosols (Hien et al., 2004; Decarlo et al., 2010).

The slightly better performance observed in Round 2 compared to Round 1 is likely connected to the differences between simulated and measured data. Round 1 was more challenging for the participants due to the inconsistencies in the uncertainties they had to deal with in a blind test with limited information about a non-European study area. On the contrary, the synthetic dataset contained internally consistent data with a lower level of noise and fewer source categories.

In the real-world, the variability of profiles in time and the chemical reactivity of organic species may affect the source/receptor relationships. Datasets from areas with complex atmospheric transport and chemistry are likely more challenging for models to quantify the sources (especially secondary and/or distant ones) than areas influenced mainly by local sources. In this study, there are no indications that the variability of profiles and degradation of markers affected the comparability of results among participants working on the same dataset. On the other hand, it was observed that the time resolution of the datasets influenced the ability of RMs to capture the time patterns of mobile sources.

5. Conclusions

The results of this study indicate that RMs are capable of estimating the contributions of the main pollution source categories within a given time window with a level of accuracy that is in line with the needs of air quality management.

Further intercomparisons evaluated with the same or comparable methodologies are needed to create a weight-of-evidence about the characteristics and capabilities of the models and tools.

Future work to improve the capacity of these models should focus on: a) the development and availability of source profiles relevant for the study area, b) better definition of the source categories, c) experimental design to improve the uncertainty estimation, d) development of speciated PM data series with appropriate time resolution and extended set of markers.

Moreover, the implementation of the existing common guidelines (Belis et al., 2014) would lead to more comparable results with recognised quality standards in line with those reported in the present work.

Acknowledgements

The authors are grateful to J. Schauer, and R. Turner for making available the organic dataset collected in St. Louis for the first round of the intercomparison. We thank C. Samara and G. Argyropoulos for sharing their data elaborations and for the fruitful discussions during this work.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.atmosenv.2015.10.068.

References

Amato, F., Pandolfi, M., Escrig, A., Querol, X., Alastuey, A., Pey, J., Perez, N., Hopke, P.K., 2009. Quantifying road dust resuspension in urban environment by Multilinear Engine: a comparison with PMF2. Atmos. Environ. 43, 2770–2780.

Analytical Methods Committee, 1989. Robust statistics – how not to reject outliers. Part 1: basic Concepts. Analyst 114, 1697–1698.

Belis, C.A., Karagulian, F., Larsen, B.R., Hopke, P.K., 2013. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. Atmos. Environ. 69, 94–108.

Belis, C.A., Pernigotti, D., Karagulian, F., Pirovano, G., Larsen, B.R., Gerboles, M., Hopke, P.K., 2015. A new methodology to Assess the Performance and Uncertainty of Source Apportionment Models. Atmos. Environ. 119, 35–44.

Belis, C.A., Larsen, B., Amato, F., El Haddad, I., Favez, O., Harrison, R.M., Hopke, P.K., Nava, S., Paatero, P., Prévot, A., Quass, U., Vecchi, R., Viana, M., 2014. European Guide on Air Pollution Source Apportionment with Receptor Models.

Publication Office of the European Union, Italy, p. 88.
Decarlo, P.F., Ulbrich, I.M., Crounse, J., De Foy, B., Dunlea, E.J., Aiken, A.C., Knapp, D., Weinheimer, A.J., Campos, T., Wennberg, P.O., Jimenez, J.L., 2010. Investigation of the sources and processing of organic aerosol over the Central Mexican Plateau from aircraft measurements during MILAGRO. Atmos. Chem. Phys. 10, 5257–5280.
Favez, O., El Haddad, I., Piot, C., Boréave, A., Abidi, E., Marchand, N., Jaffrezo, J.L., Besombes, J.L., Personnaz, M.B., Sciare, J., Wortham, H., George, C., D'anna, B., 2010. Inter-comparison of source apportionment models for the estimation of wood burning aerosols during wintertime in an Alpine city (Grenoble, France). Atmos. Chem. Phys. 10, 5295–5314.
Friedlander, S.K., 1973. Chemical element balances and identification of air pollution sources. Environ. Sci. Technol. 7, 235–240.
Galarneau, E., 2008. Source specificity and atmospheric processing of airborne PAHs: implications for source apportionment. Atmos. Environ. 42, 8139–8149.
Hennigan, C.J., Sullivan, A.P., Collett Jr., J.L., Robinson, A.L., 2010. Levoglucosan stability in biomass burning particles exposed to hydroxyl radicals. Geophys. Res. Lett. 37 (L09806), 4.
Henry, R.C., Lewis, C.W., Hopke, P.K., Williamson, H.J., 1984. Review of receptor model fundamentals. Atmos. Environ. 18, 1507–1517.
Hien, P.D., Bac, V.T., Thinh, N.T.H., 2004. PMF receptor modelling of fine and coarse PM10 in air masses governing monsoon conditions in Hanoi, northern Vietnam. Atmos. Environ. 38, 189–201.
Hopke, P.K., 2010. The application of receptor modeling to air quality data. Pollut. Atmos. 91–109.
Hopke, P.K., Ito, K., Mar, T., Christensen, W.F., Eatough, D.J., Henry, R.C., Kim, E., Laden, F., Lall, R., Larson, T.V., Liu, H., Neas, L., Pinto, J., Stölzel, M., Suh, H., Paatero, P., Thurston, G.D., 2006. PM source apportionment and health effects: 1. Intercomparison of source apportionment results. J. Expo. Sci. Environ. Epidemiol. 16, 275–286.
ISO 13528, 2005. Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparisons. (ISO) International Organization for Standardization.
Jaeckels, J.M., Bae, M.S., Schauer, J.J., 2007. Positive matrix factorization (PMF) analysis of molecular marker measurements to quantify the sources of organic aerosols. Environ. Sci. Technol. 41, 5763–5769.
Jolliff, J.K., Kindle, J.C., Shulman, I., Penta, B., Friedrichs, M.A.M., Helber, R., Arnone, R.A., 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. J. Mar. Syst. 76, 64–82.
Karagulian, F., Belis, C.A., 2012. Enhancing source apportionment with receptor models to foster the air quality directive implementation. Int. J. Environ. Pollut. 50, 190–199.
Larsen, B.R., Junninen, H., Monster, J., Viana, M., Tsakovski, P., Duvall, R.M., Norris, G.A., Querol, X., 2008. The Krakow Receptor Modelling Intercomparison Exercise *Rep.* JRC Scientific and Technical Reports, EUR 23621 EN 2008, Ispra.
Lee, J.H., Hopke, P.K., Turner, J.R., 2006. Source identification of airborne PM2.5 at the St. Louis-Midwest Supersite. J. Geophys. Res. D Atmos. 111 (D10S10), 1–12.
Paatero, P., 1997. Least squares formulation of robust non-negative factor analysis. Chemom. Intell. Lab. Syst. 37, 23–35.
Paatero, P., Hopke, P.K., 2003. Discarding or downweighting high-noise variables in factor analytic models. Anal. Chim. Acta 490, 277–289.
Paatero, P., Eberly, S., Brown, S.G., Norris, G.A., 2013. Methods for estimating uncertainty in factor analytic solutions. Atmos. Meas. Tech. Discuss. 6, 7593–7631.
Pandolfi, M., Viana, M., Minguillón, M.C., Querol, X., Alastuey, A., Amato, F., Celades, I., Escrig, A., Monfort, E., 2008. Receptor models application to multi-year ambient PM_{10} measurements in an industrialized ceramic area: comparison of source apportionment results. Atmos. Environ. 42, 9007–9017.
Schauer, J.J., Rogge, W.F., Hildemann, L.M., Mazurek, M.A., Cass, G.R., Simoneit, B.R.T., 1996. Source apportionment of airborne particulate matter using organic compounds as tracers. Atmos. Environ. 30, 3837–3855.
Thomson, M., Ellison, S.L.R., Wood, R., 2006. The international harmonized protocol for the proficiency testing of analytical chemistry laboratories. Pure Appl. Chem. 78, 145–196.
Thurston, G.D., Spengler, J.D., 1985. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. Atmos. Environ. Part A General Top. 19, 9–25.
Viana, M., Kuhlbusch, T.A.J., Querol, X., Alastuey, A., Harrison, R.M., Hopke, P.K., Winiwarter, W., Vallius, M., Szidat, S., Prévôt, A.S.H., Hueglin, C., Bloemen, H., Wåhlin, P., Vecchi, R., Miranda, A.I., Kasper-Giebl, A., Maenhaut, W., Hitzenberger, R., 2008a. Source apportionment of particulate matter in Europe: a review of methods and results. J. Aerosol Sci. 39, 827–849.
Viana, M., Pandolfi, M., Minguillón, M.C., Querol, X., Alastuey, A., Monfort, E., Celades, I., 2008b. Inter-comparison of receptor models for PM source apportionment: case study in an industrial area. Atmos. Environ. 42, 3820–3832.
Viana, M., Hammingh, P., Colette, A., Querol, X., Degraeuwe, B., Vlieger, I.D., Van Aardenne, J., 2014. Impact of maritime transport emissions on coastal air quality in Europe. Atmos. Environ. 90, 96–105.
Watson, J.G., Chen, L.W.A., Chow, J.C., Doraiswamy, P., Lowenthal, D.H., 2008. Source apportionment: findings from the U.S. supersites program. J. Air Waste Manag. Assoc. 58, 265–288.