Atmospheric Environment 180 (2018) 138–148

Contents lists available at ScienceDirect

Atmospheric Environment

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

DeltaSA tool for source apportionment benchmarking, description and sensitivity analysis

D. Pernigotti, C.A. Belis*

European Commission, Joint Research Centre, Ispra, Italy

ARTICLE INFO

Keywords:
Ensemble
Receptor models
Performance criteria
On-line

ABSTRACT

DeltaSA is an R-package and a Java on-line tool developed at the EC-Joint Research Centre to assist and benchmark source apportionment applications. Its key functionalities support two critical tasks in this kind of studies: the assignment of a factor to a source in factor analytical models (source identification) and the model performance evaluation. The source identification is based on the similarity between a given factor and source chemical profiles from public databases. The model performance evaluation is based on statistical indicators used to compare model output with reference values generated in intercomparison exercises. The references values are calculated as the ensemble average of the results reported by participants that have passed a set of testing criteria based on chemical profiles and time series similarity. In this study, a sensitivity analysis of the model performance criteria is accomplished using the results of a synthetic dataset where "a priori" references are available. The consensus modulated standard deviation *punc* gives the best choice for the model performance evaluation when a conservative approach is adopted.

1. Introduction

Despite the progress made in the latest decades, air pollution is still the primary environmental cause of premature death in Europe (Maas and Grennfelt, 2016). In order to design abatement measures, knowledge of the origin of pollutants affecting a given area is required (Directive 2008/50/EC, 2008). Source Apportionment (SA) aims to allocate shares of the measured pollutant mass to its emission sources, so called source contribution estimate (SCE). In the real-world, the actual SCEs are unknown. Due to such lack of references, a common problem in SA studies is to validate the model outputs. In the framework of the Forum for air quality modelling in Europe (FAIRMODE, 2007) the European Commission JRC launched, inter-comparison exercises for particulate matter SA among receptor models and more recently also for Chemical Transport Models. The experience gained analysing the data of such intercomparisons led to an European Guide for SA receptor models (Belis et al., 2014) and to a new methodology for evaluating SA performance (Belis et al., 2015a, 2015b: B2015 in the following).

In an intercomparison for SA (a glossary is provided in Appendix A) many practitioners run their models using the same input dataset, providing the following information for each source identified in the output (hereafter referred to as a candidate): the overall average SCE and the SCE time series (SCT) in absolute mass (e.g. $\mu g/m^3$) the source chemical profile (CP) and the contribution-to-species (the % of a given species attributed to that candidate source, C2S). In factor analytical methods the correspondence between factors with real-world sources or processes is accomplished in post processing (Hopke, 2009). Concerning source identification, particulate matter CPs measured at the source are the most reliable references. To support SA practitioners in this step, a repository for measured CPs was created. The above-mentioned SA evaluation methodology, embedded in the DeltaSA on-line tool, includes a total of 1 160 CPs from the SPECIEUROPE database developed at JRC (Pernigotti et al., 2016) and the SPECIATE database (Hsu et al., 2014). For pollutants deriving from secondary processes where measured profiles are not available (e.g. ammonium sulphate and nitrate), the stoichiometric profiles are considered.

The methodology for the intercomparison evaluation is described in B2015 and the results in Belis et al. (2015a and Belis et al. 2017. In the first steps (so called complementary and preliminary) for each source a set of screening criteria are made on the corresponding participants candidates. In real-world dataset the candidates successful to the previous test are averaged to build an ensemble reference, while this is not necessary if the reference is synthetic. In the final step each candidate is compared with the corresponding source reference to asses the participant performance. In particular, the second intercomparison is the only one performed with a synthetic dataset artificially created by the JRC where, unlike real-world data, pre-defined reference CP, SCE and

* Corresponding author.
E-mail addresses: deniperni@gmail.com (D. Pernigotti), claudio.belis@ec.europa.eu (C.A. Belis).

https://doi.org/10.1016/j.atmosenv.2018.02.046
Received 8 August 2017; Received in revised form 21 February 2018; Accepted 26 February 2018
Available online 01 March 2018

SCT are available. In all other intercomparisons no control on the reliability of the ensemble reference itself is possible.

In this paper, using the synthetic dataset, a criterion is proposed to improve the robustness of the methodology when using real-world datasets, taking into account the consensus among participants on the presence of a given source in the analyzed data. If there is a large consensus, than the model performance criteria (MPC) for that source will be more stringent. On the contrary if the consensus is low and just few participants agree on the presence of a contribution from that source, than the uncertainty will be larger and the MPC will be more stringent. The present study is divided in three sections. In the first, an updated version of the methodology for intercomparison evaluation (B2015) is summarised, as an introduction to the following sections. In the second, sensitivity tests using the abovementioned synthetic dataset to improve the model performance criteria (MPC) when using ensemble references are proposed. In the third section the Java web interface DeltaSA implementing functionalities for chemical profile similarity and model performance evaluation of the R-package with the same name is illustrated.

2. Developments in the methodology and re-evaluation of the synthetic dataset

The methodology described in B2015 has evolved as more experience with other intercomparisons was gained. In this section, the methodology for the SA model performance evaluation is summarised and the results of a sensitivity analysis using the synthetic dataset are presented. The methodology for the evaluation of the model performance comprises three steps: complementary tests, similarity tests and performance tests. In intercomparisons with real-world datasets, the objective of the first two steps is to select the candidates to be used for the ensemble reference.

2.1. The synthetic dataset

The synthetic dataset consists of artificially created PM2.5 daily average concentrations (total mass and chemical speciation for 38 species) for Milan in 2005 (Belis et al., 2015a). The 25 participants were using various receptor models (see Belis et al., 2015a for details on models): PMF (17 participants, mostly using PMF3), CMB (4), FA (2), ME2 (1), COPREM (1). They presented between 6 and 13 candidates each, with a total of 190 candidates, and 266 candidate-source couples (considering that some candidates were attributed to more than one source). There were up to four sources attributed to a single candidate while five candidates were attributed to sources that were excluded from the analysis. The most populated source was 1 (traffic), with 30 candidates.

The synthetic reference CP, SCE and SCT for each “a priori” source is shown in Fig. 1 (panels B,C and A respectively). The reference uncertainty was set to 20% for SCE and to 36% for SCT (the quadratic sum of 20%, and 30%, respectively the SCE and PM total mass uncertainties at each time step) while the CP uncertainty depends on the measurement technique.

2.2. Complementary tests

These tests provide information about the overall consistency of individual reported results. Additional checks are used to exclude participants and/or candidates whose results present macroscopic irregularities from the reference ensemble. In practice, those participants having the sum of time averaged SCT or sum of CP in absolute mass, differing by an order of magnitude from the sum of SCE or from the measured PM total mass, are excluded from the calculation of the reference. Only four candidates were excluded due to these criteria during the evaluation of the synthetic intercomparison (synthetic in the following).

In the updated methodology warnings are given for: a) participants with a difference of candidates with respect to the median for participants of more than three; b) candidates with the sum of the SCE of all the sources differing by more than 20% from the PM mass; c) candidates with SCT average total mass differing by more than 20% from SCE; d) candidates with the total reconstructed mass time series (sum of candidate SCTs) being out of the target plot (Thunis et al., 2012, in the following T2012, with the modification reported in Appendix B). Moreover, warnings are also given for candidates with less than four valid species in the CP, zero or missing SCE and/or CP.

2.3. Preliminary tests

Below we give a short summary of the tests, more extensive descriptions can be found in Appendix B, Pernigotti et al. (2016) and B2015. For each **candidate-source** couple (the couples are defined by each participant), the distances between the candidate and the source repository CPs corresponding to that source category are computed, together with the distances from all the other candidates attributed to the same source category (in the following denoted with the prefix ‘r_’ and ‘f_‘ respectively).

The distance indicators are: Pearson Distance (PD = 1-R, where R is the Pearson correlation coefficient) and the Standardised Identity Distance (SID). Only SIDcp and PDcp can be calculated against the repository CPs, while PDsct and PDc2s can be only calculated against the other candidates’ SCT and C2S. The suffix ‘-norm’ indicates that the SID has been normalised to account for the variability of the source categories (Appendix B). A source dependent coefficient $\boldsymbol{q}$ is set to the 95th percentile of the distances among repository CPs belonging to a given source category. This coefficient modulates the maximum allowed identity distance ID (MAD) for every source, so that the test is tolerant for sources with great variability in the measured chemical profiles and is stringent for those with a well-defined chemical fingerprint. In cases where $\boldsymbol{q}$ cannot be calculated (less than 3 CPs with at least 2 common species) a default value of 1 is taken. The value of $\boldsymbol{q}$ depends on the repository CPs, the considered source, as well as the intercomparison dataset, given that the calculation is only performed on the participants reported chemical species.

The values of $\boldsymbol{q}$ calculated for the synthetic dataset are: 1.15 for fuel oil, 1.12 for industrial, 1.08 for traffic, 1.08 for wood burning, 1.06 for biomass burning, 1.02 for exhaust, 1.01 for cement production, 0.94 for iron & steel production, 0.93 for road dust, 0.89 for de-icing salt, 0.88 for soil dust and 0.67 for marine aerosol. The default value of 1 is kept for the secondary sources.

The acceptability criteria for distances are $\mathbf{SIDcp\_norm} \leq \mathbf{1}$ and $\mathbf{PD} \leq \mathbf{0.4}$, where the first is given by the definition of SID_norm and the second corresponds to a Pearson coefficient above 0.6. In Fig. 2 SID distances between candidates and repository profiles are plotted for the synthetic arranged by category. The marine (12), deicing salt (66), secondary inorganic aerosol (60), ammonium nitrate (61) and ammonium sulphate (62) sources fall outside the acceptability area (green) when compared to repository profiles.

Only candidate-source couples fulfilling two out of the following three criteria pass the preliminary tests and are admitted to the ensemble reference calculation: 1) median of r-SIDcp-norm ≤ 1, 2) median of r-PDcp ≤ 0.4 (if the repository CPs are missing f distances are used) and 3) 25th percentile of f-PDsct ≤ 0.4. Criterion 3 aims at excluding candidates with an uncorrelated time trend. To avoid giving any single participant too much influence, where multiple candidates are from the same source only the candidate with the minimum r-SIDcp is kept.

In the synthetic after the application of the complementary tests all the (P = 25) participants are admitted to be ensemble members. In total 21 candidates and 50 candidate-source couples were excluded from the reference calculations.

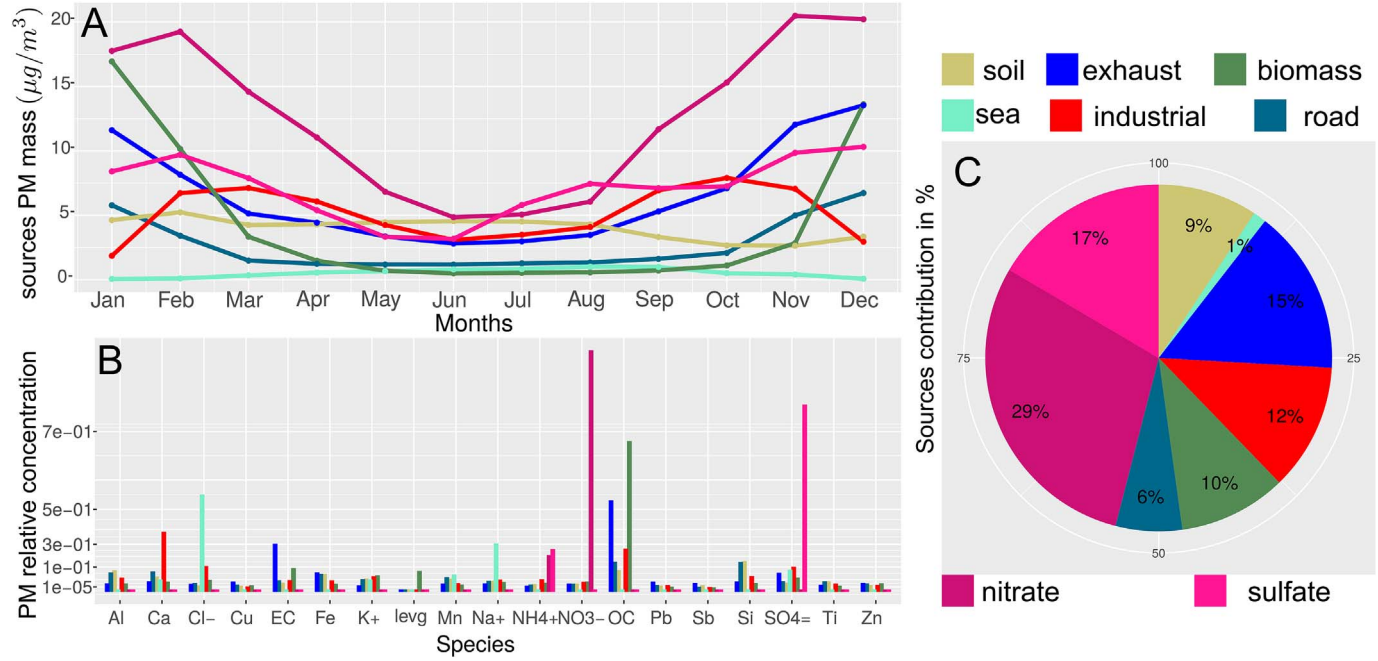

Fig. 1. Sources of the synthetic dataset used as reference: soil, marine, exhaust, industry, biomass burning, road dust, ammonium nitrate and ammonium sulphate. On panel B the chemical profiles in relative mass (CP), on panel A the seasonal trend of SCT in μg/m^3, on panel C the SCE as % of total PM mass.

2.4. Performance tests

Below a short summary of these tests is given, more extensive descriptions can be found in Appendix B, T2012 and B2015. The goal of this step is to measure the performance of each candidate using the z-score for SCE and the target approach for SCT. The uncertainty normalised root mean square error ($RMSE_u$) and associated target plot evaluates the bias, correlation and amplitude of the SCTs. The overall reference uncertainty used as a weighting factor in this model performance criterion is the source dependent combined uncertainty over the whole time series u_k, according to T2012 but without the use of the coverage factor 2 (see Appendix B):

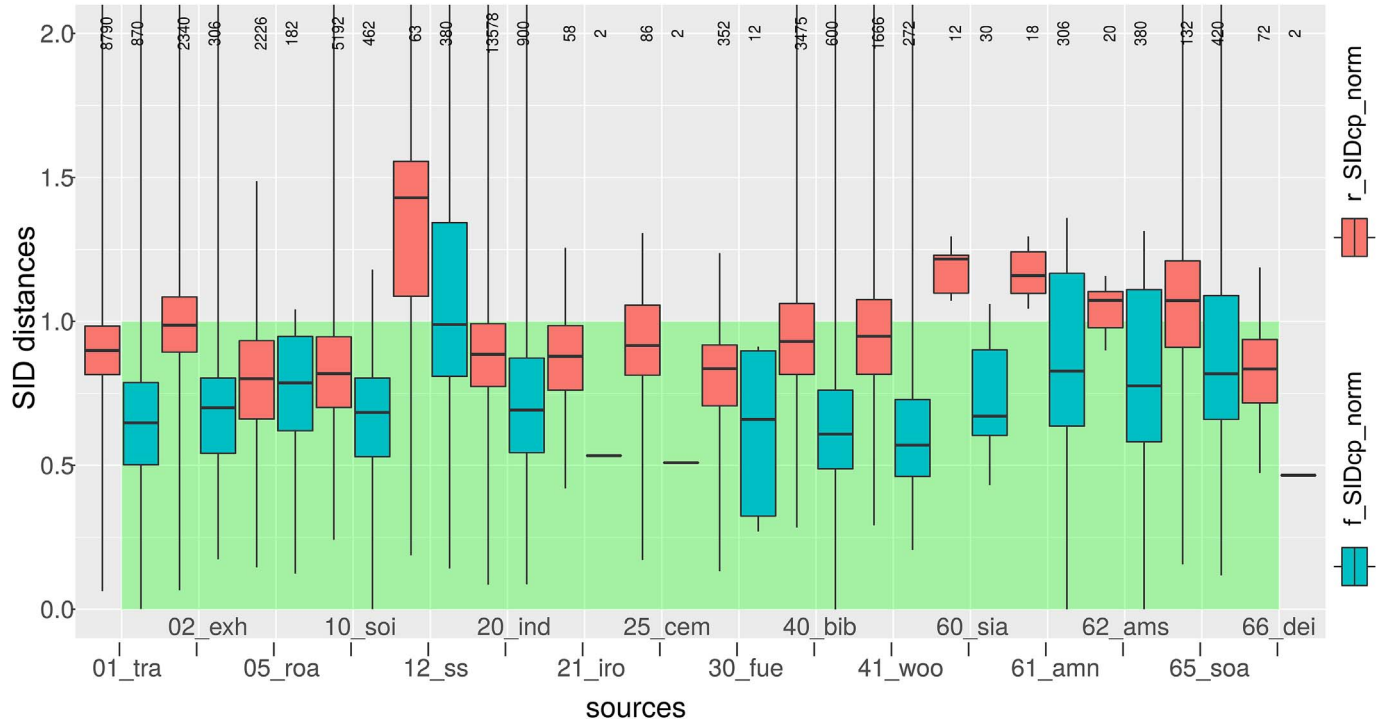

Fig. 2. SID_{norm} synthetic distance boxplots. Prefix 'r ' indicates the distances between a candidate and the repository CPs by source while 'f 'denotes the distances among candidates. On top are reported the number of computed distances by source. From left to right sources are 01:traffic, 02:exhaust, 05:road dust, 10:soil dust, 12:marine aerosol, 20:industrial, 21: iron and steel production, 25:cement production, 30:fuel oil, 40:biomass burning, 41:wood burning, 60:secondary inorganic aerosol (SIA), 61:ammonium nitrate, 61:ammonium sulphate, 66:deicing salt. The green background represents the area of acceptable distances. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

$$RMSEu_k = \frac{RMSE}{\beta u_k} \leq 1 \text{ , with } u_k = \sqrt{\frac{1}{N}\sum_{i=1}^{N} u_{ki}^2} \quad (1)$$

where k, i and βu_k stand for source, time step and model performance criterion weighting factor respectively.

3. Sensitivity tests for the ensemble reference uncertainty in the performance criteria

The synthetic reference (one for each source) uncertainty is given and can be used to judge the SCT performances of the candidates with the default β coefficient of 2 in eq. **(1)** as suggested in T2012. On the other hand, in real world applications a reference is not available and an ensemble value needs to be calculated together with an adequate model performance criterion based on its uncertainty. For this reason when using the ensemble reference a proper estimation of the equivalent to βu_k is crucial.

A sensitivity analysis of the effect of different estimations of the $RMSEu_k$ weighting factor for the performance criteria for the synthetic is proposed here. The synthetic is the only one available where the ensemble reference can be compared with an unbiased reference (synthetic reference in the following).

3.1. Ensemble bias

In Fig. 3 the z-scores (B2015 and Appendix B) of all candidates compared to the synthetic reference, aggregated by source are shown. Positive values indicate an overestimation of the SCE by the candidates while negative values indicate an underestimation. If the candidates of a given source were unbiased, z-scores would be more or less randomly distributed around zero as in sources 2, 10, 20, 40 and 62 (respectively exhaust, soil dust, industrial, biomass burning and ammonium sulphate). On the contrary, systematic underestimation is observed in source 61 (ammonium nitrate) and overestimation in sources 5 (road dust) and 12 (marine aerosol).

Fig. 4 shows the Talagrand rank histograms (Jolliffe and Stephenson, 2012; Talagrand et al., 1997) for SCT of marine (12, left) and ammonium nitrate (61, right). This kind of plot displays the distribution of the ordinal positions of the reference when ranking the members from smallest to biggest for each time step (days in this case). When the candidates are randomly distributed around the reference the bins are all equiprobable and the histogram is flat. If the histogram has a maximum in the lower extreme of the rank, it means that the members systematically overestimate the reference as in the case of marine (Fig. 4, left). Conversely if the histogram presents a maximum in the higher ranks, it means that the candidates systematically underestimate the reference as for ammonium nitrate (Fig. 4, right).

In intercomparisons using real-world datasets the reference is calculated from the participants' candidates. Calculating the ensemble reference as the average of candidates passing a series of preliminary tests is a common practice in modelling (e. g. Buizza, 1997), nevertheless, the indicators shown in Figs. 3 and 4 suggest that the ensemble references for road, marine and ammonium nitrate (5, 12 and 61) would be biased with respect to the synthetic reference. The positive bias of marine is likely to be due to the very small share of the total PM represented by this source (1%), which falls in the lower limit for the quantification of sources with factor analytical methods. Less prominent biases, falling within the acceptability thresholds, have been detected in road and ammonium nitrate (5 and 61 in Fig. 3). The positive bias of road dust is probably affected by soil dust (10), which has a similar chemical composition. A small underestimation common to all participants is observed in ammonium nitrate, which is the one with the highest SCEs in this dataset.

3.2. Ensemble reference model performance weighting factor

At each time step i (from 1 to N) and source k (from 1 to 8 for the synthetic dataset) the ensemble reference is calculated as the average $\overline{x_{ki}}$ among the candidates r (from 1 to R_k) passing the complementary and preliminary tests and having instantaneous values x_{ki}^r. The calculation of the instantaneous reference uncertainty is based on the source dependent instantaneous ensemble standard deviation defined as:

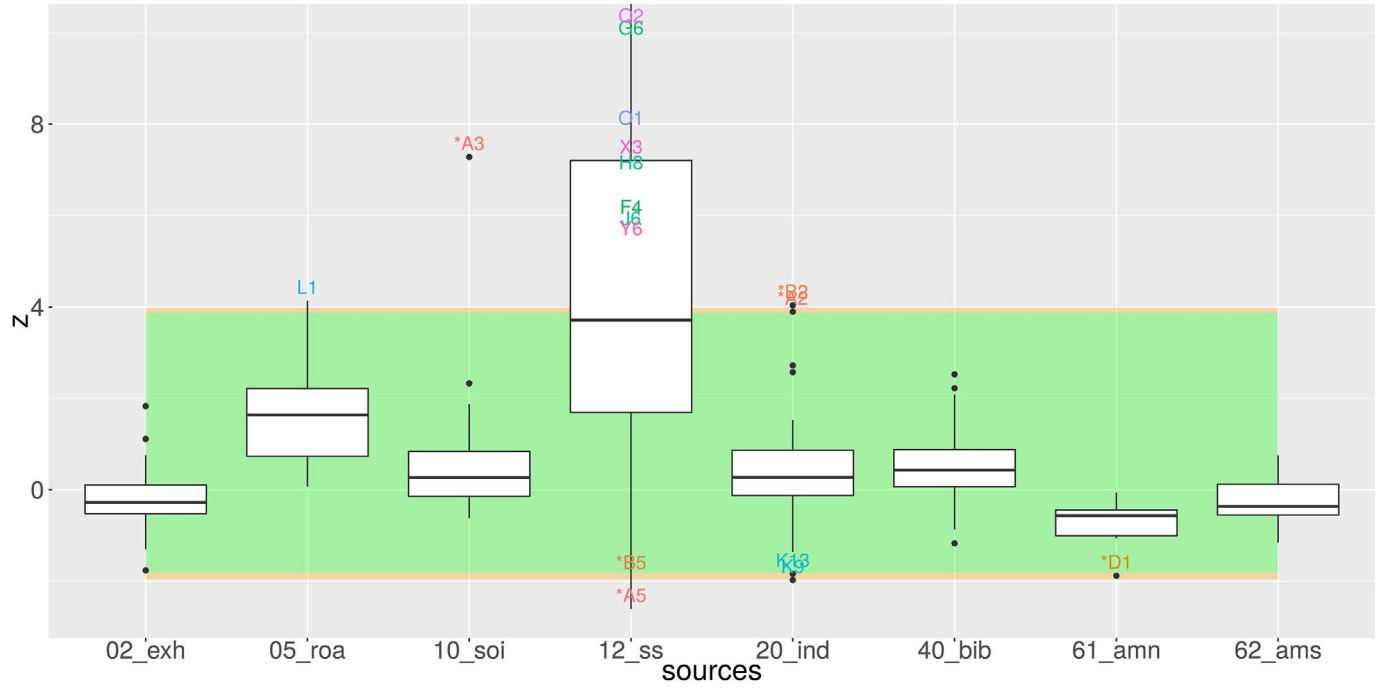

Fig. 3. z-scores for candidates' SCEs arranged per source category: exhaust (2), road dust (5), soil dust (10), marine (12), industrial (20), biomass burning (40), ammonium nitrate (61) and ammonium sulphate (62). Positive values indicate overestimation (positive bias) and negative values underestimation (negative bias) with respect to the synthetic reference. The green background represents the area of acceptability. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

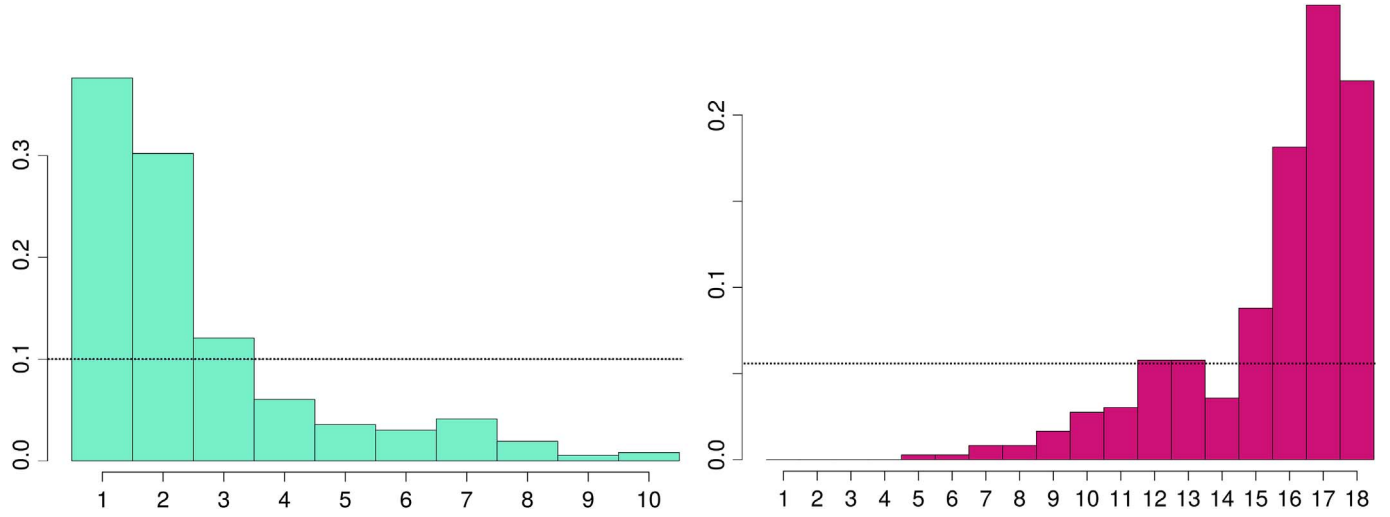

Fig. 4. Talagrand rank histograms for marine (12, on the left) and ammonium nitrate (61, on the right). Marine (left) includes 10 candidates and 11 bins while ammonium nitrate (right) includes 18 candidates and 19 bins. The dotted lines indicate the probability of the ideal flat distribution.

$$\sigma_{ki} = \sqrt{\frac{\sum_{r=1}^{R_k} (x_{ki}^r - \overline{x_{ki}})^2}{R_k}}, \quad (2)$$

In the model ensemble approach, the ensemble spread and in particular σ_{ki} is commonly adopted as a measure of the uncertainty u_{ki}. Substituting u_{ki} in equation (1) we obtain

$$u_k \simeq \sqrt{\frac{1}{N}\sum_{i=1}^{N} \sigma_{ki}^2} \quad (3)$$

As a first guess the value of $\beta = 1$ is adopted when estimating the weighting factor in Eq. **(1)**. In the target plot of Fig. 5 (see Appendix B and T2012 for details) the synthetic references are compared with the ensemble references with their uncertainties calculated according expression (3) and $\beta = 1$. The area where the criterion $RMSE/\beta u_k \leq 1$ is fulfilled has a green background. In this test, the synthetic reference of ammonium nitrate is not fulfilling the target criterion, indicating that the ensemble uncertainty estimated with expression (3) and $\beta = 1$ for the reference of this source is underestimated.

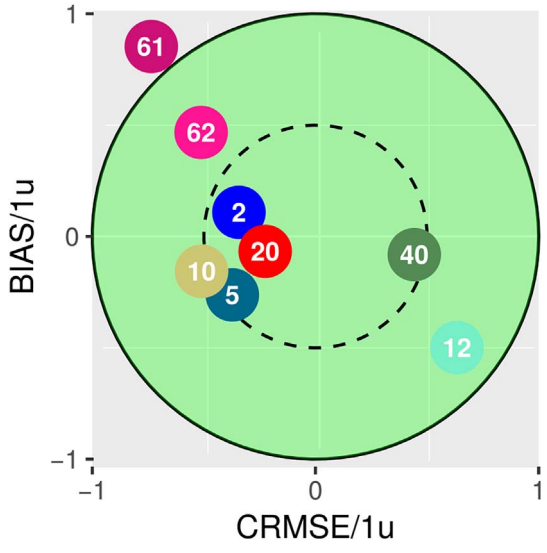

Fig. 5. Target plots of the synthetic references vs the ensemble references using its standard deviation and $\beta = 1$ as weighting factor for CRMSE and BIAS. The colour code (omitted for readability) and the white identification numbers refer to the sources and in particular source 61 is the ammonium nitrate as in Fig. 4. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

On one hand, the weighting factor for $RMSEu_k$ should be kept as small as possible in order to filter out results deviating too much from the reference, but on the other hand, the uncertainty should be large enough to avoid rejecting results which are consistent with the reference values. The sensitivity analysis on the model performance weighting factor described below aims to increase the βu_k to overcome the problem shown in Fig. 5, while keeping it as small as possible to maintain the selectivity of the method. In order to prevent penalising the candidates' sources under testing, in this work a conservative approach is adopted by giving the priority to the first criterion.

Four formulations for the model performance weighting factor equivalent to βu_k are here proposed:

1. ***sdunc:*** proportional to the u_k first guess as defined in equation (3)

$$sdunc_k = 1.23 u_k \quad (4)$$

The value $\beta = 1.23$ is the minimum enlargement of the green area in Fig. 5 for synthetic reference to be in the acceptability area for all sources when using the ensemble reference (not shown). In other terms, as it will be explained in the comment of Table 2, this value of β brings *sdunc* to 6.1 μg/m^3 for source 61 ammonium nitrate.

2. ***seunc:*** proportional to the standard error (standard deviation of the mean) $\frac{u_k}{\sqrt{R_k}}$

$$seunc_k = 5.07 \frac{u_k}{\sqrt{R_k}} \quad (5)$$

where the coefficient $\beta_k = \frac{5.07}{\sqrt{R_k}}$ is the minimum leading to all synthetic references be accepted in the target in Fig. 5, as explained for *sdunc*. The coefficient β_k depends on the number of candidates R_k used in the source ensemble calculation, being the smallest for the more populated and the biggest for less populated sources.

3. ***erms:*** the instantaneous root mean square error of the ensemble average (square root of equation (5) in van Loon et al., 2007)

$$erms_k = \sqrt{\left(1 + \frac{1}{R_k}\right) u_k^2 + b_k^2} \quad (6)$$

where the bias b_k was estimated from the comparison with the synthetic dataset. In line with the conservative approach adopted in this paper, a simple source independent upper-bound assessment of the absolute bias (for concentrations in μg/m^3) $b = 0.052x^2 + 0.55$, fitting these intercomparison data, was used for the *erms* in the sensitivity tests (not

Table 1
Modulation function for punc as a function of the proportion (here reported in % for convenience) of total participants used in the calculation of the ensemble reference.

P(%)	0%	5%	10%	15%	20%	25%	30%	35%	45%	50%	65%	85%	100%
punc	1.51	1.49	1.47	1.45	1.43	1.41	1.39	1.37	1.33	1.31	1.24	1.16	1.10

shown). This estimation of b_k is not available in the real-world datasets.

4. ***punc:*** is calculated starting from the standard deviation as follows,

$$punc_k = u_k(1.1 + 0.41(1 - P_k)) \quad (7)$$

where $P_k = \frac{R_k}{P}$ is the relative consensus among participants being R_k the participants reporting source k and P the total number of valid participants (25 for the synthetic). In this study P_k varies from 0.08 (ensemble calculated with just 8%, 2 participants out of 25) to 1 for the source populated by one candidate for each participant. In Table 1 the values of the linear function in equation (7) are summarised.

The modulation function adjusts the ensemble instantaneous standard deviation as a function of the consensus among participants on the presence of that given source in the dataset. The minimum value of *punc* corresponds to $\beta = 1.1$, when consensus is 100%. The *punc* modulating function is then increased with a power function, whose coefficient 0.41 has been found optimising (not shown) the agreement with the % of successful candidates obtained with the synthetic reference (last four columns in Table 2, described later). The *punc* inherits a similar but smoothed behaviour with respect to *seunc*, depending on P_k instead of R_k

In Table 2 the models' performances for the synthetic dataset using the synthetic reference and its uncertainty with $\beta = 2$ in Eq. (1) are compared with those using the ensemble reference and the four abovementioned methodologies. To assess the impact of the different approaches the performance corresponding to each methodology is also shown in terms of the percentage of successful candidates. The best weighting factor will be the one which are closer to the performance obtained using synthetic reference (green values), while approaching it by excess. For improve the results readability the best values for each source are also marked in green in the table, while in red are the values which are too low. The sources are ordered by decreasing P_k with the sources represented in the synthetic reference on the top.

The scaling coefficients of the ensemble weighting factors *seunc*, *sdunc* and *punc* (Table 2, columns 7 to 10) are set on the basis of the minimum required value (6.1 μg/m^3) that allows the most problematic source (61, ammonium nitrate) to pass the $RMSE_u$ test (Fig. 5). In the following, the performances obtained using the ensemble reference with and the four proposed methods (last four columns of Table 2) are compared with the performance obtained with the synthetic reference (column 5).

For the first six sources, for which the synthetic reference is also available, *seunc* and *erms* are not conservative for biomass burning with *seunc* being such also for ammonium sulphate. The opposite is true for *sdunc* and *punc* approaches. As a whole, the best results are those of *punc* as it gives the same % of successful candidates as the synthetic references, with the exception of industrial. Among the sources for which a reference is available, the last two sources (road and marine) show completely different performances when compared with the synthetic reference. These are the sources where a systematic overestimation of the synthetic reference was observed in the ensemble members (Fig. 3). This behaviour may be linked to the fact that in the synthetic dataset the uncertainty of the reference is proportional to its value. Thus, sources with very small contributions ($\leq$6%) tend to have unrealistically small uncertainties what makes the RMSEu test too stringent.

The last nine rows in Table 2 represent the sources that are only present in the ensemble reference. The presence of wood and de-icing salt is justified by their chemical affinity with biomass burning and marine, respectively. In other cases, sources are hierarchically related: traffic encompasses exhaust and road. Sources in the last six rows are those represented by less than 25% of the participants. This percentage could be considered as a threshold for flagging the scarce representativeness of sources, or even for their exclusion from the analysis when using and ensemble approach.

Table 2
Comparison of different MPC weighting factor for RMSE in Eq. (1). The total number of candidates assigned to each source by the participants is indicated by 'nc_all'. Synthetic: 'β *unc' is the synthetic MPC (μg/m^3) with $\beta = 2$ and '% successful' are the percentage of corresponding successful candidates relative to nc_all. Ensemble: 'P_k' is the relative number of participants accepted in the ensemble reference relative to nc_all; 'MPC weighting factors' (μg/m^3) calculated for the ensemble as described in the text; '% successful' percentage of corresponding successful candidates relative to nc_all. See text for green and red colours meaning.

Sources		nc all	Synthetic		Ensemble							
ID	Name		β*unc (μg/m^3)	% Successful	P_k	MPC weighting factor (μg/m^3)				% Successful			
						sdunc	seunc	erms	punc	sdunc	seunc	erms	punc
40	biom. burn.	25	5.4	76	96%	5.7	4.8	5.0	5.2	76	68	72	76
62	amm. sulph.	21	6	81	80%	4.1	3.8	4.1	4.0	81	77	81	81
10	soil	22	3.3	60	76%	3.2	3.0	3.1	3.1	64	60	60	60
20	industrial	31	5.4	42	76%	7.3	6.9	6.4	7.1	68	62	55	65
61	amm. nitr.	19	11.3	85	68%	6.1	6.1	6.7	6.1	85	85	85	85
2	exhaust	18	5.8	78	52%	6.3	7.3	5.9	6.7	78	89	78	78
5	road	14	2.5	22	52%	4.7	5.4	4.2	5.0	79	79	72	79
12	marine	21	0.7	24	40%	1.5	1.9	1.4	1.6	53	67	53	58
1	traffic	30			100%	6.1	5.1	5.5	5.5	87	60	67	67
41	wood burn.	17			64%	6.1	6.3	5.4	6.2	77	77	59	77
66	deicing salt	22			64%	2.3	2.4	2.0	2.3	73	73	69	73
60	SIA	6			24%	8.9	15.0	8.6	10.2	67	100	67	100
65	SOA	5			16%	4.1	8.4	4.3	4.8	60	100	60	80
30	fuel oil	4			12%	5.2	12.4	5.1	6.2	75	100	75	100
21	iron&steel	2			8%	2.1	6.1	2.5	2.5	100	100	100	100
25	cement	2			8%	2.4	7.1	3.1	2.9	100	100	100	100
74	combustion	2			8%	2.1	6.1	2.5	2.5	100	100	100	100

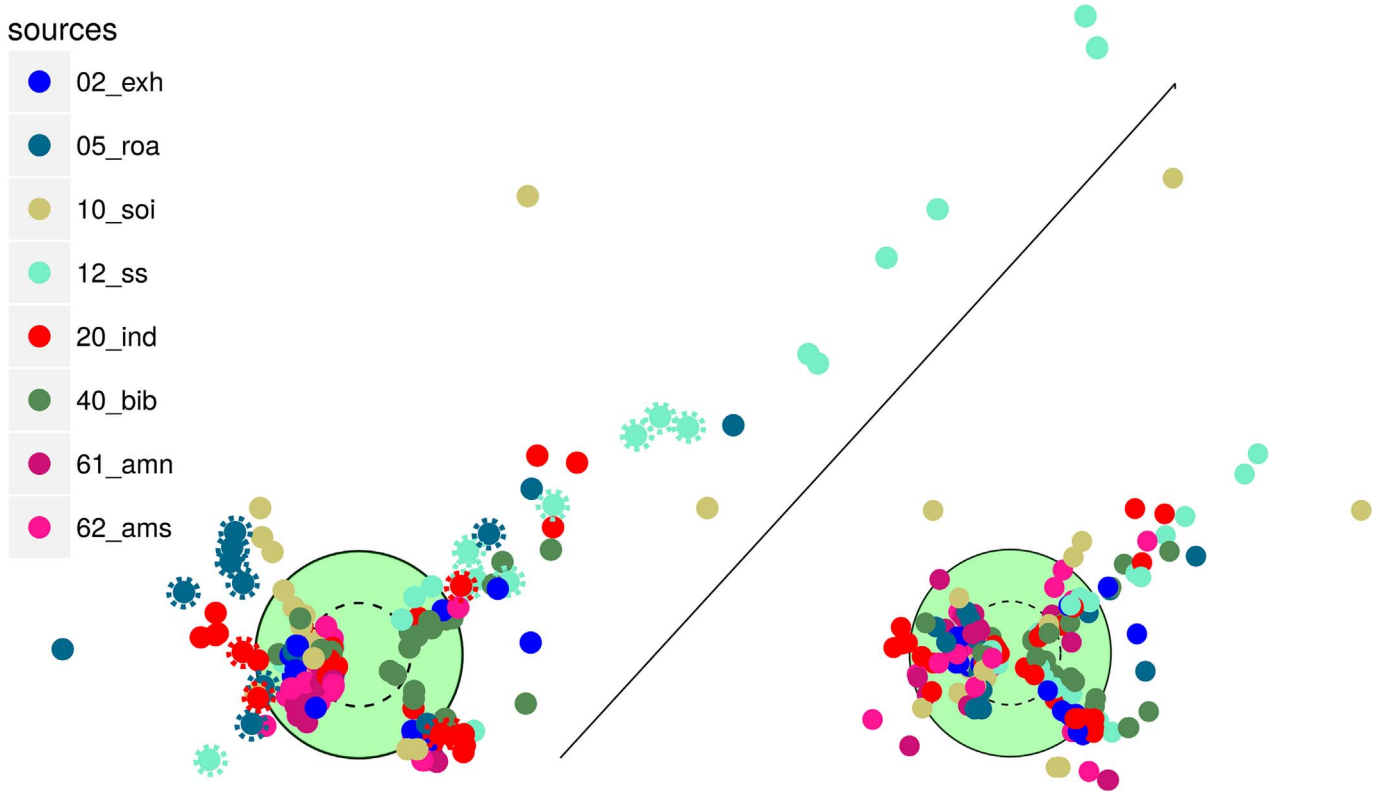

Fig. 6. Compared target plots with all candidates (dots) attributed to the various sources (colour code). On the left the reference is the synthetic one (with $\beta = 2$) and on the right candidates are assessed vs the ensemble reference with *punc* as RMSEu weighting factor. The enhanced candidates on the left are false positives: the would be out of the target while using the synthetic reference while they are in the target while using the ensemble reference. On the right there are not enhanced candidates because applying a conservative approach we avoid false negatives. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3.3. Performance criteria

In this section the focus is on the comparison between the performances of the candidates calculated with the synthetic reference and the ones calculated using the ensemble average and *punc* as RMSEu weighting factor, which proved to be the most conservative and most selective one.

In Fig. 6 a detailed picture of the results summarised in columns 5 and 14 of Table 2, respectively, is shown. A total of 171 candidates were attributed to sources for which a reference is available (first 8 rows in Table 2). Out of them 42% (71) do not pass the RMSEu test (Fig. 6 left) using the synthetic reference. The rejected candidates are reduced to 49 (29%) when using the ensemble reference with *punc* (Fig. 6 right). As reported also in Table 2, the 22 candidates not rejected using the ensemble reference that are rejected with the synthetic reference (false positives, where positive stands for inside the acceptability area) are all attributed to industrial (20_ind, 6 candidates), road dust (05_roa, 8 candidates) and marine (12_mar, 8 candidates). All the candidate-sources rejected using the ensemble reference are also rejected using the synthetic reference (no false negatives), as required by the conservative approach used. The absence of false negatives demonstrate the consistency between the *punc* approach and the synthetic reference and provides evidence in favour of its application in the evaluation of SA results when using real-world testing datasets.

4. Delta tool for source apportionment

The methodology for source apportionment intercomparison and performance measurement is available in an R-package and on-line through the Java on-line interface DeltaSA (DeltaSA, 2016). The on-line interface works in two ways: a) to compare user CPs with those in the repositories to support the identification of PM sources and b) to execute a complete test of user SA results using existing intercomparisons' datasets for which references SCE are available.

4.1. Chemical Profile Similarity (CPS)

One (or more) user CP, expressed as relative mass of the species to the total PM mass, is evaluated by comparison with 1 160 source chemical profiles taken from the combined SPECIEUROPE and SPECIATE public databases, hereafter referred to as the DeltaSA set. The distances PD and SID and their 95% confidence intervals are calculated for all the CPs in the DeltaSA set. The user is not requested to provide any a priori relationship between the tested CP and real source categories.

An example of CPS output plot is provided in Fig. 7 for the synthetic reference corresponding to road dust (source 5). The distances' average (centroid) from all the sources in the same source category is represented by a dot where the size is proportional to the source population and the error bars represent the 95% confidence interval. The sources are identified both with a colour code and with the corresponding identification number in the DeltaSA set (Table 2 first column for a subset of them). The output is also provided as a table where the sources are ordered by proximity taking into account the average distance plus its confidence interval. These data can help the user in the final identification of the source.

4.2. Source apportionment Model Performance (MP)

In this section, an assessment of the performance of the user SA application using as input existing intercomparison datasets is accomplished. All the complementary and preliminary tests described in section 2 are executed in this mode. To that end, the user is requested to link the candidates to one or more predefined source categories using a selection window. The first MP output is a summary of the uploaded

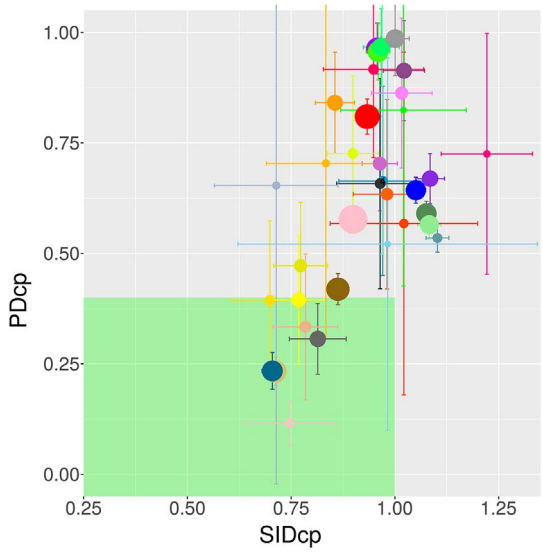

Fig. 7. Example of CP similarity plots for the synthetic reference for road dust (5). The dots represent the SID versus PD distances of the candidate source to sources with more than 3 reference profiles in SPECIEUROPE/SPECIATE grouped by source category (average and 95% confidence interval). The legend with all sources colour codes is omitted here for readability. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

data for user checking purposes similar to the one in Fig. 1. The second output is a composite graph including boxplots and a scatter plots summarising the distances of the user candidates to the pre-loaded intercomparison' candidates and to the DeltaSA set (Fig. 8). The third and main MP output synthetises the performance of the user SA application evaluated on the basis of the intercomparison reference (Fig. 9).

In Fig. 8 an example of boxplots and scatter plot for one participant to the second intercomparison is shown on the left: for each candidate (abscissa) the distribution of its distances from the candidates attributed to the same source which have passed the preliminary and complementary tests is shown. The colour code is used to identify the sources. The user has the chance to test one single candidate in more than one source. For instance, candidate 9 is tested for source traffic (grey) and exhaust (blue). These boxplots provide an overall view of the similarities between the user candidates and those of the intercomparison. For instance, it is evident that candidate 1 is not similar to other candidates in source road dust (5, dark green), while the opposite is true for candidate 3 attributed to soil dust (10, khaki). The right panel of Fig. 8 is a scatter plot of PD and SID_{norm} (see B2015 and Appendix B) distances between the user candidates and the CP in the DeltaSA set. The white numbers on the dots denote the candidate numbers. In this example, the unlikely attribution of candidate 1 to road dust (5, dark green) observed in the left panel is confirmed.

In Fig. 9 an example of performance plots: the target plot (left panel) and z-score bar plot (right panel) are presented. The candidate-sources excluded from the analysis (not classified) because the candidate has been attributed to a source for which the reference is missing are indicated in the bottom-left corner. In this case, the synthetic reference for wood burning (41), fuel oil (30), de-icing salt (66) and traffic (1) are not defined, therefore, measuring the performances of candidates 2, 5, 7 and 9 assigned to them is not possible. In the example of Fig. 9, the candidate 1 attributed to road dust is not in the acceptability range for any of the criteria, while candidate 3 attributed to soil dust is performing well.

5. Conclusions

In this study, refinements to the methodology for evaluating source apportionment applications performance are proposed. The new features encompass new checks and warnings in the preliminary tests to identify outliers and a better definition of the criteria to accept candidates for the calculation of the source specific ensemble references.

The features described in this study are implemented in the recently released DeltaSA on-line and R-package versions. The tool provides key functionalities in support of the identification of the PM sources taking advantage of public repositories. In addition, the performance testing module gives practitioners the chance to check their skills and to test the impact of different model set ups on the performance. The

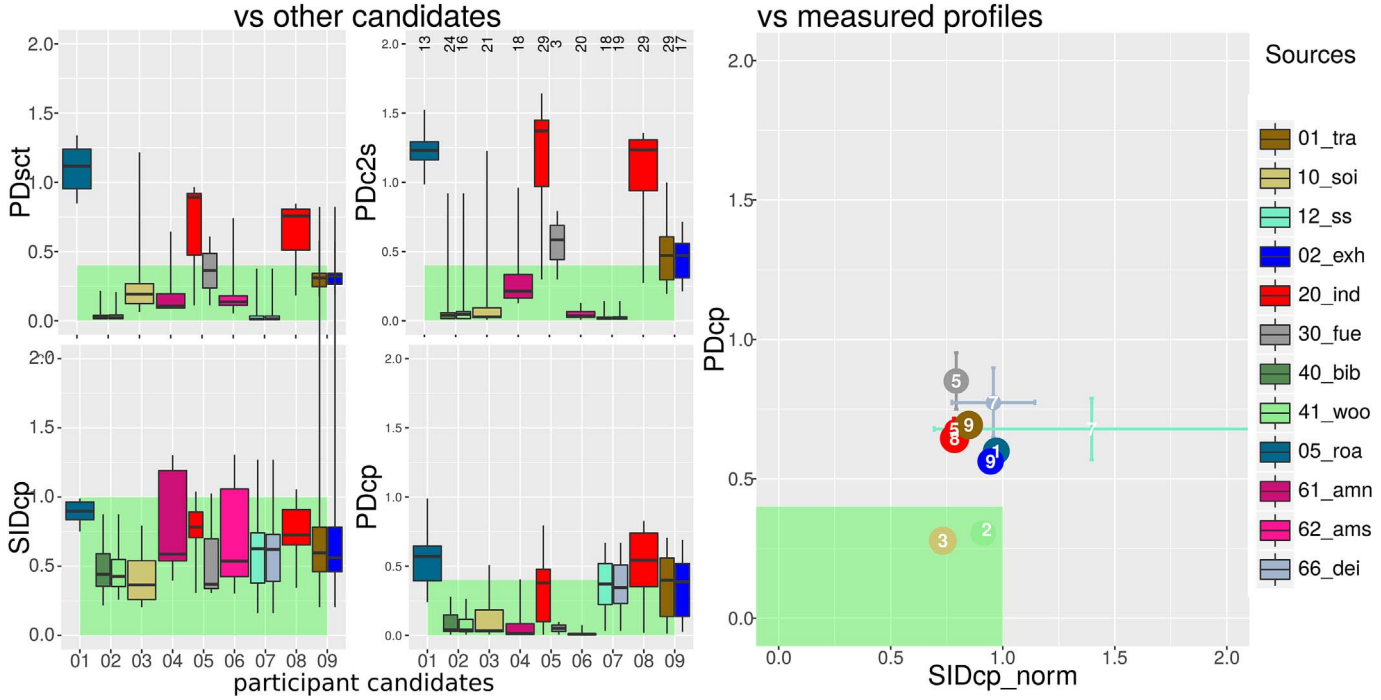

Fig. 8. Example of boxplots for one synthetic participant displaying distances of the user candidates to the other candidates (four boxplots on the left); scatter plot of SID_norm and PD distances between the user candidates and DeltaSA set of measured profiles (right). More details in the text.

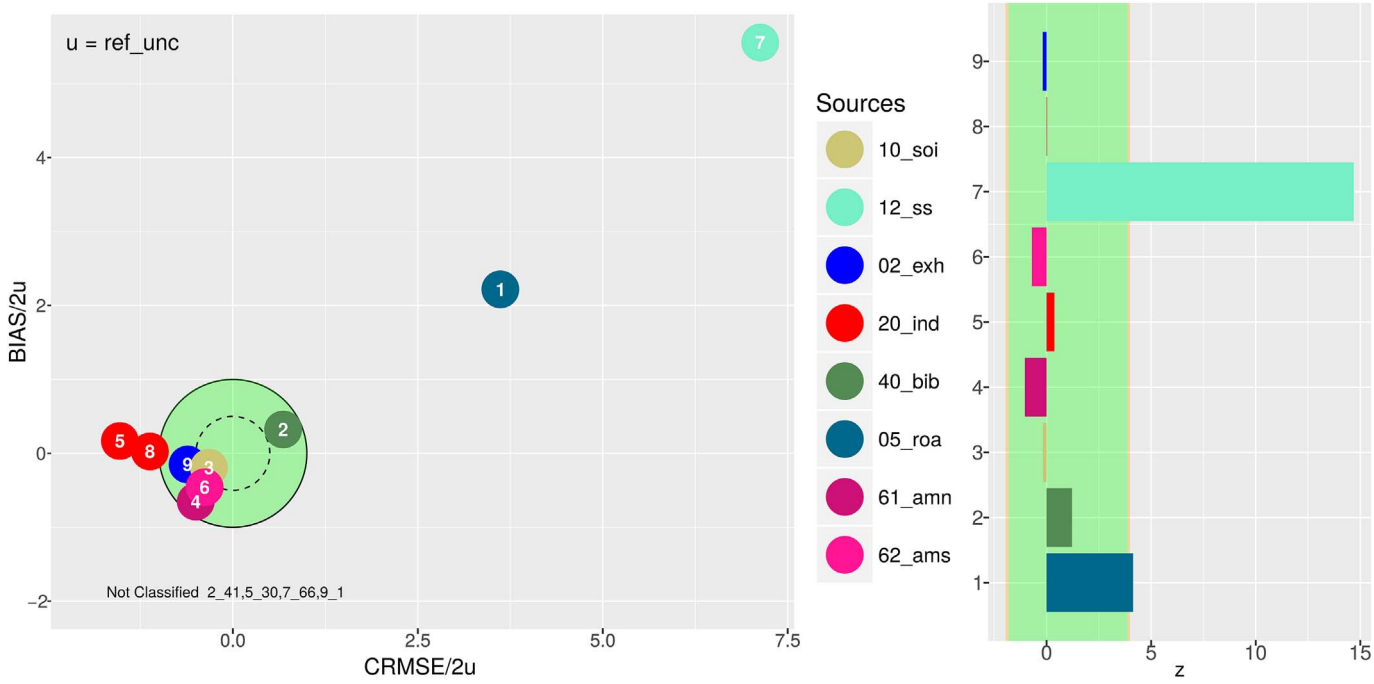

Fig. 9. Example of performance plot for one synthetic participant against the synthetic dataset (β = 2). On the left is shown the target plot and the z-score is displayed in the right panel. More details in the text.

advantages of the on-line option are that the user does not need to install any application and that updates are readily available. In addition to supporting SA applications and testing intercomparisons, DeltaSA can also be regarded as a valuable training resource for practitioners with little experience in this field. The availability of DeltaSA is expected to contribute to improving the harmonisation and traceability of the PM sources reported in future SA studies.

Considerable bias may occur for sources with very small or very high contributions (e.g. marine aerosol, ammonium nitrate in the synthetic dataset) and to sources which are very similar to others (e.g. road and dust in the synthetic dataset). This is a limit to the use of the ensemble approach in testing the participants' performances under certain conditions. A sensitivity analysis on different methods to estimate the weighting factor for the model performance criterion RMSEu is presented. The consensus modulated standard deviation *punc* results the best choice as it closely reproduces the performances obtained using the synthetic references. This method adopts a conservative approach to avoid false negatives (wrongly excluded from the green circle of the target plot) and shows a slightly lower selectivity than the synthetic reference in the RMSEu test. This study highlights the degree of consensus among participants on the effective presence of a particular source is key information in the evaluation of SA intercomparison. It may be also used to flag weak sources by setting a threshold to the share of results in which the sources is reported (e.g. 25%). Testing the proposed methodology and parameterisation on other synthetic datasets is needed to provide the basis for a more general application in SA model performance evaluation.

Appendix A. Glossary

SA: source apportionment.
participant: a practitioner participating in a SA intercomparison.
candidates: sources in a SA result reported by a participant.
source: short for source category, which indicated a group of air pollution sources of the same type that emit pollutants with similar chemical composition.
candidate-source: a candidate attributed to a source by a participant using factor analytic methods; multiple candidate-sources couples are possible for a single candidate.
verified candidate-source: candidate passing the complementary and preliminary test when attributed to a given source.
SCE: source contribution estimate time averaged (absolute mass of particulate matter attributed to the candidate).
SCT: source contribution estimate time series (SCE time series).
CP: chemical profile (each species mass relative to the candidate SCE).
C2S: contribution to species (each species % mass attributed to the candidate).
PD: Pearson Distance, 1-Pearson product-moment correlation coefficient.
SID: standardised identity distance.
r_: prefix to indicate distances from repository CP of a given source.
f_: prefix to indicate distances from verified candidates-source CP, SCT or C2S of given source.

Appendix B. Statistical indicators

Over bars indicate time averaged variables, M stands for model data and O for observation (reference data), for time i going from 1 to N:

$$BIAS = \overline{M} - \overline{O};$$

$$CRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((O_i - \overline{O}) - (M_i - \overline{M}))^2};$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - M_i)^2} = \sqrt{BIAS^2 + CRMSE^2};$$

Model Performance Criteria (T2012):

$$RMSE_U = \frac{RMSE}{\beta U} \leq 1 \text{ with } U = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(ku_i)^2}$$

where u_i is the uncertainty of the reference (the observations in this case) at each time step and $U = ku$, is the expanded uncertainty (coverage factor of $k = 2$ for a confidence level of approximately 95%). The T2012 target is a modified version of Jolliff et al. (2009) with CRMSE/βU as abscissa and BIAS/βU as ordinate, so that the green circle represents the area where the Model Performance Criterion (MPC) RMSEU is fulfilled.

In this study we use the modified version of the RMSE (B2015):

$$RMSE_U = \frac{RMSE}{\beta U} \leq 1 \text{ , with } u = \sqrt{\frac{1}{N}\sum_{i=1}^{N}u_i^2}$$

without the use of the coverage factor. This approach is in line with the current praxis for uncertainty manipulation, where the coverage factor is only added at the end of the calculations in order to transform an uncertainty in a confidence interval (JCGM, 2008).

A value of the $\beta = 2$ is adopted when reference data were available, as in the synthetic dataset, while the sensitivity test presented in the paper aims to estimate a value of βu which mimic the model performance when using an ensemble reference.

As the CRMSE is positive by definition (so that only the right part of the diagram would be populated) so that the left part of the Target plot is also used the model performance problems are more related to standard deviation problem (right part of the plot) or to Pearson correlation problem (left part of the plot). We leave the reader to T2012 and to the recent manual of the Delta Version 5.5 (Thunis and Cuvelier, 2016) for further details.

Z-score

In ISO 13528 (2005) the z-score is defined as:

$$z = \frac{x - X}{\sigma_p},$$

where X is the reference value and σ_p is the standard deviation for proficiency assessment.

In the methodology (B2015) x is the SCE of a candidate-source, X is the reference SCE for that source and $\sigma_p = 0.5X$ as the PM10 model quality objective in Directive 2008/50/EC (2008). The z-scores critical values were determined according to ISO 13528:2015 on proficiency testing. To that end, the statistical distribution of the z-scores was estimated fitting a kernel distribution (R package ks v. 1.9.2) to a dataset with more than 200 unbiased z-scores after removing outliers. Extracting the 0.005, 0.025, 0.975 and 0.995 percentiles from the obtained distribution lead to the definition of areas with the same probability density as those specified in the abovementioned standard for normal distributions. In this study, the following critical values were used: −1.96, −1.81, 3.87 (B2015).

The Standardised Identity Distance (SID)

This distance was introduced in the methodology (B2015, Pernigotti et al., 2016) to measure the proximity between two CPs (in relative mass) x and y, with j = 1 to m common species.

The identity distance ID (B2015) measures the average distance from the identity

$$ID = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{\sqrt{2}}|x_j - y_j|$$

Geometrically for each species j, $\frac{1}{\sqrt{2}}|x_j - y_j|$ corresponds to the distance of x_j and y_j form the identity line.

The MAD is a threshold for the maximum accepted distance of the two species, defined by the user as a function of the identity (the average of the two species relative mass)

$$MAD_j = k\frac{1}{2}(x_j + y_j)$$

The standardised distance is defined as the fraction of ID to MAD:

$$SID = \frac{1}{m}\sum_j \frac{ID_j}{MAD_j} = \frac{1}{m}\sum_j \frac{\frac{1}{\sqrt{2}}|x_j - y_j|}{k\frac{1}{2}(x_j + y_j)} = \frac{\sqrt{2}}{qm}\sum_{j=1}\frac{|x_j - y_j|}{(x_j + y_j)}$$

The q coefficient for a given source category is calculated as the 95th percentiles of the SID with q = 1 among repository CP and they range from 0.67 (marine) to 1.15 (fuel oil). In the text SID is used when q = 1 while SID_norm is used when q is calculated.

References

Belis, C.A., Karagulian, F., Amato, F., Almeida, M., Artaxo, P., Beddows, D.C.S., Bernardoni, V., Bove, M.C., Carbone, S., Cesari, D., Contini, D., Cuccia, E., Diapouli, E., Eleftheriadis, K., Favez, O., El Haddad, I., Harrison, R.M., Hellebust, S., Hovorka, J., Jang, E., Jorquera, H., Kammermeier, T., Karl, M., Lucarelli, F., Mooibroek, D., Nava, S., Nøjgaard, J.K., Paatero, P., Pandolfi, M., Perrone, M.G., Petit, J.E., Pietrodangelo, A., Pokorná, P., Prati, P., Prevot, A.S.H., Quass, U., Querol, X., Saraga, D., Sciare, J., Sfetsos, A., Valli, G., Vecchi, R., Vestenius, M., Yubero, E., Hopke, P.K., 2015a. A new methodology to assess the performance and uncertainty of source apportionment models II: the results of two European intercomparison exercises. Atmos. Environ. 123, 240–250. http://dx.doi.org/10.1016/j.atmosenv.2015.10.068.

Belis, C.A., Larsen, B.R., Amato, F., Haddad, I.E., Favez, O., Harrison, R.M., Hopke, P.K., Nava, S., Paatero, P., Prévôt, A., Quass, U., Vecchi, R., Viana, M., 2014. European Guide on Air Pollution Source Apportionment with Receptor Models. (JRC reference reports No. Report EUR 26080 EN). JRC.

Belis, C.A., Pernigotti, D., Karagulian, F., Pirovano, G., Larsen, B.R., Gerboles, M., Hopke, P.K., 2015b. A new methodology to assess the performance and uncertainty of source apportionment models in intercomparison exercises. Atmos. Environ. 119, 35–44. http://dx.doi.org/10.1016/j.atmosenv.2015.08.002.

Belis, C.A., Pirovano, G., Pernigotti, D., 2017. Preliminary Results of the first European source apportionment intercomparison for receptor and chemical transport models. Geophys. Res. Abs. 19 EGU2017-16905.

Buizza, R., 1997. Potential forecast skill of ensemble prediction and spread and skill distributions of the ECMWF ensemble prediction system. Mon. Weather Rev. 125, 99–119. http://dx.doi.org/10.1175/1520-0493(1997)125<0099:PFSOEP>2.0.CO;2.

DeltaSA [WWW Document], 2016. URL http://source-apportionment.jrc.ec.europa.eu/DeltaSA/index.html (accessed 11.22.16).

Directive 2008/50/EC, 2008. Directive 2008/50/EC of the european parliament and of the council of 21 may 2008 on ambient air quality and cleaner air for Europe. [WWW Document]. Ambient Air Qual. Clean. Air Eur. 4. paragraph 10. http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32008L0050 (accessed.6 24.15).

FAIRMODE, 2007. Forum for Air Quality Modelling in Europe.

Hopke, P.K., 2009. Chapter 1 Theory and Application of Atmospheric Source Apportionment.

Hsu, Y., Divita, F., Jonathan, Dorn, 2014. SPECIATE Version 4.4 Database Development Documentation (No. EPA/600/R-13/307). Abt Associates Inc.

ISO 13528, 2005. Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparisons. International Standard.

JCGM 100, 2008. Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement. Bureau International des Poids et Mesures, Paris, pp. 134.

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. http://dx.doi.org/10.1016/j.jmarsys.2008.05.014.

Jolliffe, I.T., Stephenson, D.B. (Eds.), 2012. Forecast Verification: a Practitioner's Guide in Atmospheric Science, 2. ed. Wiley-Blackwell, Chichester.

Maas, R., Grennfelt, P. (Eds.), 2016. Towards Cleaner Air. Scientific Assessment Report 2016. EMEP Steering Body and Working Group on Effects of the Convention on Long-Range Transboundary Air Pollution, Oslo.

Pernigotti, D., Belis, C.A., Spanò, L., 2016. SPECIEUROPE: the European data base for PM source profiles. Atmospheric Pollut. Res. 7, 307–314. http://dx.doi.org/10.1016/j.apr.2015.10.007.

Talagrand, O., Vautard, R., Strauss, B., 0ctober 1997. Evaluation of probabilistic prediction systems. In: Proceeding of workshop on predictability. European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, Berkshire RG2 9AX, UK, pp. 1–25.

Thunis, P., Cuvelier, C., 2016. DELTA Version 5.5.

Thunis, P., Pederzoli, A., Pernigotti, D., 2012. Performance criteria to evaluate air quality modeling applications. Atmos. Environ. 59, 476–482. http://dx.doi.org/10.1016/j.atmosenv.2012.05.043.

van Loon, M., Vautard, R., Schaap, M., Bergström, R., Bessagnet, B., Brandt, J., Builtjes, P.J.H., Christensen, J.H., Cuvelier, C., Graff, A., Jonson, J.E., Krol, M., Langner, J., Roberts, P., Rouil, L., Stern, R., Tarrasón, L., Thunis, P., Vignati, E., White, L., Wind, P., 2007. Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble. Atmos. Environ. 41, 2083–2097. http://dx.doi.org/10.1016/j.atmosenv.2006.10.073.