European
Commission

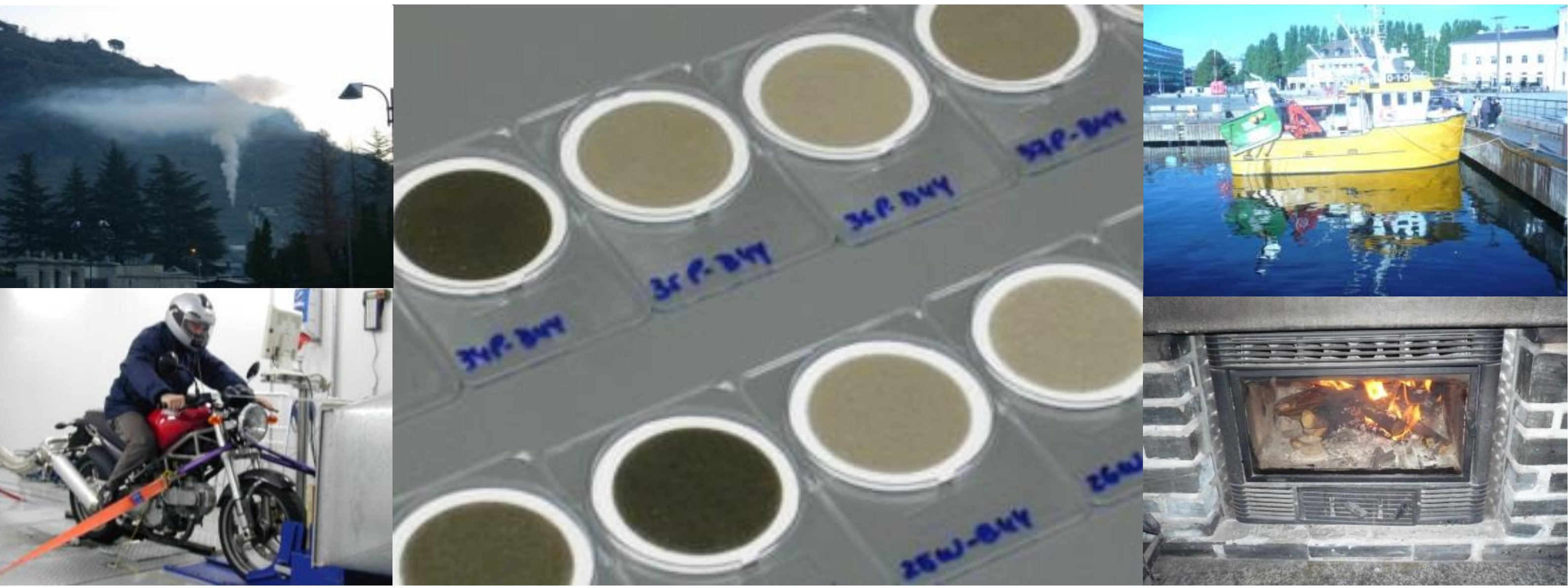
JRC TECHNICAL REPORTS

European guide on air pollution source apportionment with receptor models

Revised version 2019

Belis C.A., Favez O., Mircea M.,
Diapouli E., Manousakas M-I., Vratolis S.,
Gilardoni S., Paglione M., Decesari S.,
Mocnik G., Mooibroek D., Salvador P.,
Takahama S., Vecchi R., Paatero P.

2019



EUR 29816 EN

Joint
Research
Centre

Contact information
Name: Claudio A. Belis
Address: European Commission, Joint Research Centre, via Fermi 2749, 21027 Ispra, Italy
Email: claudio.belis@ec.europa.eu
Tel.: +39 0332 786644

EU Science Hub
https://ec.europa.eu/jrc

JRC117306

EUR 29816 EN

PDF	ISBN 978-92-76-09001-4	ISSN 1831-9424	doi:10.2760/439106

Luxembourg: Publications Office of the European Union, 2019

How to cite this report:
Belis C.A., Favez O., Mircea M., Diapouli E., Manousakas M-I., Vratolis S., Gilardoni S., Paglione M., Decesari S., Mocnik G., Mooibroek D., Salvador P., Takahama S., Vecchi R., Paatero P., *European guide on air pollution source apportionment with receptor models - Revised version 2019*, EUR 29816 EN, Publications Office of the European Union, Luxembourg, 2019, ISBN 978-92-76-09001-4, doi:10.2760/439106, JRC117306.

European guide on air pollution source apportionment with receptor models

Revised version 2019

Keywords: pollution sources, harmonisation, receptor models, air quality directives, pollution abatement measures, thematic strategy on air pollution

TABLE OF CONTENTS

ACKNOWLEDGMENTS 1

SUMMARY 6

INTRODUCTION TO SOURCE APPORTIONMENT WITH RECEPTOR MODELS 11

1. PRELIMINARY EVALUATION OF THE STUDY AREA 22

2. DEFINING A METHODOLOGICAL FRAMEWORK 26

3. EXPERIMENT DESIGN - CRITERIA FOR SITE AND SPECIES SELECTION AND ESTIMATION OF MINIMUM NUMBER OF SAMPLES 28

4. DATA COLLECTION / FIELD WORK / CHEMICAL ANALYSIS 36

5. KNOWING YOUR DATASET: BASIC STATISTICS 42

6. PRELIMINARY DATA QUALITY CHECKS 46

7. INPUT DATA UNCERTAINTY CALCULATION 52

8. CHEMICAL MASS BALANCE MODELS 58

9. FACTOR ANALYSIS I: SELECTION OF THE NUMBER OF FACTORS AND DEALING WITH ROTATIONAL AMBIGUITY (PMF) 63

10. FACTOR ANALYSIS II: EVALUATION OF SOURCE CONTRIBUTION ESTIMATION AND MODEL PERFORMANCE INDICATORS 68

11. FACTOR ANALYSIS III: CRITERIA FOR FACTOR ASSIGNMENT 74

12. FACTOR ANALYSIS IV: CONSTRAINED AND EXPANDED MODELS 76

13. FACTOR ANALYSIS V: DATA PROCESSING OF IN SITU THERMO-DESORPTION ELECTRON IMPACT AEROSOL MASS SPECTROMETRY 82

14. AETHALOMETER MODELS 94

15. RADIOCARBON AND TRACER ANALYSIS FOR THE APPORTIONMENT OF THE PM CARBONACEOUS FRACTION 102

16. PROTON-NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY FOR THE SOURCE APPORTIONMENT OF WATER SOLUBLE ORGANIC CARBON 109

17. SOURCE APPORTIONMENT BY FOURIER-TRANSFORM INFRARED (FTIR) ANALYSIS 114

18. WIND AND TRAJECTORY ANALYSIS IN SOURCE APPORTIONMENT 120

19. TESTS FOR MODEL PERFORMANCE VALIDATION 147

20. REPORTING RESULTS AND METHODOLOGY 159

ACKNOWLEDGMENTS

The authors are grateful to F. Karagulian and G. Calzolai for their contributions to sections 7 and 12, respectively. Many thanks also to G. Mulhern for proofreading the drafts, and to E. Vignati, M. Bressi, L. Bontoux, and D. Buzica for reviewing and commenting on the manuscript.

Authors (original and revised versions)

Claudio A. Belis (chair and editor), Favez Olivier (editor), Fulvio Amato, Stefano Decesari, Evangelia Diapouli, Imad El Haddad, Stefania Gilardoni, Roy Harrison, Philip K. Hopke, Bo R. Larsen, Manousos-Ioannis Manousakas, Grisa Mocnik, Dennis Mooibroek, Silvia Nava, Pentti Paatero, Marco Paglione, André Prévôt, Ulrich Quass, Pedro Salvador, Satoshi Takahama, Roberta Vecchi, Mar Viana, Stergios Vratolis.

Section contributors

Section 1. P. K. Hopke, C. A. Belis
Section 2. C. A. Belis
Section 3. C. A. Belis, U. Quass
Section 4. C. A. Belis, O. Favez
Section 5. C. A. Belis, B. R. Larsen, U. Quass
Section 6. C. A. Belis, M. Viana, U. Quass, P. Paatero, R. M. Harrison
Section 7. C. A. Belis, B. R. Larsen, P. Paatero
Section 8. I. El Haddad, R. M. Harrison, C. A. Belis
Section 9. R. Vecchi, P. Paatero, U. Quass
Section 10. M. Viana, R. Vecchi
Section 11. R. Vecchi
Section 12. F. Amato
Section 13. A. Prévôt, S. Gilardoni, O. Favez, C. A. Belis
Section 14. O. Favez, L. Diapouli, G. Mocnik
Section 15: S. Nava
Section 16: M. Paglione, S. Decesari
Section 17: S. Takahama
Section 18: P. Salvador, S. Vratolis, D. Mooibroeck, U. Quass, F. Amato
Section 19. F. Amato, B. R. Larsen, P. Paatero, M.-I. Manousakas
Section 20. C. A. Belis, B. R. Larsen

INITIAL DRAFTING COMMITTEE

Claudio A. Belis[1] (chair and editor), Bo R. Larsen[2], Fulvio Amato[3,4], Imad El Haddad[5], Olivier Favez[6], Roy M. Harrison[7], Philip K. Hopke[8], Silvia Nava[9], Pentti Paatero[10], André Prévôt[5], Ulrich Quass[11],Roberta Vecchi[12],Mar Viana[3]

REVISION COMMITTEE

Olivier Favez[6] (editor), Claudio A. Belis[1] (chair and co-editor), Mihaela Mircea[13], Evangelia Diapouli[14], Manousos-Ioannis Manousakas[14], Stergios Vratolis[14], Stefania Gilardoni[15], Marco Paglione[15], Stefano Decesari[15], Grisa Mocnik[16], Dennis Mooibroek[17], Pedro Salvador[18], Satoshi Takahama[19], Roberta Vecchi[12], Pentti Paatero[10]

[1] European Commission – Joint Research Centre, Directorate Energy, Transport and Climate. Ispra, Italy.

[2] European Commission – Joint Research Centre, Institute for Health and Consumer Protection. Ispra, Italy.

[3] Institute of Environmental Assessment and Water Research (IDAEA), Barcelona, Spain

[4] TNO, Built Environment and Geosciences, Dept Air Quality and Climate, Utrecht, The Netherlands

[5] Laboratory of Atmospheric Chemistry (LAC), Paul Scherrer Institut, Villigen, Switzerland

[6] Institut National de l'Environnement Industriel et des Risques (INERIS), Verneuil-en-Halatte, France

[7] National Centre for Atmospheric Science, University of Birmingham, Birmingham, UK

[8] Center for Air Resources Engineering and Science, Clarkson University, Potsdam, NY, USA

[9] Department of Physics, Università degli Studi di Firenze and INFN-Firenze, Sesto Fiorentino, Italy

[10] Department of Physics, University of Helsinki, Helsinki, Finland

[11] Air Quality and Sustainable nanotechnology (IUTA), D- Duisburg, Germany

[12] Department of Physics, Università degli Studi di Milano and INFN, Milan, Italy

[13] National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), Rome, Italy.

[14] National Center for Scientific Research "Demokritos" (NCSR-DEM), Athens, Greece

[15] Institute of Atmospheric Sciences and Climate (ISAC), Bologna, Italy

[16] Josef Stefan Insitute (JSI), Ljubjana, Slovenia

[17] National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands

[18] Center for Energy, Environment and Technology (CIEMAT), Dept of Environment, Spain

[19] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

Introduction

This report contains a guide and a European harmonised protocol prepared within the framework of the Working Group 3 on source apportionment of the Forum for Air Quality Modelling in Europe (FAIRMODE). It has been initiated as a JRC initiative for the harmonisation of source apportionment with receptor models, in collaboration with FAIRMODE as well as with the European networks in the field of air quality measurements (AQUILA), and then further with the European COLOSSAL (Chemical On-Line cOmpoSition and Source Apportionment of fine aerosol) COST action. The JRC initiative also included a review of the methodologies used in Europe for source identification and intercomparison exercises for the quantitative assessment of the performance of source apportionment models.

The document, drafted and then reviewed by a group of international experts, is the revision of the version drafted by Belis et al. (2014). It is organised following the logical sequence of steps to be carried out in a source apportionment study. Sections with increasing levels of complexity make it accessible to readers with different degrees of familiarity with this topic, from air quality managers to air pollution experts and modellers. It has been conceived as a reference document that includes tutorials, technical recommendations and check lists.

EU policy context and importance of the issue

The abatement of pollution at its source is one of the overarching principles of the Thematic Strategy on Air Pollution (TSAP; Dir. 2008/50/EC, preamble). Reliable and quantitative information on pollution sources is essential for the implementation of the Air Quality Directives (AQD: Dir. 2008/50/EC and Dir. 2004/107/EC). For instance, pollution source information is required for identifying whether exceedances are due to natural sources or to road salting and sanding (arts. 20 and 21), preparing air quality plans (Annex XV A), quantifying transboundary pollution (Annex IV A), informing the public (Annex XVI) and, in the past, for demonstrating eligibility for postponement of PM_{10} and NO_2 limit value attainment (COM/2008/403).

Source Apportionment (SA) is the identification of ambient air pollution sources and the quantification of their contribution to pollution levels. This task can be accomplished using different approaches: emission inventories, source-oriented models and receptor-oriented models.

Goals/objectives of the report

The objective of this document is to disseminate and promote the best available methodologies for source identification using receptor models, and to harmonise their application across Europe.

In addition, it aims at making results of source apportionment studies more accessible to experts involved in the development and assessment of pollution source abatement measures.

Methodology

Receptor models (RMs) apportion the measured mass of an atmospheric pollutant at a given site to its emission sources by solving a mass balance equation. These models have the advantage of providing information derived from real-world measurements, including estimations of output uncertainty. However, their applicability to very reactive species is limited. RMs are extensively used for source contribution quantification at local and regional scales all over the world. In the past decade, the number of scientific publications and applications in this field has been increasing steadily, and tools have been developed with constantly improving capabilities in terms of source resolution and the accuracy of source contribution quantification (Belis et al., 2013).

The protocol presented in this document focuses on the most commonly used RMs: Chemical Mass Balance (CMB) and Positive Matrix Factorization (PMF). The CMB model is a ‘least squares’ model which estimates source contributions on the basis of the emissions' chemical composition (fingerprints) and the concentration of pollutants. The PMF model is based on uncertainty-weighted factor analysis which relies on pollutant measurements.

In addition, to promote the application of more specific methodologies under continuous development, some of the last sections are dedicated to constrained and expanded models, air mass trajectory- and wind-based models, as well as methodologies for the data treatment of aerosol mass spectrometry, multi-wavelength aethalometer, radiocarbon, proton-nuclear magnetic resonance spectroscopy, and Fourier transform infrared measurements.

Key results, deliverables, key messages

Due to the complexity of source apportionment studies, it is essential to support the final results with an appropriate description of the methodological choices available and with documentation of the objective qualitative or quantitative information that supports expert decisions. In this way, reviewers and final users (e.g. air quality managers) are provided with the elements they need to assess the relevance of the study, and other modellers have the possibility to reproduce the same approach. It is essential that only methodologies fulfilling quality standards that are in line with the objectives of the study are adopted. To that end, the information about models' performance collected in the above-mentioned intercomparison exercises provides the necessary complement to the procedures described in this document. These exercises have demonstrated that RMs provide quantitative estimations of contributions by source category with at most 50% uncertainty (Karagulian et al., 2012). It follows that SA studies that are consistent with the present protocol, in particular with regard to the quality assurance steps, can claim state-of-the-art performance in line with that observed in European-wide intercomparison exercises.

Real/potential impact and benefits to customers, users, and stakeholders

1. Quantitative estimations of pollution sources obtained with reliable and harmonised methods across Europe constitute a fundamental input for the different actors involved in the implementation of the Air Quality Directives at the local and regional scales. The present document helps to streamline the technical criteria required for accomplishing such a complex task according to the best available standards, with a view to improving the transparency and comparability of results obtained by different practitioners in different areas of Europe.
2. Expected benefits of the report for different target groups are listed hereafter:

   - the report is intended to be a reference for practitioners, providing them with clear and widely accepted criteria for model execution and the interpretation of results;

   - final users of pollution source data, such as authorities involved in air quality management, would have access to transparent and comparable information obtained with known quality standards that can be used as input data in scenario- or cost-benefit analyses;

   - harmonisation would have a positive impact on the quality and comparability of data reported by Member States to the Commission under the scheme for reciprocal exchange of information and reporting on ambient air quality (Commission Implementing Decision 2011/850/EU);

   - the report is an information dissemination tool for air quality managers and atmospheric scientists that are not familiar with this methodology.
3. In addition, the synergy between the harmonised technical protocol and the intercomparison exercises provides the basis for the continuous improvement of source identification approaches in order to keep abreast of the scientific developments in this field.
4. All the methodologies for source identification have strengths and limitations. Considering that RMs deliver independent estimates of source contributions at a given site, they can

also be used for the validation of other methodologies such as emission inventories and air quality models.

References

2011/850/EU. Commission implementing decision of 12 December 2011 laying down rules for Directives 2004/107/EC and 2008/50/EC of the European Parliament and of the Council as regards the reciprocal exchange of information and reporting on ambient air quality. (notified under document C(2011) 9068)(2011/850/EU).

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. *Atmospheric Environment* 69, 94-108.

Belis C.A., Larsen B.R. Amato F. El Haddad I, Favez O., Harrison R.M., Hopke P.K., Nava S., Paatero P. Prevot A.S.H., Quass U., Vecchi R., Viana M., 2014. European guide on air pollution source apportionment with receptor models. European Union, JRC reference reports, ISBN 9789279325144 doi: 10.2788/9307.

COM(2008) 403. Communication from the Commission on notifications of postponements of attainment deadlines and exemptions from the obligation to apply certain limit values pursuant to Article 22 of Directive 2008/50/EC on ambient air quality and cleaner air for Europe.

Karagulian F., Belis C. A., Borowiak A., 2012. Results of the European Intercomparison exercise for receptor models 2011-2012. Part I. JRC Report EUR 25727 EN.

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.

Directive 2004/107/EC of the European Parliament and of the Council of 15 December 2004 relating to arsenic, cadmium, mercury, nickel and polycyclic aromatic hydrocarbons in ambient air.

SUMMARY

Abatement of pollution at its source is one of the overarching principles of the Thematic Strategy on Air Pollution (TSAP; Dir. 2008/50/EC, preamble). Reliable and quantitative information on pollution sources is essential for the implementation of the Air Quality Directives (AQD: Dir. 2015/1480/EC, Dir. 2008/50/EC and Dir. 2004/107/EC). For instance, pollution source information is required for identifying whether exceedances are due to natural sources or to road salting and sanding (arts. 20 and 21), preparing air quality plans (Annex XV A), quantifying transboundary pollution (Annex IV A), informing the public (Annex XVI), and demonstrating eligibility for the postponement of PM_{10} and NO_2 limit value attainment (COM/2008/403).

Source Apportionment (SA) is the practice of deriving information about pollution sources and the amount they contribute to ambient air pollution levels. This task can be accomplished using three main approaches: emission inventories, source-oriented models and receptor-oriented models. The objective of this document is to present the receptor-oriented methodology, explaining its role in the identification of sources with particular reference to particulate matter, and to describe the best practices for the available and emerging methodologies with a view to promoting their harmonisation across Europe.

Receptor-oriented models (also known as receptor models (RMs)) apportion the measured mass of an atmospheric pollutant at a given site, called the receptor, to its emission sources by using multivariate analysis to solve a mass balance equation. These tools have the advantage of providing information derived from real-world measurements, including estimations of output uncertainty. However, there are limitations in their application to very reactive species. RMs are extensively used for the quantification of source contributions at local and regional scales all over the world. In the past decade, the number of scientific publications and applications in this field has been increasing steadily, and tools have been developed with improved capabilities in terms of source resolution and the accuracy of source contribution quantification.

This report is the result of the work of a group of international experts carried out within the framework of the JRC initiative for the harmonisation of source apportionment with receptor models. This initiative was launched in collaboration with the European networks in the field of air quality modelling (FAIRMODE) and measurements (AQUILA). The initiative also includes a review of the methodologies used in Europe for source identification, and intercomparison exercises for the quantitative assessment of the performance of SA models.

The structure of this document follows the logical sequence of steps to be carried out in a SA study. The organisation of the report in sections of increasing levels of complexity makes it accessible to readers with different degrees of familiarity with this topic: from air quality managers to air pollution experts and modellers. The report has been conceived as a reference document that includes tutorials, technical recommendations and check lists. However, it is not intended to substitute practitioners' experience and competence, which can only be acquired through training and working under the supervision of experts.

The report focuses notably on the most commonly used RMs: Chemical Mass Balance (CMB) and Positive Matrix Factorization (PMF) models. The CMB model is a 'least squares' model which estimates source contributions on the basis of the chemical fingerprints of the source and the concentration of pollutants. The PMF model is based on uncertainty-weighted factor analysis which relies on pollutant measurements.

In addition, to promote the application of more specific methodologies under continuous development, some sections are specifically dedicated to constrained and expanded models, air mass trajectory- and wind-based models, as well as methodologies for the data treatment of aerosol mass spectrometry, multi-wavelength aethalometer, radiocarbon, proton-nuclear magnetic resonance spectroscopy, and Fourrier transform infrared measurements.

Due to the high number of variables to be considered, SA studies are complex. Therefore, it is essential to support the final results with an appropriate description of the methodological choices made and documentation of the qualitative or quantitative information that supports expert decisions. In this way, reviewers and final users, such as local air quality managers, are provided with the elements they need to assess the relevance of every study, and other modellers have the possibility to reproduce the methodology.

Moreover, it is essential that only methodologies fulfilling quality standards that are in line with the objectives of the study are adopted. To that end, the information about models' performance collected in the above-mentioned intercomparison exercises provides the necessary complement to the procedures described in this document. These exercises have demonstrated that RMs provide quantitative estimations of contributions by source categories that are consistent with a 50% standard uncertainty criterion. It follows that SA studies consistent with the present protocol, especially with the steps concerning quality assurance, can claim state-of-the-art performance when documenting European-wide intercomparison exercises.

GLOSSARY

Chemical mass balance (CMB): models that solve the mass balance equation using effective variance least square. These are applied when the number and composition of sources are known.

Degrees of freedom: the number of independent observations minus the number of parameters estimated using them.

Factor: an independent theoretical variable calculated by linearly combining many measured dependent variables in order to describe their relationship patterns.

Factor analytical methods: multivariate techniques which do not require information on the number and composition of sources in the model input. In this document, factor analysis (FA) refers to techniques without intrinsic constraints.

Factor/source: the pollution-emitting entity identified in a SA study. Depending on the type of model used, the output may be a factor (multivariate analysis type) or a source (CMB type).

Kronecker product (denoted by $\otimes$)**:** an operation performed on two matrices which, unlike the classical matrix multiplication, does not impose limitations on the dimension of the matrices being multiplied.

Multivariate analysis: methods used to deal with datasets consisting of several measurements (variables) for each object (sample unit).

Positive matrix factorization (PMF): a specific type of factor analytical method which uses experimental uncertainty for scaling matrix elements and constrains factor elements to be non-negative.

PM_{10}, $PM_{2.5}$: particulate matter with aerodynamic diameter equal to or less than 10 and 2.5 micrometres, respectively.

Receptor models (RMs): methodology to apportion the measured mass of air pollutants in one or more sites to their emission sources by solving a mass balance equation using multivariate analysis.

Source: a source of air pollution is any human activity or natural process that causes pollutants to be released into the atmosphere.

Source apportionment (SA): the practice of deriving information about pollution sources and the amount they emit from ambient air pollution data.

Source category: a group of sources that emit pollutants with similar chemical composition and time trends.

Source contribution estimate (SCE): quantitative output of an RM expressed as mass ($\mu g\ m^{-3}$) that represents the amount of a pollutant that can be attributed to a specific source or source category.

Source profile or fingerprint: the average relative chemical composition of the particulate matter deriving from a pollution source, commonly expressed as the ratio between the mass of every species to the total PM mass.

GENERAL ACRONYMS

AMS:	Aerosol Mass Spectrometer
ACSM:	Aerosol Chemical Speciation Monitor
APS:	aerodynamic particle sizer
BDL:	below the detection limit
CEN:	European Committee for Standardisation
CC:	carbonatic carbon
CTMs:	chemical transport models
DL:	detection limit
DRUM/RDI:	Davis rotating-drum Universal-size-cut Monitoring impactor
EC:	elemental carbon
EMEP:	European Monitoring and Evaluation Programme
EPA:	Environmental Protection Agency (US)
GC-MS:	gas chromatography coupled with mass spectrometry
GF-AAS:	graphite furnace - atomic absorption spectrometry
HPLC:	high-performance liquid chromatography
IC:	ion chromatography
ICP-MS:	inductively coupled plasma- mass spectrometry
LS:	least squares
LOD:	limit of detection
OC:	Organic carbon
OM:	organic matter
OPC:	optical particle counter
PAHs:	polyaromatic hydrocarbons
PBL:	planetary boundary layer
PIXE:	particle-induced X-ray emission
PM:	particulate matter
POC:	primary organic carbon
RM:	receptor model
SA:	source apportionment
SMPS:	scanning mobility particle sizer,
SOC:	secondary organic carbon
TC:	total carbon
TOR:	Thermo Optical Reflectance
TOT:	Thermal Optical Transmission
VOCs:	volatile organic compounds
XRF:	energy dispersive X-ray fluorescence

RECEPTOR MODEL ACRONYMS

APCFA:	absolute principal components factor analysis
APCA:	absolute principal component analysis
CMB:	chemical mass balance
COPREM:	constrained physical receptor model
CPF:	conditional probability function
FA:	factor analysis
ME-2:	multilinear engine version 2
NWR:	non-parametric wind regression
PCA:	principal components analysis
PMF:	positive matrix factorization
PDRM:	pseudo deterministic receptor model
PSCF:	potential source contribution function
SoFi:	Source Finder
SQTBA:	simplified quantitative transport bias analysis
TSA:	trajectory sector analysis
TRMB	trajectory mass balance
TMBR:	trajectory mass balance regression

INTRODUCTION TO SOURCE APPORTIONMENT WITH RECEPTOR MODELS

European Guide and Harmonised Receptor Model Protocol: driving elements

This document is the result of the collaboration of leading European experts in the field of atmospheric pollution (with the additional support of the P. K. Hopke of Clarkson University, New York, United States). Its objective is to disseminate and promote the best available operating procedures for source apportionment (SA) with receptor models (RMs) and to harmonise their application across Europe. The target audiences are:

- practitioners involved in the model execution and in the interpretation of results,
- air quality managers interested in the output of RMs for the design of abatement measures,
- air quality experts and atmospheric scientists not familiar with this methodology.

The structure of this document follows a logical sequence of steps to be carried out in SA studies, with different levels of complexity accessible to readers with different levels of expertise.

This document has been conceived as a guide (including tutorials, technical recommendations and check lists) that provides relevant references to the original information sources. However, it is not meant to be comprehensive, nor intended to substitute experience and competence. Although the guide aims to promote the highest quality standards, it is subject to the intrinsic limitations of any SA methodology, which lie in the fact that the “true” contribution of sources to atmospheric pollution at a given point cannot be measured directly.

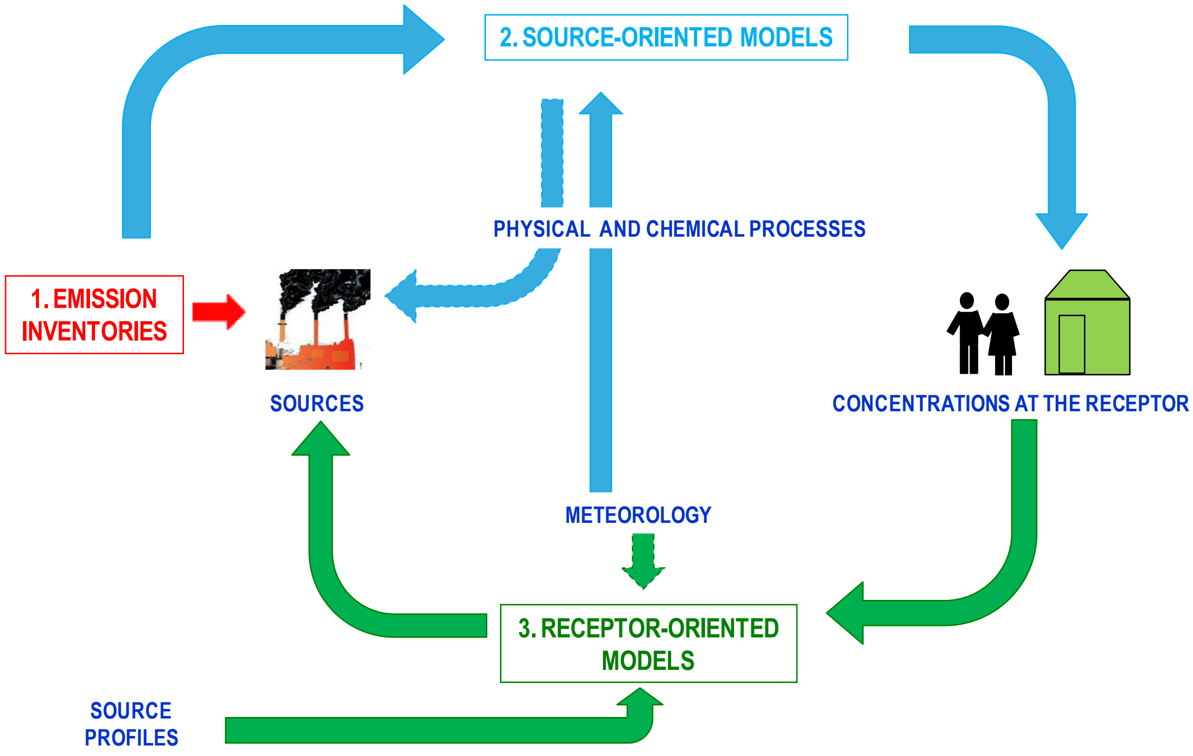
Identification of pollution sources

Source Apportionment (SA) is the practice of deriving information about pollution sources and the amount they contribute to ambient air pollution levels.

Information on pollution sources is essential to the design of air quality policies and, therefore, SA is required explicitly or implicitly for the implementation of the Air Quality Directives (IPR guidance). Activities for which identification of pollution sources is relevant include:

- Drawing up action plans
- Assessment of the effectiveness of abatement measures (before and after)
- Application for the postponement of attaining limit values (PM_{10}, NO_2)
- Quantification of pollution arising from:
    - Long-range transport
    - transboundary transport
    - natural sources
    - winter sanding and salting
- Identification of sources of pollutants that are of particular interest, e.g. polycyclic aromatic hydrocarbons (PAHs), ozone precursor hydrocarbons, elemental carbon (black carbon).

Figure I.1. Schematic representation of the different methods for source identification.



Different approaches are used to determine and quantify the impacts of air pollution sources on air quality. Commonly used SA techniques are:

- Explorative methods
- Emission inventories
- Inverse modelling
- Artificial neural networks
- Lagrangian models
- Gaussian models
- Eulerian models
- Receptor models

Exploratory methods use simple mathematical relationships and several assumptions to achieve a preliminary estimation of the source contribution.

Emission inventories are detailed compilations of the emissions from all source categories in a certain geographical area and within a specific year. Emissions are estimated by multiplying the intensity of each relevant activity (activity rate) by a pollutant-dependent proportionality constant (emission factor).

In inverse modelling, air quality model parameters are estimated by fitting the model to the observations. The inverse technique consists of a least squares optimisation with an objective function defined as the sum of squared deviations between modelled and observed concentrations.

Artificial neural networks (ANN) are sets of interconnected simple processing elements (artificial neurons) which can exhibit complex global behaviour. In order to produce a desired signal flow, algorithms designed to modulate the weights of the connections in the network are applied.

Lagrangian models use a moving frame of reference to describe the trajectories of single or multiple particles as they move in the atmosphere.

Gaussian plume models assume that turbulent dispersion can be described using a Gaussian distribution profile. This type of model is often used to estimate emissions from industrial sources.

Eulerian models encompass equations of motion, chemistry and other physical processes that are solved at points arranged on a 3D grid.

Often, the terms 'dispersion models' or 'source oriented models' are used to refer to the latter three categories. Nevertheless, there are relevant differences in how these models are applied for source identification purposes.

Receptor models (RMs) focus on the properties of the ambient environment at the point of impact, as opposed to the source-oriented dispersion models which account for transport, dilution, and other processes that take place between the source and the sampling or receptor site (Figure I.1).

What are receptor models (RMs)?

The fundamental principle of receptor modelling is that mass conservation between the emission source and the study site can be assumed, and a mass balance analysis can be used to identify and apportion sources of atmospheric pollutants. Table I.1 summarises the main characteristics of RMs.

RMs identify sources by solving the following mass balance equation:

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij} \tag{I.1}$$

where x_{ij} is the concentration of the j^{th} species in the i^{th} sample, g_{ik} the contribution of k^{th} source to i^{th} sample, f_{kj} the concentration of the j^{th} species in the k^{th} source, and e_{ij} is the residual (i.e. the difference between the measured and fitted value) term.

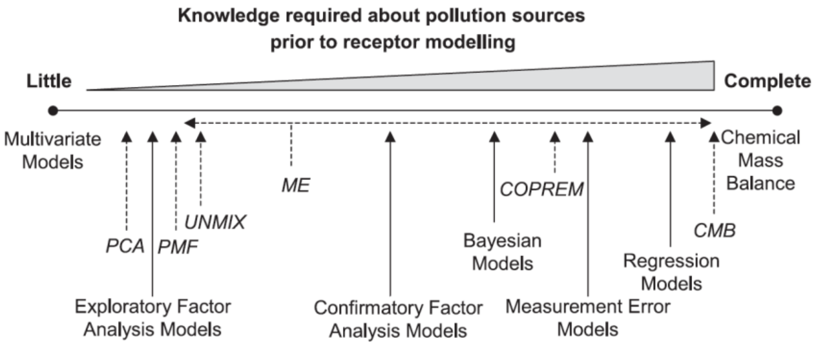
Table I.1. Main characteristics of RMs.

Use measured concentrations at the receptor (sampling site)
Make reference to the chemical mass balance principle
Are based on the solution of multilinear equations
At the first step do not consider physical and chemical processes, but evolved hybrid models can process additional information to constrain rotational uncertainty
Do not depend on emission inventories; source profiles (fingerprints) are required by certain kinds of RMs
Do not require complex meteorological and chemical processors
Require low computational intensity
Their application with reactive species requires correcting terms
Mainly used on particulate matter (PM) and seldom on hydrocarbons and inorganic gases
Appropriate for urban and regional scales

In order to find the solution, a dataset with a rather large amount of data consisting of chemical constituents (such as elemental concentrations) gathered from a number of observations (samples) is required. The larger the data matrix, the higher the chances that the model will identify distinct factors that can be identified as sources. Note that the number of required samples can vary upon a priori knowledge of the sources and RM's methodology chosen accordingly (e.g., CMB vs. PMF).

If the number and nature (composition profiles/fingerprints) of the sources in the study area are known (f_{kj}), then the only unknown term of equation (I.1) is the mass contribution of each source to each sample, g_{ik}. Solving the mass balance equation in this way was first independently suggested by Winchester and Nifong (1971) and by Miller et al. (1972). The problem is typically solved using an effective-variance least-squares approach that is now generally referred to as the chemical mass balance (CMB) model (Watson, 1979, 1984). Since then, many models and methodologies have been developed and are still under continuous evolution. RMs have been traditionally classified into those which explicitly use information about the emission fingerprints (described above) and those which do not use any a priori information on source chemical profiles (factor analysis methods).

Figure I.2. Approaches to estimate pollution sources with RMs (from Viana et al., 2008)



The main types of RMs are presented in Figure I.2 and Table I.2. A more detailed description and discussion of the most common RMs can be found in Watson et al. (2008), Viana et al. (2008), Hopke (2010), and Belis et al. (2013).

Table I.2. Types of RM (adapted from Belis et al., 2013).

Type	**Examples**
Exploratory methods	Enrichment factor, tracer method, incremental approach
Chemical Mass Balance	EPA CMB 8.2
Eigenvector-based models	PCA, UNMIX
Factor analysis without constraints	FA, APCFA
Positive matrix factorization	PMF2, EPA PMF v3, EPA PMF v5
Hybrid trajectory-based models	CPF, PSCF
Hybrid expanded models	PMF solved with ME-2 using constraints, COPREM

Legend: CMB, chemical mass balance; PCA, principal components analysis; FA, factor analysis; APCFA, absolute principal component factor analysis; PMF, positive matrix factorization; ME, multilinear engine; CPF, conditional probability function; PSCF, potential source contribution function; COPREM: Constrained physical receptor model.

In the US, RMs are officially recognised and promoted as tools for air quality management (US-EPA SCRAM), notably in the frame of dedicated monitoring networks, and a number of tools have been developed and are freely distributed by the US-EPA. RMs are also used extensively in Europe, although the lack of a common approach and documented performance limits their application to air quality policy.

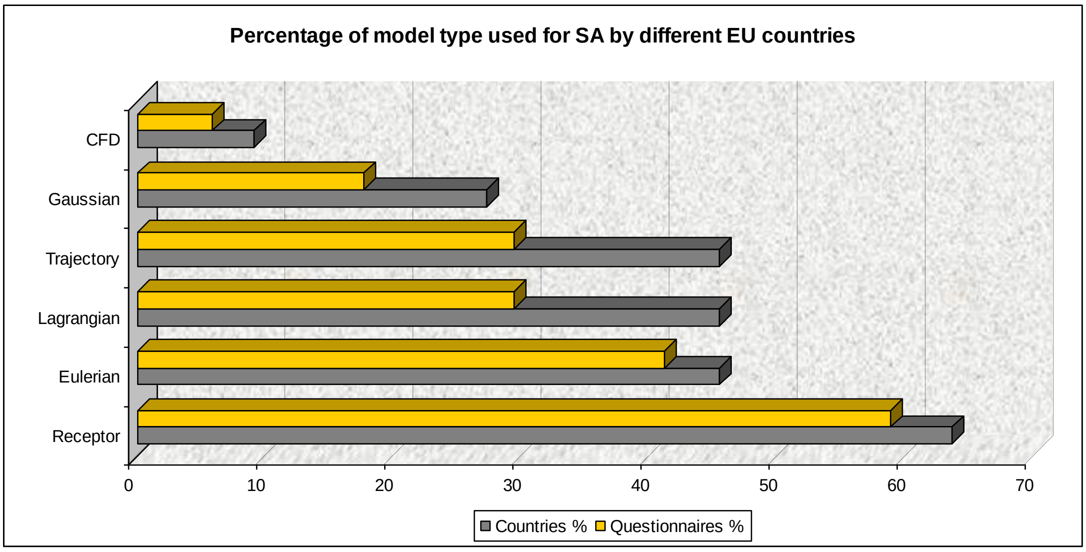
The role of RMs in the identification of pollution sources

Within the activities of the Forum for Air Quality Modelling in Europe (FAIRMODE) group on “Contribution of natural sources and source apportionment”, two surveys were carried out on the type and frequency of modelling tools that are used in Europe for source apportionment (Fragkou et al., 2012). The most recent of these surveys collected information on the use of models for the source apportionment of regulated pollutants and on the procedures used to evaluate the applied methodologies. The use of the different tools for source identification ranged from less than 20% for Gaussian models to almost 60% for receptor models (Figure I.3). Lagrangian (e.g. Lagrangian particle dispersion models) and Trajectory models were less frequently used and always complementary to other models. The use of CFD models was only reported in one case.

A study by Viana and co-authors carried out an overview of source apportionment studies in Europe from 1987 to 2007 by compiling meta-data on 71 studies (see Table 1 page 831 of Viana et al., 2008) based on a questionnaire and existing publications.

According to this study, PCA was the most frequently used model up to 2005 (30% of the studies), followed by the 'Lenschow approach' or incremental concentrations approach (11%) and back-trajectory analysis (11%). An increase in the use of PMF (13%) and the mass balance analysis of chemical components (19%) was observed from 2006 onwards.

Figure I.3: Percentage of model types used for SA by different EU countries (from Fragkou et al., 2012).



PM_{10} was the preferred target metric (46%) followed by $PM_{2.5}$ (33%) and coarse fraction ($PM_{2.5-10}$; 9%).The majority of the studies were carried out in urban background locations (53% of the studies) while industrial or kerbside sites represented 11% and 20% of the studies, respectively.

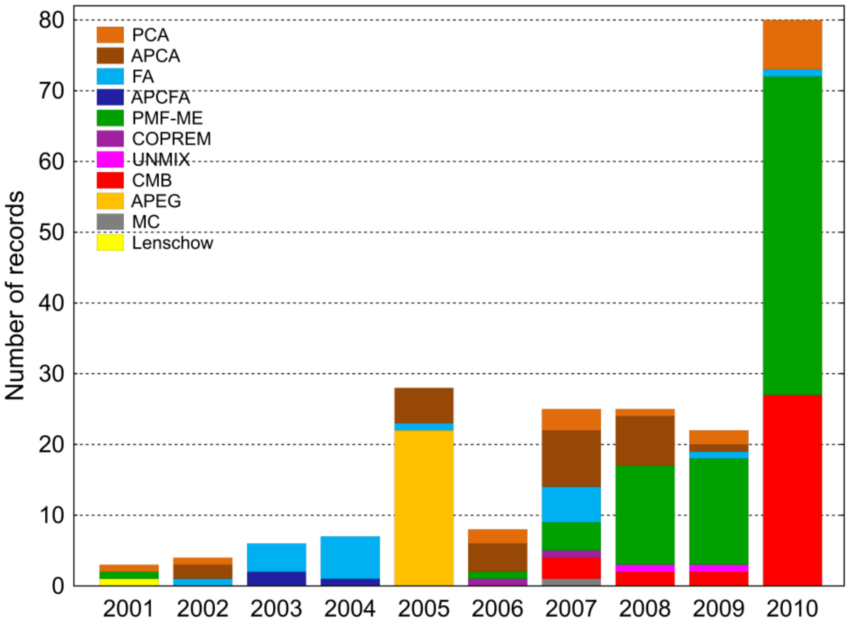
Overall, a generally good spatial coverage of SA studies over Europe, especially regarding the northern, south-eastern and south-western dimensions, was observed.

In this review, four main source categories across Europe were identified:

- Traffic sources, characterised by Carbon/Fe/Ba/Zn/Cu, often including road dust;
- Mineral/crustal matter sources with Al/Si/Ca/Fe as distinctive components;
- Sea-salt, sea-spray and marine sources associated with high Na/Cl/Mg concentrations;

- Regional-scale pollution and long-range transboundary anthropogenic pollution sources rich in either vanadium/nickel/sulphate or sulphate/nitrate/ammonium.

Figure I.4. Time trend of RM studies in Europe between 2001 and 2010/11 (from Karagulian & Belis, 2012).



A survey on the use of receptor models (RMs) for particulate matter (PM) source apportionment in Europe between 2001 and 2010, including 79 studies and 243 reported records (Karagulian and Belis, 2012), found evidence of a dramatic increase in the number of scientific publications on this topic during the past decade and an increasing number of ready-to-use tools (Figure I.4). The highest rate of increase in the number of studies coincides with the entry into force of the limit value for PM_{10} (1999/30/EC) and the target value for $PM_{2.5}$. About 60% of the studies were carried out in urban background sites, 16% in source-oriented sites (sites mainly affected by a single source), and 15% in rural sites.

In contrast with the tendency observed between 1987 and 2005, the majority of the studies were performed using Positive Matrix Factorization and Chemical Mass Balance models in the period 2001-2010 (Figure I.4).

Most of the studies were conducted in Spain, Italy and the UK. Many recent studies completed or in progress were also carried out in France.

A detailed meta-analysis of data available from previous studies is presented in the most recent review of source identification studies, which covers the period until 2012 (Belis et al., 2013). In order to compare all the SA results and to attain useful conclusions, sources have been pooled into six major categories covering those most frequently observed in the individual studies: Sea/Road Salt, Crustal/Mineral Dust, Secondary Inorganic Aerosol (SIA), Traffic, Point Sources and Biomass Burning. In addition, residential heating by coal (or coal substitutes) combustion proved to be a major PM pollution source in many areas of the new EU Member States. Residential coal combustion in small stoves and boilers has also been found to be a main source of PM_{10} and benzo(a)pyrene in certain areas of Europe (Junninen et al., 2009).

The main results of the above-mentioned review show that the field of receptor models is developing swiftly, with Positive Matrix Factorization and Chemical Mass Balance (which are the most used models) evolving towards tools with refined uncertainty treatment.

The review demonstrates that, aside from mineral dust and sea/road salt, PM_{10} and $PM_{2.5}$ derive from the same sources. Secondary pollution deriving from gas-to-particle conversion is the main PM mass and particulate organic carbon source. Therefore, in order to reduce the concentration of these pollutants it is necessary to abate the sources of secondary inorganic aerosol deriving mainly from traffic emissions and agriculture. Primary emissions from traffic and biomass burning have also been identified as causes of exceedances, especially during the cold seasons.

The review stresses the need for long-term speciated PM datasets and the characterisation of source fingerprints to further improve source identification studies. In addition, harmonisation of the different approaches would facilitate the interpretation and comparison of the results and their application in the design of abatement measures.

When to use receptor models (RMs)?

The application of RMs requires quantitative data on air pollutant concentrations, good knowledge about atmospheric processes, good command of the chemical nature of the source emissions, and competence in the use of computational tools.

RMs have mainly been used to apportion airborne particulate matter sources. Therefore, the protocol presented in this report will mostly deal with this type of pollutant. However, it is also possible to use this methodology on volatile organic compounds (VOCs), polyaromatic hydrocarbons (PAHs), inorganic gaseous pollutants and particle size distribution.

If very little information is available on the study area or if skilled staff are not available for running the standard applications, exploratory methods can be used to obtain a preliminary picture of the most relevant sources. Nevertheless, in order to achieve more accurate estimations of the source contributions and their uncertainties, a well-designed study is necessary, including field work, laboratory analyses for the chemical characterisation and data processing with standard tools.

Hybrid trajectory-based methods provide information about the geographical origin of pollutants. Advanced tools such as hybrid expanded models introduce a priori physical constraints in the model or combine different types of data (e.g. chemical and physical parameters, meteorology), making it possible to identify sources with small contributions and to better resolve similar or collinear ones.

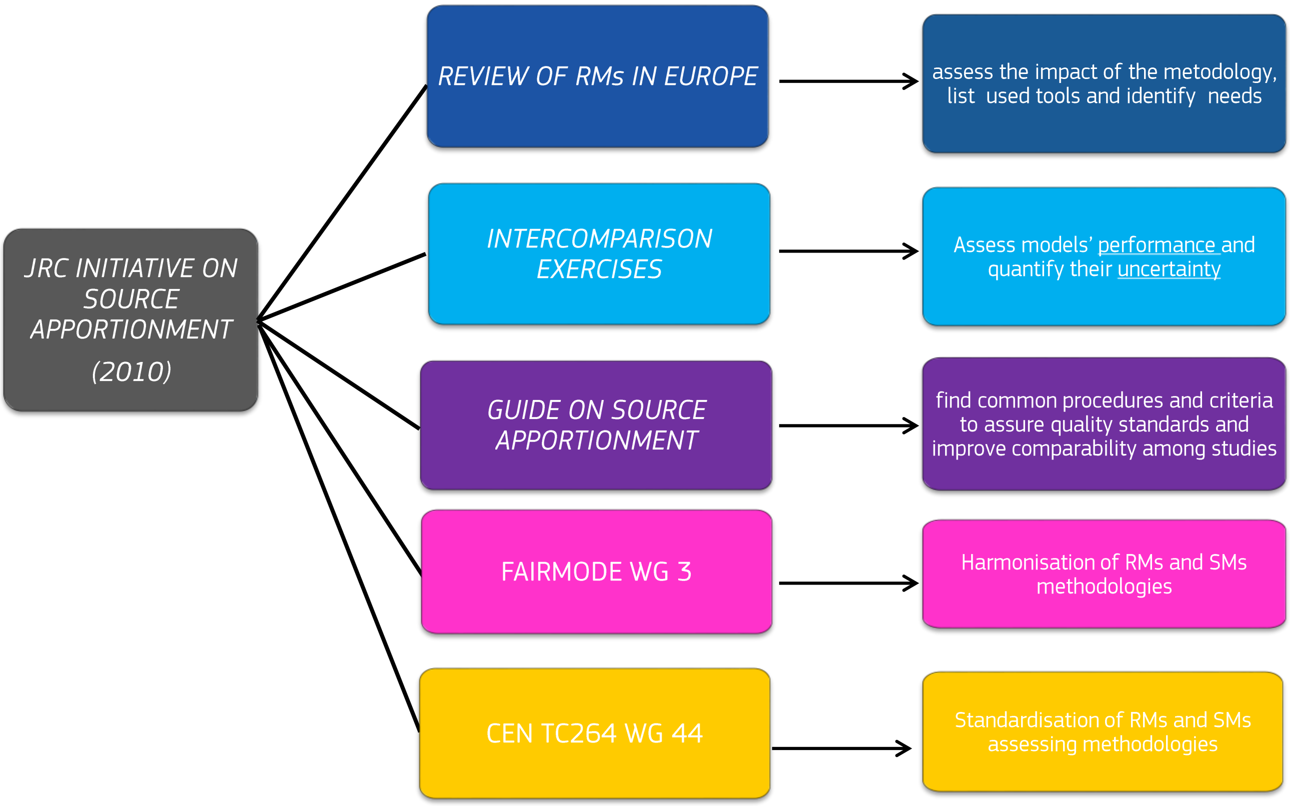
Moreover, RMs can be used in combination with independent methodologies (e.g. emission inventories, chemical transport models (CTMs)) to achieve more robust estimations by mutual validation of the outputs.

Harmonisation of receptor models

Different methodologies for identifying sources are available. However, it is difficult to establish to what extent a methodology is appropriate for a specific purpose and to quantitatively express the reliability of the results. This is mainly because the actual source contributions at a specific point are unknown. In addition, the techniques used by experts with different backgrounds need to be harmonised so as to make the results of the different studies comparable. In order to address the challenges related to the use of modelling techniques in estimating pollution sources, the JRC launched an initiative – starting from 2010 – for the harmonisation of RMs used to identify pollution sources in Europe (Figure I.5) (Belis et al., 2015a). This initiative, which involved experts from many European countries, consisted of two main activities:

- organising European-wide intercomparsion exercises for RMs and,
- developing a European harmonised technical protocol for RMs.

Figure A.5. JRC initiative on Source Apportionment.



The information about model performances collected in the above-mentioned intercomparison exercises provides the necessary complement to the procedures described in this document.

These exercises have demonstrated that RMs provide quantitative estimations of the contributions of source categories with 50% or lower standard uncertainty (Karagulian et al., 2012; Belis et al., 2015b, Belis et al., 2019).

References

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. *Atmospheric Environment* 69, 94-108.

Belis C.A., Pernigotti D., Karagulian F., Pirovano G., Larsen B.R., Gerboles M., Hopke P., 2015a. A new methodology to assess the performance and uncertainty of source apportionment models in intercomparison exercices. *Atmospheric Environment* 119, 35-44.

Belis C.A., Karagulian F., Amato F., Almeida M., et al., 2015b. A new methodology to assess the performance and uncertainty of source apportionment models II: the results of two European intercomparison exercices. *Atmospheric Environment* 123, 240-250.

Belis C. A., D. Pernigotti, G. Pirovano ,O. Favez, J.L. Jaffrezo, J. Kuenen, H. Denier van Der Gon, M. Reizer, V. Riffault, L. Y. Alleman, M. Almeida, F. Amato, A. Angyal, G. Argyropoulos, S. Bande, I. Beslic, J-L. Besombes, M.C. Bove, P. Brotto, G. Calori, D. Cesari, C. Colombi, D. Contini, G. De Gennaro, A. Di Gilio, E. Diapouli, I. El Haddad, H. Elbern, K. Eleftheriadis, J. Ferreira, M. Garcia Vivanco,, S. Gilardoni, B. Golly, S. Hellebust, P.K. Hopke, Y. Izadmanesh , H. Jorquera, K. Krajsek, R. Kranenburg, P. Lazzeri, F. Lenartz, F. Lucarelli, K. Maciejewska, A. Manders, M. Manousakas, M. Masiol, M. Mircea, D. Mooibroek, S. Nava, D. Oliveira,, M. Paglione, M. Pandolfi, M. Perrone, E. Petralia, A. Pietrodangelo, S. Pillon, P. Pokorna, P. Prati, D. Salameh, C. Samara, L. Samek, D. Saraga, S. Sauvage, M. Schaap, F. Scotto, K. Sega, G. Siour, R. Tauler, G. Valli, R. Vecchi, E. Venturini, M. Vestenius, A. Waked,, E. Yubero, 2019. Evaluation of receptor and chemical transport models for PM_{10} source apportionment. Atmospheric Environment (under review).

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.

Directive 2004/107/EC of the European Parliament and of the Council of 15 December 2004 relating to arsenic, cadmium, mercury, nickel and polycyclic aromatic hydrocarbons in ambient air.

Fragkou, E., Douros I., Moussiopoulos N., Belis C. A., 2012. "Current Trends in the use of Models for Source Apportionment of Air Pollutants in Europe." *International Journal of Environment and Pollution* 50 (1-4), 363-375.

Hopke, P. K., 2010. The application of receptor modeling to air quality data. *Pollution Atmospherique*, Special Issue 91-109.

Junninen, H., Mønster, J., Rey, M., Cancelinha, J., Douglas, K., Duane, M., Forcina, V., Müller, A., Lagler, F., Marelli, L., Borowiak, A., Niedzialek, J., Paradiz, B., Mira-Salama, D., Jimenez, J., Hansen, U., Astorga, C., Stanczyk, K., Viana, M., Querol, X., Duvall, R.M., Norris, G.A., Tsakovski, S., Wåhlin, P., Horak, J., Larsen, B.R., 2009. Quantifying the impact of residential heating on the urban air quality in a typical European coal combustion region. *Environmental Science and Technology* 43, 7964-7970.

Karagulian F. Belis C.A., 2012. Enhancing Source Apportionment with receptor models to Foster the Air Quality Directive Implementation. *International Journal of Environmental Pollution* 50, 190-199.

Karagulian F., Belis C.A., Borowiak A., 2012. Results of the European Intercomparison exercise for receptor models 2011-2012. Part I. JRC Scientific and Policy Reports. Report EUR 25727 EN. ISBN 978-92-79-28130-3.

Miller, M. S., Friedlander S. K., and Hidy G. M., 1972. A Chemical Element Balance for the Pasadena Aerosol. *Journal of Colloid and Interface Science* 39 (1), 165-176.

US-EPA Support Center for Regulatory Atmospheric Modelling (http://www.epa.gov/scram001/aqmindex.htm).

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., 2008. Source apportionment of particulate matter in Europe: A review of methods and results. *Aerosol Science* 39, 827–849.

Watson, J.G., 1979. Chemical element balance receptor model methodology for assessing the sources of fine and total suspended particulate matter in Portland, Oregon. Oregon Graduate Center. PhD Thesis.

Watson, J.G., 1984. Overview of Receptor Model Principles*. Air Pollution Control Association Journal*, Vol. 34, No.6, June 1984

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. *Journal of the Air and Waste Management Association* 58, 265–288.

Winchester, J. W., Nifong G. D., 1971. Water Pollution in Lake Michigan by Trace Elements from Pollution Aerosol Fallout. *Water, Air, and Soil Pollution* 1 (1), 50-64.

1. PRELIMINARY EVALUATION OF THE STUDY AREA

Collection of available data on atmospheric pollution

A sound source apportionment study requires careful preparation. The most important task in this step is the collection of all the relevant existing information about atmospheric pollution in the area under examination or in areas with similar characteristics. Bibliographic research should concentrate on both scientific publications and reports issued or sponsored by official bodies in charge of environmental monitoring that concern:

- emission inventories with a level of detail appropriate to the study (at least municipality or town level),
- local source profiles,
- time series at different time resolutions (daily, yearly averages) and daily profiles of pollutant levels and exceedances of legal thresholds,
- spatial distribution of pollutants, hotspots,
- meteorology at local and synoptic scale,
- previous source apportionment studies.

This step is essential to understand the nature and number of sources and the factors influencing pollutant dispersion (e.g. advection) and transformation (e.g. gas-to-particle processes). Examples of commonly-assessed sources and factors are given in Table 1.1.

The preliminary evaluation will be of great help in defining the objectives of the project and in planning the experimental work. To that end, it is also recommended that the local authorities be interviewed to understand the kind of information on pollution sources for air quality assessment and planning they are interested in, gather information on the measures that have been proposed or implemented, and understand the limitations they have encountered in their enforcement.

Table 1.1. Common source categories of particulate matter (PM) in ambient air.

marine salt	industrial emissions
crustal material	secondary ammonium sulphate
road dust	secondary ammonium nitrate
gasoline vehicle exhaust	biomass burning / wood burning
diesel vehicle exhaust	maritime transport
power plants	secondary organic aerosol

More details on the most common sources of PM in Europe can be found in Viana et al. (2008) and Belis et al. (2013)

Description of the physical system

In addition to having a good conceptual understanding of the sources in the study area, it is important to understand the physical nature of the system. The topography, natural or artificial, has a significant influence on the local source-receptor relationships (e.g. Chow et al., 2007; Belis et al., 2008), and a lack of understanding of the physical system can lead to problems in interpreting and understanding the source apportionment results. A number of aspects of the physical system should be identified and incorporated into the planning and execution of a project as well as in the analysis of the subsequently generated data, in particular:

- mountain / valley terrain,
- tall buildings,
- water bodies,
- local source complexes (grouped sources),
- isolated local sources,
- major transportation information,
- prevailing wind directions,
- distant sources.

Obstructions interfere with the direction of wind flows. People generally live in low-lying areas and thus often occupy valleys surrounded by obstructions (hills or mountains) that limit the wind directions to those found within the valley. Mountains can give rise to day-time upslope winds and night-time downslope winds. Tall buildings produce urban street canyons or block specific wind directions so that local meteorological measurements can be biased away from the actual wind directions. Water bodies also affect air flow locally (e.g. the influence of landsea breezes). Thus, understanding the geography as well as the natural and anthropogenic topography will be important in understanding source/receptor relationships for a given site.

Sources can be contained in an industrial area that is well-delimited and/or isolated from other major sources. In the case of complex sources, the emitted pollutants come from roughly the same location and, if the temporal patterns of emissions are similar from multiple sources, methods that use the covariation of measured chemical species to identify specific source types will be confused by the simultaneous variation of the receptor-site impacts of emissions from disparate sources. Isolated sources can provide the opportunity to carry out some local sampling in areas known to be highly affected by that specific source and thereby get an indication of the nature of that source. For any source, it is important to understand the nature of the activities being conducted at the site and thus, what materials are likely to be released to the environment.

Transportation systems are sources of particles and other pollutants. Vehicles with combustion engines clearly produce significant tailpipe emissions along with emissions from tyre and brake wear, re-suspension of road dust, and other related materials. Electrified systems such as trams, trains and electric buses also produce particulate emissions from the ablation of the runners that pick up the electricity from the wires and transfer it to the moving vehicle. There may also be ablation from the steel wheels rolling and stopping on the steel rails. The location of highways and other transportation systems, the nature of the vehicles operating in the vicinity of the sampler, their operating pattern (highway speed, stop-and-go, etc.) and the prevailing wind directions may all influence the measurements at the receptor site.

Prevailing wind directions determine the probability of emitted materials being transported to the measurement site. Sources with low probability wind directions are unlikely to make a large impact on a site (on the long-term average) even if they are significant emitters of the measured pollutant(s).

Although primary emissions are diluted over time and distance, secondary pollutants, e.g. produced by gas-to-particle conversion processes, can increase the concentrations over relatively long distances, particularly for species such as secondary sulphate and secondary organic aerosols that take time to form in the atmosphere.

References

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. *Atmospheric Environment* 69, 94-108.

Belis C. A., Gianelle V., De Stefani G., Colombi C.and Magnani T., 2008. The influence of orography and meteorology on the spatial distribution of pollutants in the alpine city of Sondrio (Northern Italy). *Chemical Engineering Transactions* 16, 153-160. http://www.aidic.it/cet/08/16/021.pdf

Chow, J.C., Watson, J.G., Chen, L.-A. Magliano, K., 2007. Particulate matter measurement in central California, *Air and Waste Management Association - Symposium on Air Quality Measurement Methods and Technology 2007*, pp. 179-183.

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., 2008. Source apportionment of particulate matter in Europe: A review of methods and results. *Aerosol Science* 39, 827–849.

2. DEFINING A METHODOLOGICAL FRAMEWORK

Source apportionment studies should be planned in advance according to:

- the preliminary evaluation (section 1),
- the objectives of the study,
- the available resources (funds, staff skills, time),
- the model and software to be used,
- the input data source (already available or data collection is needed),
- the required qualification of the operator and training needs.

Appropriate study planning prevents or reduces the risk of collecting useless information, missing relevant information for model execution or data interpretation, using resources inefficiently and/or building up a delay with respect to the scheduled deadlines. Useful advice for the definition of the methodological framework is available from Kim Oanh et al. (2009), Johnson et al. (2011), Watson et al. (2002) and Watson et al. (2008). The adoption of a quality management system (QMS) for the project could be useful to identify and document procedures, deliverables, responsibilities and deadlines (e.g. ISO 9001:2008).

The preliminary evaluation achieved in the previous step provides the basis for defining the objectives of the study.

At this point, the expert shall define the main questions he/she intends to answer. Subsequently, the main questions are translated into operational hypotheses and how the experimental work will contribute to test those hypotheses is clearly explained.

The objectives of the study must be in line with the available resources in terms of equipment, staff, and software. It is important to evaluate whether the required technical skills are present in the team, to make sure there is access to the technical and methodological information and, if possible, to collaborate with experts in institutions with demonstrated expertise in the field of source apportionment that can provide professional advice.

Selecting the type of model early in the planning process is also important as the kind of information to be collected depends on the model input variables:

- a chemical mass balance (CMB) model requires local source profiles as input;
- principal components analysis (PCA) and factor analysis do not require source profiles as input, but do require a very good knowledge of the study area in order to be able to interpret the output factors in terms of source categories;
- positive matrix factorization (PMF) and CMB models need an uncertainty estimation for each data entry;

- advanced models also process other types of data: e.g. meteorological variables, trajectories, day of the week, size distribution.

The choice of the model should take into account the fact that running more than one model on the same dataset can mutually validate their outputs and lead to more robust results. This may require additional time and skills.

If the input data for the selected model is not available (as is almost always the case) it is necessary to plan field activity in order to collect information on the ambient concentration of the pollutants of interest and the chemical profiles of local sources (see section 3). You may also need to collect meteorological data if this kind of information is not available close to your study site.

References

ISO 9001:2008. Quality management systems – Requirements. http://www.iso.org/iso/home/standards/management-standards/iso_9000.htm

Johnson T. M, Guttikunda S., Wells G. J., Artaxo P., Bond T. C., Russell A. G., Watson J. G., West J., 2011. Tools for Improving Air Quality Management. Formal Report 339/11. Energy Sector Management Assistance Program. Washington DC, USA.

Kim Oanh N.T., Pongkiatkul P., Upadhyay N., Hopke P.K., 2009. Designing ambient particulate matter monitoring program for source apportionment study by receptor modeling. *Atmospheric Environment* 43(21), 3334-44.

Watson J. G. , Zhu T., Chow J. C., Engelbrecht J., Fujita E. M., Wilson W. E., 2002. Receptor modeling application framework for particle source apportionment. *Chemosphere* 49 1093–1136.

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. *Journal of the Air and Waste Management Association* 58, 265–288

3. EXPERIMENT DESIGN - CRITERIA FOR SITE AND SPECIES SELECTION AND ESTIMATION OF MINIMUM NUMBER OF SAMPLES

Site Selection

For source apportionment, sites representative of the mixture of sources in a given area are preferable to sites influenced by specific sources. To establish the number and location of sources, it is necessary to study emission source distribution, wind roses and typical dispersion patterns (upwind, downwind of major sources). According to Kim Oanh et al. (2009), several sites are required to represent the different sub airsheds in a city. Furthermore, a combination of multiple sites with different characteristics is often used to obtain estimations of source contributions in an area. Using the incremental or 'Lenschow' approach, the differences in contributions from traffic, urban background and rural or regional backgrounds might be used to estimate sources in appropriate conditions (i.e., flat terrain, insignificant influence of the urban heat island effect on boundary layer height, …; Lenschow et al., 2001). A more complex option is the combination of independent source contribution estimations for different sites (e.g. Larsen et al., 2012). The orientation of sites according to the main wind directions makes it possible to assess the contributions from medium- to long-range transport (e.g. AIRPARIF and LSCE, 2012).

Representativeness of monitoring sites and heterogeneity of the study areas can be tested using geostatistical methods. These techniques assess the relationship between the difference of concentrations in and distance between different sites by fitting functions known as "variograms" (Clark & Harper, 2002; Kim et al., 2005, Hwang et al., 2008, Lagudu et al.; 2011, Kumar et al., 2012).

For point sources, stack height, temperature, mechanical buoyancy, and temporal variation of emissions are important pieces of information. Additional knowledge to evaluate the distribution of pollutants can be obtained from basic meteorological parameters and the levels of primary gaseous pollutants (Kim Oanh et al., 2009).

Species selection

The chemical species to include in the analysis should be selected according to the study objectives, the site characteristics and expected sources, considerin the available human, technical, and financial resources. Since RMs have mainly been used to apportion sources of airborne particulate matter, this document focuses on this type of pollutant (Table 3.1). Nevertheless, this methodology has also been used on datasets containing volatile organic compounds (VOCs; e.g. Elbir et al., 2007; Niedojadlo et al., 2007; Lanz et al., 2009; Baudic et al., 2016), polycyclic aromatic hydrocarbons (PAHs; e.g. Mari et al., 2010; Okuda et al., 2010; Hanedar et al., 2011; Belis et al., 2012; Tomaz et al., 2017) and inorganic gaseous pollutants (e.g. Ogulei et al., 2006; Alleman et al., 2010).

Table 3.1 Examples of input data for source apportionment with RMs

Ions	sulphate, nitrate, ammonium, chloride, Na+, Mg^{++}, K^+, Ca^{++}
Carbonaceous fractions	Total carbon (TC), elemental carbon (EC)/organic carbon (OC) total or fractions obtained in every analytical step
Elements	Na, Mg, Al, Si, P, S, Cl, K, Ca, Sc, Ti, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, Ga, Ge, As, Se, Br, Rb, Sr, Zr, Mo, Rh, Pd, Ag, Cd, Sn, Sb, Te, I, Cs, Ba, La, W, Au, Hg, Pb
Organic markers	n-alkanes, alkanoic (carboxylic) acids (especially fatty acids), aromatic carboxilic acids, levoglucosan/mannosan, PAHs, hopanes, resin acids, syringols, cholesterol
Aerosol size distribution	scanning mobility particle sizer (SMPS), optical particle counter (OPC), aerodynamic particle sizer (APS), cascade impactors, streakers, Davis rotating-drum Universal-size-cut Monitoring impactor (DRUM/RDI)
Mass fragments (m/z) concentrations	obtained with aerosol mass spectrometer (AMS) or aerosol chemical speciation monitor (ACSM) techniques and used to apportion the organic fraction (see section 13).
Optical properties	absorption coefficients to apportion C_{ff} * and C_{wb} *, light scattering at multiple wavelengths (see section 14).
Isotopic ratios	$^{14}C/^{12}C$ ratios to apportion fossil and modern C fractions (see section 15)
Radon	indicator of planetary boundary layer (PBL) mixing and long-range pollution transport

* C_{ff}: carbonaceous fraction deriving from fossil fuel and C_{wb}: carbonaceous fraction deriving from wood burning.

Chemical species that are difficult to analyse or that yield anomalous values (commonly referred to as “weak constituents” in PMF) tend to result in physically meaningless factors (Huang et al., 1999). For that reason, certain authors recommend that species considered unsuitable as source tracers be excluded. According to Ito et al. (2004), species that are not indicative of any source, or that are indicative of sources which are not relevant to the objectives of the study, can be discarded. However, the exclusion of species may lead to a loss of relevant information if we consider that the concomitant variation of a set of species could be indicative of a source even though none of them is exclusively emitted by that source. Quite often, analytical protocols such as those of X-ray fluorescence (XRF) or gas chromatography coupled with mass spectrometry (GC-MS) are able to provide multiple species output at little or no additional cost. The opportunity to take advantage of these “additional” species should not be ignored.

Using a reduced number of species could limit the number of sources that can be identified. Many multivariate methods like PMF and CMB are sensitive to collinearity. Increasing the number of species may help to reduce the collinearity between different source or factor profiles, thereby increasing the number of sources that can potentially be resolved.

In order to prevent double mass counting, redundant species should be avoided. This could be the case with sulphur (S) and sulphate, between elements and their corresponding cations or between organic carbon / elemental carbon (OC/EC) and total carbon (TC). However, soluble potassium (K) can sometimes be a useful indicator of biomass burning and thus, soluble and insoluble K can both be included in the model where insoluble K = total K – soluble K. More generally, if the two species are proportional to each other throughout the dataset, then it does not matter which one is used. However, if they do not track each other, a better separation of sources could be achieved by keeping both species in the dataset during the analysis. Double mass counting should be corrected at a later stage by retaining only one of the species in the computed factor profiles.

The traditional approach in receptor models relies on a basic set of chemical species that represents most of the particulate mass such as major ions (sulphate, nitrate, and ammonium) and the carbonaceous fraction (total organic carbon (TOC), OC/EC) plus a number of elements whose absolute and relative concentrations or specific ratios are used to identify sources (Miller et al., 1972). Although organic matter constitutes a considerable share of PM and has relevant influence on the physical and chemical properties and effects of the aerosol on health, the analytical techniques used in the past were not suitable for describing this fraction in full. The development of mass spectrometry made it possible to determine and identify organic compounds that are characteristic of certain sources called **molecular markers**, such as:

- levoglucosan (and isomers), syringol and metoxyphenols are markers for **biomass burning** (Simoneit, 2002),
- hopanes and steranes for **vehicle emissions** (Cass, 1998; Schauer et al., 2002),
- cholesterol and fatty acids for **cooking emissions** (Chow et al., 2007; Zhao et al., 2007, Schauer et al., 1999),
- benzene, di-, tri and tetra carboxylic acids, phthalates, branched ketones for **secondary organic aerosols** (Jaeckels et al., 2007; Subramanian et al., 2007).

The inclusion of molecular markers in the set of species is often desirable but requires specific sampling and analytical techniques (Wang et al., 2012; Waked et al., 2014; Srivastava et al., 2018a & 2018b). Moreover, the development and availability of instruments to measure the optical properties of the aerosol (light scattering, light absorption) and its size distribution has led to studies in which this information is combined with the chemical composition in order to better constrain the sources on the basis of their properties and the processes that pollutants undergo in the atmosphere.

Mass concentration or number concentrations in particle size bins can be used as species together with chemical species (e.g. Gu et al. 2011, Pere-Trepat et al., 2007; Pey et al., 2009; Zhou et al., 2005). For example, Ogulei et al. (2006) reported NO_3^- associated with particles larger than those associated with SO_4^{2-}, and ultra-fine particles (UFP) associated with gasoline and diesel exhausts but not with the burning of vegetation.

Number and frequency of samples

From the mathematical point of view, CMB can be executed with just one sample. In practice, many samples are required to obtain results that are representative of the variety of conditions in the study area, including the variability of sources over time. By contrast, multivariate techniques only work properly with large numbers of samples as input. According to the EPA PMF v3 User Guide (Norris et al., 2008), this method is often used on speciated $PM_{2.5}$ datasets with over 100 samples. At least 100 samples of 24-hour data of at least 20 species are recommended by Brown & Hafner (2005). Johnson et al. (2001) claim that at least 50 chemically characterised ambient samples are required for running multivariate models. According to Henry et al. (1984), the minimum number of samples (N) is the one that yields a ratio between degrees of freedom (D) and number of variables (V) that is higher than 60, while the optimal is one that leads to values above 100, according to the following equation:

$$D/V = N-(V/2-1.5) \qquad (3.1)$$

On the other hand, Thurston and Spengler (1985) propose that the number of samples should exceed the number of variables by at least a factor of three.

In practice, the minimum number of samples required to detect the latent variables cannot be established a priori as it depends on the amount of information contained in the dataset. If the relative contribution of sources were the same in all samples, analysing new samples would not add any new information to the model. Therefore, there should be enough samples to catch the variability of the sources, including samples where some sources are absent or negligible.

Small datasets simultaneously obtained for several sites may be made usable for PMF analysis by pooling them into one matrix. This can be done, for example, for several urban background sites within a city (Xie et al., 2012), a larger region with a homogeneous aerosol climate (Belis et al., 2011; Larsen et al., 2012; Mooibroek et al., 2009) or even for different site types to assess contributions of common sources (natural sources, Beuck et al., 2011).

In filter-based systems, the most common configuration is the collection of 24-hour samples. This is in part due to the requirements of reference gravimetric methods for the determination of the PM mass. In addition, a 24-hour period is considered to be representative of all the sources occurring in one day-to-night cycle and hence an appropriate unit for data elaboration. A practical reason for selecting 24-hour sampling also derives from the need to collect enough

PM for chemical analysis. This limitation is especially true for low-volume samplers when PM levels are low such as in areas that are located far from the sources or in seasons during which the main sources are not active. In urban areas, four- to six-hour sampling times usually allow for the collection of enough material for major component analyses (e.g. Vecchi et al., 2009; Bernardoni et al., 2011). This configuration provides the opportunity to detect the daily trend of most sources, making their identification with receptor modelling more feasible. With high-volume samplers, two- to four-hour samples can be sufficient.

Higher time resolutions can be achieved using semi-continuous systems for chemical analysis: particle-into-liquid samplers (PILS), semi-continuous elements in aerosol systems (SEAS), monitoring instrument for aerosols and gasses (MARGA), semicontinuous OC/EC, with resolutions ranging from a few minutes to one hour (see section 4). Streakers or DRUM/RDI samplers also provide the opportunity to select the time resolution of the analysis on size-resolved samples. Physical parameters associated with particle size or optical properties (scattering, absorption) can be obtained with time resolutions close to a minute or less.

Time resolutions in the order of seconds and minutes can be obtained with online aerosol mass spectrometers (Pratt and Prather, 2011; Drewnick, 2012).

References

AIRPARIF and LSCE, 2012. Etude de la contribution des sources des particules en Ile-de-France-Rapport Final, Edited, 171 pp. AIRPARIF, Paris. English version: http://www.airparif.asso.fr/_pdf/publications/rapport-particules-anglais-120829.pdf

Alleman L.Y., Lamaison L., Perdrix E., Robache A., Galloo J.-C., 2010. PM10 metal concentrations and source identification using positive matrix factorization and wind sectoring in a French industrial zone, *Atmospheric research* 96(4), 612-625.

Baudic A., Gros V., Sauvage S., Locoge N., et al., 2016. Seasonal variability and source apportionment of volatile organic compounds (VOCs) in the Paris megacity (France). *Atmospheric Chemistry and Physics* 16, 11961-11989.

Belis, C. A., Cancelinha, J., Duane, M., Forcina, V., Pedroni, V., Passarella, R., Tanet, G., Douglas, K., Piazzalunga, A., Bolzacchini, E., Sangiorgi, G., Perrone, M. G., Ferrero, L., Fermo, P., Larsen, B. R., 2011. Sources for PM air pollution in the Po Plain, Italy: I. Critical comparison of methods for estimating biomass burning contributions to benzo(a)pyrene. *Atmospheric Environment* 45, 7266-7275.

Bernardoni, V., Vecchi, R., Valli, G., Piazzalunga, A., Fermo, P., 2011. PM10 source apportionment in Milan (Italy) using time-resolved data. *The Science of the Total Environment* 409, 4788-4795.

Beuck H., Quass U., Klemm O., Kuhlbusch T.A.J., 2011. Assessment of sea salt and mineral dust contributions to PM10 in NW Germany using tracer models and positive matrix factorization, *Atmospheric Environment* 45, 5813-5821.

Brown S.G., Hafner H., 2005. Multivariate Receptor Modeling Workbook. Sonoma Technology Inc. Petaluma, CA. U.S.A. 236 pp. http://www.sonomatechdata.com/sti_workbooks/

Cass, G. R. 1998. Organic Molecular Tracers for Particulate Air Pollution Sources. *TrAC - Trends in Analytical Chemistry* 17 (6): 356-366.

Chow, J.C.; Watson, J.G.; Lowenthal, D.H.; Chen, L.-W.A., Zielinska, B.; Mazzoleni, L.R.; Magliano, K.L., 2007. Evaluation of organic markers for chemical mass balance source apportionment at the Fresno supersite. *Atmos. Chem. Phys.* 7(7), 1741-2754.

Chow, J.C.,1995. Critical review: Measurement methods to determine compliance with ambient air quality standards for suspended particles. *Journal of the Air and Waste Management Association.* 45(5):320-382.

Clark I., Harper W., 2002. Practical Geostatistics, 2000. Ecosse North America Llc. Columbus. Ohio, USA. 342pp ISBN 0-9703317-0-3

Drewnick, F. Speciation Analysis in On-line Aerosol Mass Spectrometry. 2012. *Analytical and Bioanalytical Chemistry* 404:2127-2131.

Elbir, T., Cetin, B., Cetin, E., Bayram, A., Odabasi, M., 2007. Characterization of volatile organic compounds (VOCs) and their sources in the air of Izmir, Turkey. *Environmental Monitoring and Assessment* 133, 149-160

Gu, J.W., Pitz, M., Schnelle-Kreis, J., Diemer, J., Reller, A., Zimmermann, R., Soentgen, J., Stoelzel, M., Wichmann, H. E., Peters, A., Cyrys, J., 2011. Source apportionment of ambient particles: Comparison of positive matrix factorization analysis applied to particle size distribution and chemical composition data. *Atmospheric Environment* 45, 1849-1857.

Hanedar, A., Alp, K., Kaynak, B., Baek, J., Avsar, E., Odman, M. T., 2011. Concentrations and sources of PAHs at three stations in Istanbul, Turkey. *Atmospheric Research* 99, 391-399

Henry R.C., Lewis C.W., Hopke P.K., Williamson H.J., 1984. Review of receptor model fundamentals. *Atmospheric Environment* 18(8), 1507-1515.

Huang S., Rahn K.A., Arimoto R., 1999. Testing and optimizing two factor-analysis techniques on aerosol at Narragansett, Rhode Island. *Atmospheric Environment* 33(14), 2169-85.

Hwang I.J, Hopke P.K., Pinto J.P., 2008. Source Apportionment and Spatial Distributions of Coarse Particles During the Regional Air Pollution Study (RAPS), *Environ. Sci. Technol.* 42, 3524–3530.

Ito K, Xue N, Thurston G., 2004. Spatial variation of PM2.5 chemical species and source-apportioned mass concentrations in New York city. *Atmospheric Environment* 38(31), 5269-82.

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. *Environmental Science and Technology* 41 (16), 5763-5769.

Johnson T.M., Suttikunda S., Wells G.J., Artaxo, P. Bond T. C., Russell A.G., Watson J.G., West J., 2011. Tools for improving Air Quality Management REPORT339/11 ESMAP.

Kim E., Hopke P.K., Pinto J.P., Wilson W.E., 2005.Spatial Variability of Fine Particle Mass, Components, and Source Contributions During the Regional Air Pollution Study in St. Louis, *Environmental Science and Technology* 39, 4172-4179.

Kim Oanh N.T., Pongkiatkul P., Upadhyay N., Hopke P.K., 2009. Designing ambient particulate matter monitoring program for source apportionment study by receptor modeling. *Atmospheric Environment* 43(21), 3334-44.

Kumar P., Hopke P.K., Raja S., Casuccio G., Lersch T.L., West R.R., 2012. Characterization and Heterogeneity of Coarse Particles Across An Urban Area, *Atmospheric Environment* 46, 339-459

Lagudu U.R.K., Raja S., Hopke P.K., Chalupa D.C., Utell M.J., Casuccio G., Lersch T.L., West R.R., 2011. Heterogeneity of Coarse Particles in an Urban Area, *Environmental Science and Technology* 45, 3288–3296.

Lanz, V. A., Henne, S., Staehelin, J., Hueglin, C., Vollmer, M. K., Steinbacher, M., Buchmann, B., and Reimann, S., 2009. Statistical analysis of anthropogenic non-methane VOC variability at a European background location (Jungfraujoch, Switzerland). *Atmospheric Chemistry and Physics* 9, 3445-3459.

Lenschow P., Abraham H.-J., Kutzner K., Lutz M., Preuss J.-D., Reichenbacher W., 2001. Some ideas about the source of PM10. *Atmospheric Environment* 35, S23-S33.

Larsen, B.R.; Gilardoni S., Stenström K., Niedzialek J., Jimenez J., Belis C. A., 2012. Sources for PM air pollution in the Po Plain, Italy: II. Probabilistic uncertainty characterization and sensitivity analysis of secondary and primary sources. *Atmospheric Environment* 50, 203-213.

Mari, M., Harrison R. M., Schuhmacher M., Domingo J. L., Pongpiachan S., 2010. Inferences Over the Sources and Processes Affecting Polycyclic Aromatic Hydrocarbons in the Atmosphere Derived from Measured Data. *Science of the Total Environment* 408 (11), 2387-2393.

Miller, M.S., S.K. Friedlander and G.M. Hidy, 1972. A chemical element balance for the Pasadena aerosol, *Journal of Colloid Interface Science* 39, 65-176.

Mooibroek D., Schaap M., Weijers E.P., Hoogerbrugge R., 2011. Source apportionment and spatial variability of PM2.5 using measurements at five sites in the Netherlands. *Atmospheric Environment* 45: 4180-4191.

Niedojadlo, A., Becker, K. H., Kurtenbach, R., Wiesen, P., 2007. The contribution of traffic and solvent use to the total NMVOC emission in a German city derived from measurements and CMB modelling. *Atmospheric Environment* 41, 7108-7126

Norris G., Vedantham R., Wade K., Brown S., Prouty J., Foley C., 2008. EPA Positive MatrixFactorization (PMF 3.0) Fundamentals & User Guide. EPA 600/R-08/108.

Ogulei D., Hopke P.K., Zhou L., Patrick Pancras J., Nair N., Ondov J.M.., 2006. Source apportionment of Baltimore aerosol from combined size distribution and chemical composition data. *Atmospheric Environment* 40(SUPPL. 2), 396-410.

Okuda, T., Okamoto, K., Tanaka, S., Shen, Z., Han, Y., and Huo, Z., 2010. Measurement and source identification of polycyclic aromatic hydrocarbons (PAHs) in the aerosol in Xi'an, China, by using automated column chromatography and applying positive matrix factorization (PMF). *Science of the Total Environment* 408, 1909-1914

Pey, J., Querol, X., Alastuey, A., Rodriguez, S., Putaud, J.P., Van Dingenen, R., 2009. Source apportionment of urban fine and ultra-fine particle number concentration in a Western Mediterranean city. *Atmospheric Environment* 43 (29), 4407-4415.

Pratt, K. A., Prather, K. A., 2012. Mass spectrometry of atmospheric aerosols—Recent developments and applications. Part II: On-line mass spectrometry techniques. *Mass Spectrometry Review* 31: 17–48.

Pere-Trepat E., Kim E., Paatero P., Hopke P.K., 2007. Source apportionment of time and size resolved ambient particulate matter measured with a rotating DRUM impactor. *Atmospheric Environmen*t 41 (28), 5921-5933.

Schauer, J. J., Kleeman M. J., Cass G. R., Simoneit B. R. T., 1999. Measurement of Emissions from Air Pollution Sources. 1. C1 through C29 Organic Compounds from Meat Charbroiling. *Environmental Science and Technology* 33 (10), 1566-1577.

Schauer, J. J., Kleeman M. J., Cass G. R., Simoneit B. R. T., 2002. Measurement of Emissions from Air Pollution Sources. 5. C 1 - C 32 Organic Compounds from Gasoline-Powered Motor Vehicles. *Environmental Science and Technology* 36 (6), 1169-1180.

Simoneit B.R.T., 2002. Biomass burning - A review of organic tracers for smoke from incomplete combustion. Applied Geochemistry 17(3), 129-62.

Srivastava D., Tomaz S., Favez O., Lanzafame G.M., Golly B., Besombes J.-L., Alleman L.Y., Jaffrezo J.-L., Jacob E., Perraudin E., Villenave E., Albinet A., 2018a. Speciation of organic fraction does matter for source apportionment. Part 1: a one-year campaign in Grenoble (France). *Science of the Total Environment*, 624, 1598-1611.

Srivastava D., Favez O., Bonnaire N., Lucarelli F., Haeffelin M., Perraudin E., Gros V., Villenave E., Albinet A., 2018b. Speciation of organic fraction does matter for source apportionment. Part 2: intensive short-term campaign in the Paris area (France). *Science of the Total Environment*, 634, 267-278.

Subramanian, R.; Donahue, N. M.; Bernardo-Bricker, A.; Rogge, W. F.; Robinson, A. L., 2007. Insights into the primary-secondary and regional-local contributions to organic aerosol and PM2.5 mass in Pittsburgh, Pennsylvania. *Atmospheric Environment* 41, 7414-7433.

Thurston, G.D., Spengler, J.D., 1985. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston, *Atmospheric Environment* 19, 9-25.

Tomaz S., Jaffrezo J.-L., Favez O., Perraudin E., Villenave E., Albinet A., 2017. Sources and atmospheric chemistry of oxy- and nitro-PAH in the ambient air of Grenoble (France), *Atmospheric Environment* 161, 144-154.

Vecchi R, Bernardoni V, Fermo P, Lucarelli F, Mazzei F, Nava S, Prati P, Piazzalunga A, Valli G., 2009. 4-hours resolution data to study PM10 in a "hot spot" area in Europe. *Environmental Monitoring and Assessment* 154(1-4), 283-300.

Waked A., Favez O., Alleman L.Y., Piot C., Petit J.-E., Delaunay T., Verlinden E., Golly B., Besombes J.-L., Jaffrezo J.-L., Leoz-Garziandia E., 2014. Source apportionement of PM10 in a north-western Europe regional urban background site (Lens, France) using positive matrix factorization and including primary biogenic emissions. *Atmospheric Chemistry and Physics*, 14, 3325-3346.

Wang, Y. Hopke, P.K., Xia, X., Rattigan, O., Chalupa, D.C., Utell, M.J., 2012. Source apportionment of airborne particulate matter using inorganic and organic species as tracers. *Atmospheric Environment* 55, 525-532.

Xie, M., Coons, T.L. Hemann, J-G., Dutton S.J., Milford, J.B. Peel. J.L. Miller S.L Kim, S-Y. Vedal, S., Sheppard, L., Hannigan, M.P., 2012. Intra-urban spatial variability and uncertainty assessment of PM2.5 sources based on carbonaceous species, *Atmospheric Environment* 60, 305-315.

Zhao Y, Hu M, Slanina S, Zhang Y., 2007. Chemical compositions of fine particulate organic matter emitted from chinese cooking. *Environmental Science and Technology* 41(1), 99-105.

Zhou, L., Hopke P. K., Stanier C. O., Pandis S. N., Ondov J. M., Pancras J. P., 2005. Investigation of the relationship between chemical composition and size distribution of airborne particles by partial least squares and positive matrix factorization, *Journal of Geophysical. Research* 110, D07S18 1-15.

4. DATA COLLECTION / FIELD WORK / CHEMICAL ANALYSIS

The analytical techniques are selected on the basis of particulate matter (PM) concentrations, required detection limits (DLs) and uncertainties, costs, access to laboratory facilities, and time resolution of the processes under study.

Sampling systems

Offline chemical analysis of PM is commonly performed using filter-based methods. Different size fractions and sampling flow rates are available. Since PM_{10} and $PM_{2.5}$ are regulated under Directive 2008/50/EC, a reference method exists (EN 12341) and most experience and types of instruments are available in Europe (Lagler et al., 2011). The advantage of using these methods is that data can be compared with those in a wide number of sites. On the other hand, the current European legislation focuses more on total PM mass concentration than on the analysis of its chemical composition (analysis of major carbon fractions and ions are requested only for few rural/remote sites). Therefore, reference methods are not always the most appropriate for source apportionment. In the US there are samplers that are specially designed for PM speciation: "RAAS" (Andersen), "MASS" (URG), "SASS" (Met One), "Partisol 2300" (Thermo), among others (Solomon et al., 2000).

The high-volume polyurethane foam (PUF) sampler, which has a large volumetric flow (hundreds L min^{-1}), may be used in parallel with low-volume PM samplers to collect samples of semi-volatile organic compounds (SVOCs) in both PM and gaseous phases (Kim Oanh et al., 2009).

Filter choice

The selection of filters is guided by the following criteria: limited artefacts, compatibility with the analytical techniques, no interactions with the sample, low level of impurities, and high efficiency.

Commonly used filter matrices are pure quartz, coated quartz and Teflon, nylon, polycarbonate, glass fibre and cellulose esters. For a detailed discussion, see Chow (1995).

Significant differences are possible between sampling systems for organic carbon and nitrates due to loss of nitrate or either deposition or loss of organic carbon. In order to test the influence of deposition and loss of semi-volatile compounds in filter-based methods, relatively complex sampling systems equipped with denuders and double filters (front filter and backup filter) are required (e.g. Subramanian et al., 2004: Vecchi et al., 2009). Unlike CEN standards, the EMEP protocols recommend the application of these methodologies for the limitation and/or the estimation of positive and negative sampling artefacts.

It is worth mentioning that quartz fibre filters are suitable for the determination of ions, and carbonaceous fractions (organic, elemental) on the same sampling support, as carbonaceous aerosols have to be analysed at elevated temperatures. For elemental composition quartz-fibre filters may present variability in filter blanks and other issues depending on the specific analytical technique used. PTFE/polycarbonate filters are the most convenient option to assess elements with a proper accuracy

Most common analytical techniques

Organic carbon and elemental carbon (OC, EC), either total or by single temperature steps, are commonly measured using thermal-optical methods. These methods take advantage of the different behaviour of the various carbonaceous fractions (i.e. OC, EC, and carbonate carbon) when exposed to elevated temperatures and to light. OC evolves at lower temperatures than EC, while the latter absorbs light more importantly than the former. The main differences between the existing thermal-optical methods (e.g. "NIOSH", "IMPROVE" and "EUSAAR") rely mainly on the temperature programs and on the devices used for optical measurements: Thermal Optical Transmission (TOT) or Thermo Optical Reflectance (TOR). More information can be found in Chow et al. (2004), Cavalli et al. (2010), and the European standard EN 16909.

The methods most used for anions and cations are ion chromatography (IC) or automated colorimetric analysis (EN 16913).

For inorganic elements, inductively coupled plasma - mass spectrometry (ICP-MS) and graphite furnace - atomic absorption spectrometry (GF-AAS) which are the reference methods for the determination of metals (As, Cd, Ni and Pb) in PM_{10} (Standard EN 14902). Although some alteration cannot be excluded due to vacuum and slight heating (Yatkin and Gerboles, 2012), energy dispersive X-ray fluorescence (XRF) is commonly used in source apportionment because it covers many elements (from Na to U), does not require sample pre-treatment and does not destroy the samples. It also has good accuracy and repeatability, and automatisation of the analysis makes it possible to treat high numbers of samples with reduced costs. A similar technique, particle-induced X-ray emission (PIXE) is also suitable. Differences in detection limits (DL), when compared to XRF, are due to intrinsic features of the two techniques, such as different ionisation cross-sections for photons or protons and differences in the intensities of the continuous background (Calzolai et al., 2008). PIXE is more powerful than XRF in analysing very small samples (i.e. size-segregated samples, high time-resolution samples or those collected in remote areas). The main limitation is due to the availability of beam time at the accelerator facility where PIXE analysis can be carried out.

Organic compounds

PAH levels in PM_{10} are regulated under Directive 2004/107/EC. The application of ISO standard 12884 is recommended but there is no reference method in this case. Either GC-MS or high-performance liquid chromatography (HPLC) methods are used for these compounds. Offline GC-MS is used to characterise a wide range of organic compounds (see table 3.1). More recently, the sensitivity of thermal desorption GC-MS methods has improved and, when combined with in situ derivation, enables the identification of polar and non-polar components (Laskin et al., 2012).

There is a variety of techniques for the determination of anhydrosugars (e.g. levoglucosan). GC-MS-based methods have been extensively used, but different types of HPLC and IC techniques have also been proposed (Schkolnik and Rudich, 2006).

Advanced User Box

Online monitoring techniques for PM mass concentration and chemical composition

In online systems, sampling and analysis are integrated into a single instrument.

The automated determination of PM mass concentrations (EN 16450) relies mostly on three operating principles: Tapered Element Oscillating Microbalance equipped with Filter Dynamics Measurement System (TEOM-FDMS), beta gauges, and optical methods.

Examples of online analytical instrumentation for ions are:

a) Particle-Into-Liquid Sampler (PILS) developed by the Georgia Institute of Technology (Weber et al., 2001) coupled with ion chromatography.

b) Monitoring instrument for aerosols and gasses (MARGA; Khylstov et al., 1995).

For the analysis of elements, the following methods are available: a) Semi-continuous elements in aerosol system (SEAS), developed by University of Maryland (Kidwell and Ondov, 2001), and b) a modified version "SEAS II" and "KSEAS" (Lee et al., 2011). Field-deployable systems also exist to collect airborne particles and apply semi-continuous XRF analysis to the samples (e.g., Xact 620, Cooper Environmental Services).

For carbonaceous fractions, a semi-continuous OC/EC analyser is available from Sunset (Bauer et al., 2009), although the measurements are not always fully comparable with those obtained with the offline method (Belz et al., 2012). A Total Carbon Analyzer system (TCA) is also currently developed by Magee Scientific.

Optical techniques for monitoring the aerosol light absorbance, such as the Multi-Angle Absorption Photometer (Petzold and Schönlinner, 2004) or the Aethalometers (Hansen et al. 1984), are commonly used to estimate equivalent carbon, i.e. the main absorber of visible light among aerosol components.

Finally, as described in section 13, organic mass spectra can be obtained routinely using Aerosol Mass Spectrometers (AMS) or Aerosol Chemical Speciation Monitors (ACSM).

Local source profiles (fingerprints)

Chemical fingerprints of local sources are of utmost importance in SA studies. The characterisation of the most relevant sources in the study area should be included in the work programme. Considering that local source characterisation is resource consuming, it is possible to adopt fingerprints available from previous work in similar areas or obtained from source profile repositories (such as SPECIATE/US-EPA: http://www.epa.gov/ttnchie1/software/speciate/ and SPECIEUROPE/JRC: http://source-apportionment.jrc.ec.europa.eu/specieurope/index.aspx, Pernigotti et al., 2016). The sample collection varies from source to source. For pollutants deriving from combustion processes, samples collected directly from the stack or exhaust at temperatures much higher than that of ambient air may lead to biases due to the absence of the condensed fraction in the particulate phase. To overcome this pitfall, it is possible to dilute the emissions with a known volume of clean air. An alternative is to sample the plume at a distance that allows the effluent to dilute and cool down to near ambient temperatures.

Source-oriented monitoring stations can be used to characterise the source emissions if periods in which other sources influence the sample are excluded from the analysis. Characterisation of mobile sources can be obtained with samples collected in the lab (e.g. Montero et al., 2010; Adam et al., 2011), on the road (e.g. Georgios et al., 2012) or in tunnel experiments (e.g. El Haddad et al., 2009).

Re-suspension of road dust and contributions deriving from industrial dust can be estimated by sampling deposits directly from the ground (Ashbaugh et al.,2003; Amato et al., 2009; Colombi et al., 2010). Samples of vehicle parts' wear (tyres, brakes, clutch) can also be obtained directly by abrasion in the laboratory (e.g. Sjödim et al., 2010).

References

Adam T., A. Farfaletti, L. Montero, G. Martini, U. Manfredi, B. R. Larsen, G. De Santi, A. Krasenbrink, C. Astorga., 2010. Chemical characterization of emissions from modern two-stroke mopeds complying with legislative regulation in Europe (EURO-2). *Environmental Science and Technology* 44, 505-512.

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. Atmospheric Environment 43.

Ashbaugh, L.L., Carvacho, O.F., Brown, M.S., Chow, J.C., Watson, J.G., Magliano, K.L., 2003. Soil sample collection and analysis for the Fugitive Dust Characterization Study. *Atmospheric Environment*, 37(9-10),1163- 1173.

Bauer JJ, Yu XY, Cary R, Laulainen N, Berkowitz C., 2009. Characterization of the sunset semi-continuous carbon aerosol analyzer. *Journal of the Air and Waste Management Association* 59(7), 826-33.

Belz G., A. Cinieri, P.R. Dambruoso, B.E. Daresta, G. de Gennaro, A. Giove, G. Miglietta, R.M. Nacci and C. Tortorella. 2012. Comparison between offline and semi-continuous analyzers for aerosol organic and elemental carbon detection. European Aerosol Conference 2012 Granada. Abstract A-WG08S1P20.

Calzolai, G., Chiari, M., Lucarelli, F., Mazzei, F., Nava, S., Prati, P., Valli, G., Vecchi, R., 2008. PIXE and XRF analysis of particulate matter samples: an inter-laboratory comparison. *Nuclear Instruments and Methods in Physics Research B*, 266, 2401–2404.

Cavalli, F., Viana, M., Yttri, K.E., Genberg, J., Putaud, J.P., 2010. Toward a standardised thermal-optical protocol for measuring atmospheric organic and elemental carbon: the EUSAAR protocol. *Atmopheric Measurement Techniques* 3, 79–89.

Chow, J.C.,1995. Critical review: Measurement methods to determine compliance with ambient air quality standards for suspended particles. *Journal of the Air and Waste Management Association*, 45(5):320-382.

Chow J., J. Watson, Chen L.W. A, Patrickarnott W., Moosmüller H., 2004. Equivalence of Elemental Carbon by Thermal/Optical Reflectance and Transmittance with Different Temperature Protocols, *Environmental Science and Technology* 38, 4414-4422.

Cooper Environmental Services. http://cooperenvironmental.com.

El Haddad I., Marchand N., Dron J., Temime-Roussel B., Quivet E., Wortham H., Jaffrezo J. L., Baduel C., Voisin D., Besombes J. L., Gille G., 2009. Comprehensive primary particulate organic characterization of vehicular exhaust emissions in France, *Atmospheric Environment* 39, 6190-6198.

EN 12341:2014: Ambient air quality – Standard gravimetric measurement method for the determination of the PM$_{2.5}$ mass fraction of suspended particulate matter, European Standard, CEN, Brussels

EN 16450:2017: Ambient air quality – Ambient air - Automated measuring systems for the measurement of the concentration of particulate matter (PM10; PM2.5), European Standard, CEN, Brussels

EN 16909:2017: Ambient air quality – Measurement of elemental carbon (EC) and organic carbon (OC) collected on filters, European Standard, CEN, Brussels

EN 16913:2017: Ambient air quality – Standard method for measurement of $NO3^-$, $SO4^{2-}$, Cl^-, NH4+, Na+, K+, Mg^2+, Ca^2+ in PM2,5 as deposited on filters, European Standard, CEN, Brussels

Georgios F., Martini G., Manfredi U., Marotta A., Krasenbrink A., Maffioletti F., Terenghi R., Colombo M., 2012. Assessment of on-road emissions of four Euro V diesel and CNG waste collection trucks for supporting air-quality improvement initiatives in the city of Milan. *Science of the Total Environment* 426, 65–72.

Hansen, A. D. A., Rosen H., Novakov T. 1984. "The Aethalometer - an Instrument for the Real-Time Measurement of Optical Absorption by Aerosol Particles." *Science of the Total Environment,* 36 (C): 191-196.

Khlystov, A., Wyers, G.P., Slanina, J., 1995. The steam-jet aerosol collector. *Atmospheric Environment* 29, 2229-2234.

Kidwell, C.B., Ondov, J.M., 2001. Development and evaluation of a prototype system for collecting sub-hourly ambient aerosol for chemical analysis. *Aerosol Science and Technology* 35, 596–601.

Kim Oanh N.T., Pongkiatkul P., Upadhyay N., Hopke P.K., 2009. Designing ambient particulate matter monitoring program for source apportionment study by receptor modeling. *Atmospheric Environment* 43 (21), 3334-44.

Lagler F., Belis C.A., Borowiak A., 2011. A Quality Assurance and Control Program for PM2.5 and PM10 measurements in European Air Quality Monitoring Networks. EUR Report EUR 24851 EN. http://bookshop.europa.eu/en/a-quality-assurance-and-control-program-for-pm2.5-and-pm10-measurements-in-european-air-quality-monitoring-networks-pbLBNA24851/

Laskin, A., Laskin J., Nizkorodov S. A., 2012. Mass Spectrometric Approaches for Chemical Characterisation of Atmospheric Aerosols: Critical Review of the most Recent Advances. *Environmental Chemistry* 9 (3), 163-189.

Lee D S., Lee B., Eom J. W., 2011. A compact semi–continuous atmospheric aerosol sampler for elemental analysis: A preliminary result. *Atmospheric Pollution Research* 2, 506-512

Montero L., Duane M., Manfredi U., Astorga C., Carriero M.; Martini G., Krasenbrink A., Larsen B.R., 2010. Hydrocarbon Emission Fingerprints from Contemporary Vehicle/Engine Technologies with Conventional and new Fuels. *Atmospheric Environment* 44, 2167-2175.

Pernigotti D., Belis C., Spano L., 2016. SPECIEUROPE: The European data base for PM source profiles, *Atmospheric Pollution Research* 7, 307-314.

Petzold, A. Schönlinner, M., 2004. Multi-angle absorption photometry- A new method for the measurement of aerosol light absorption and atmospheric black carbon, Journal of Aerosol Science, 35, 421–441.

Schkolnik G. and Y. Rudich, 2006. Detection and quantification of levoglucosan in atmospheric aerosols: A review. Analytical and Bioanalytical Chemistry 385, 26–33.

Sjödim A., Ferm M., Björk A., Rahmberg M., Gudmindsson A., Swietlicki E., Johanson C., 2010. Wear particles from road traffic – a field, laboratory and modeling study. Final Report. Swedish Environmental Research Institute. 97 pp.

Solomon P. A., Mitchell W., Tolocka M., Norris G., Gemmill D., Wiener R., Vanderpool R., Murdoch R., Natarajan S., Hardison E., 2000. Evaluation of PM2.5 Chemical Speciation Samplers for Use in the EPA National PM2.5 Chemical Speciation Network. EPA Rep. EPA-454/R-01-005, Off. of Air Qual. Plann. and Stand., Research Triangle Park, N. C.

SPECIATE (http://www.epa.gov/ttn/chief/software/speciate/)

SPECIEUROPE (http://source-apportionment.jrc.ec.europa.eu/specieurope/index.aspx)

Subramanian R., Khlystov A.Y, Cabada J.C., Robinson A.L., 2004. Positive and Negative Artifacts in Particulate Organic Carbon Measurements with Denuded and Undenuded Sampler Configurations. *Aerosol Science and Technology*, 38:27–48

Vecchi R., Valli G., Fermo P., D'Alessandro A., Piazzalunga A., Bernardoni V., 2009. Organic and inorganic sampling artefacts assessment. Atmospheric Environment, 43, 1713-1720.

Weber, R.J., Orsini, D., Duan, Y., Lee, Y.-N., Klotz, P., Brechtel, F., 2001. A particle-into liquid collector for rapid measurements of aerosol chemical composition. *Aerosol Science and Technology* 35, 718–727.

Yatkin S. and M. Gerboles. 2013. Investigation of the mass loss of water soluble ions caused by EDXRF analysis of PM10 loaded filters. *Atmospheric Environment* 73, 159-168.

5. KNOWING YOUR DATASET: BASIC STATISTICS

Before starting any kind of data treatment, it is good practice to make some summary plots and run some simple tests to gain an overview of the relationships between variables and how they change from sample to sample. Many commercial and free software applications are available that can carry out routine statistic tests (e.g. Statistica, Matlab, R, SPSS).

Central and dispersion statistics

Box and whisker plots are useful to visualise central values of your variables (mean, median) and the dispersion of your data around the central values (quartiles, minimum and maximum values).

Check the statistical distribution that best describes your data

Quite often the air pollution data can be better described using a log-normal distribution rather than a normal one. Many statistical tests assume that data is normally distributed even though small deviations from normality are acceptable. In order to better assess the results of standard statistical tests, knowing the statistical distribution of your data could be useful. Box and whisker plots give a visual overview of the data spread that enables a preliminary assessment of the distribution. Visual tests of normality such as histograms, probability plots and normal probability plots are also useful. For an in-depth evaluation, normality tests such as Kolmogorov-Smirnov (Massey, 1951) or Shapiro-Wilk W (Shapiro and Wilk, 1965; Royston, 1982) can be applied.

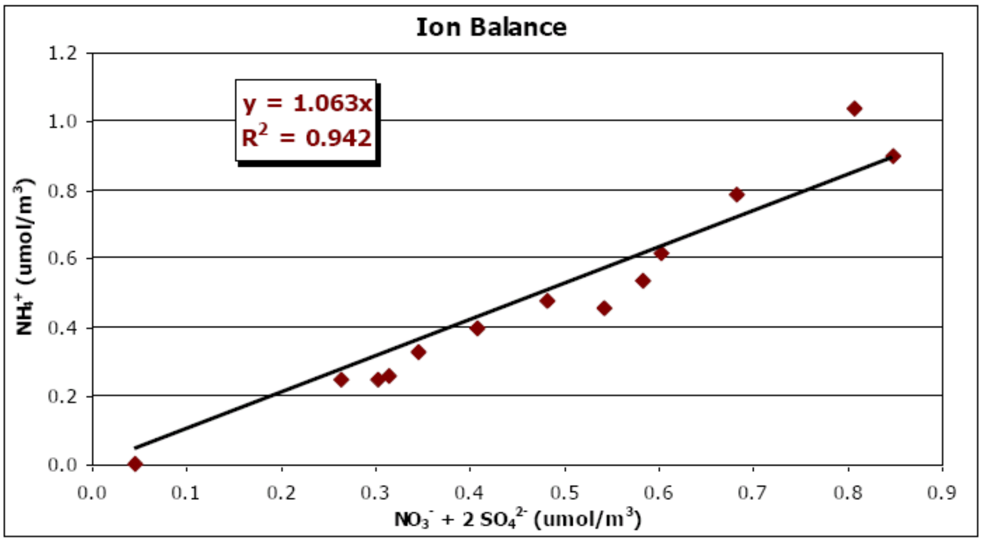
Correlation matrices

The correlation between variables can be visually assessed using scatter plots. This is particularly useful to identify anomalous data points (suspected outliers) that may affect the correlation. However, when many variables are involved the use of correlation matrices reporting the Pearson correlation coefficient (r) and related statistics for every possible pair of variables is a useful exploratory technique, provided the influence of outliers has been evaluated (see below).

Linear regression

Ordinary Least Squares regression is the simplest and quickest technique to more in-depth exploration of the association between two variables. The evaluation of the curve parameters (intercept and slope and the determination coefficient (r^2)) provides useful preliminary information to describe the (linear) relationship between the variables considered. Precaution should also be taken concerning outliers.

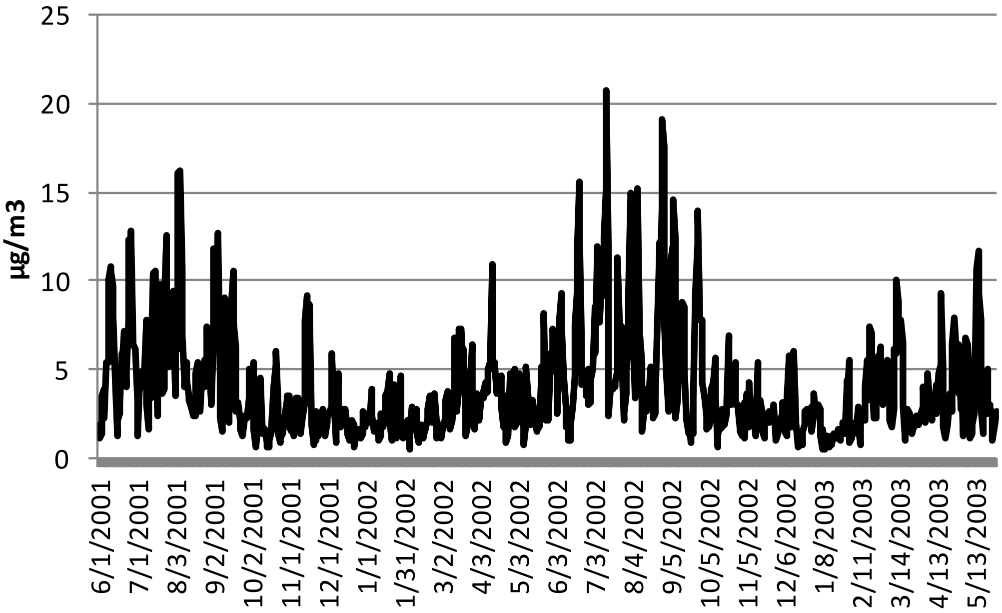
Figure 5.1 Linear regression to test ion balance in PM (B. Larsen, unpublished).



Time trends

Plotting time trends of the variables makes it possible to identify regular patterns in data (e.g. seasonality, influence of the day of the week) or extraordinary events that probably indicate the influence of specific sources influencing the study area for short periods (e.g. Saharan events, wild fires). In addition, when hourly data are available, characteristic daily profiles of certain species can be used to identify specific sources (e.g. a peak of traffic markers during rush hours).

Figure 5.2 Seasonal time trends in sulphates at Saint Louis supersite (Lee et al., 2006).



Outliers

Values that do not follow the distribution of data with similar characteristics are referred to as outliers. They may reflect genuine properties of the studied system or derive from measurement errors or anomalies that are not relevant for the model. Outliers can be extreme values or values with unusual relationships with other variables (e.g. ratio).

In some statistics software applications, values above or below the quartiles at a distance of 1.5 - 2 times the interquartile range (height of the box) are labelled as outliers. Of the analytical tests to identify outliers (in normal distributions), the most commonly used is the Grubbs test. This is based on the difference between the mean of the sample and the most extreme or the two most extreme data values, considering the standard deviation (Grubbs, 1950, 1969). These tests help the practitioner to decide whether these data provide useful information on sources or whether they only introduce noise into the model. It is good practice to report the outliers excluded from the analysis and the reason for their exclusion.

Identify samples of special interest

In source apportionment (SA), it is important to distinguish between a one-off event with a unique profile and an episode that occurs due to the increased contribution from a source with a known profile that is already present in other samples. The analysis of ancillary data is useful to investigate the possible causes of anomalous samples identified with the previous techniques. Meteorological variables such as wind direction, precipitation, or extraordinary events such as forest fires, fireworks, of volcano eruptions may influence the levels of the studied pollutant for short periods.

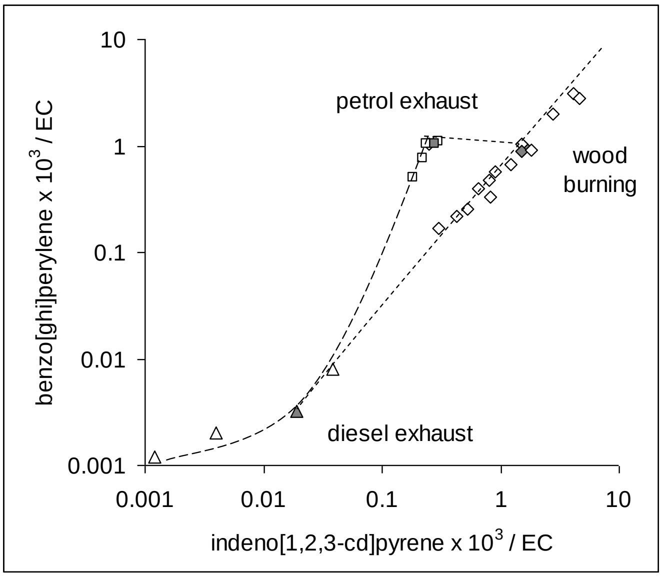
Spatial distribution

Spatial patterns can be only assessed when many sites are available. At this stage of the study, it can be checked whether the spatial variations of the chemical and physical properties of the aerosol are coherent with geographical gradients in variables that influence the emission of concentrations of atmospheric pollutants (e.g. NaCl is expected to be higher in sites close to the coast; e.g. Schaap et al., 2010).

Ratio-ratio scatter plotting

By representing the concentration (in ambient PM) of two receptor species in a scatter plot, descriptive information can be obtained for a SA dataset in which only one or two sources (or source types) contribute to these species. The data points in the plot will be distributed in an ordered manner between edges, delimited by the emission factors of these pairs of species for each source (type). The advantage of visualising concentrations of receptor compounds in PM that are normalised to concentrations of reference compounds (e.g. EC) in two-dimensional scatter plots was first demonstrated by Robinson et al. (2006).

Figure 5.3. Ratio–ratio plots using data on B(ghi)P, Ind(123)P and EC to visualise the potential contribution of three source scenarios for ambient PM.



Emission source data – triangles: diesel exhausts; squares: petrol exhausts; rhombs: wood burning (from Belis et al., 2011)

References

Belis, C. A., Cancelinha, J., Duane, M., Forcina, V., Pedroni, V., Passarella, R., Tanet, G., Douglas, K., Piazzalunga, A., Bolzacchini, E., Sangiorgi, G., Perrone, M. G., Ferrero, L., Fermo, P., and Larsen, B. R., 2011. Sources for PM air pollution in the Po Plain, Italy: I. Critical comparison of methods for estimating biomass burning contributions to benzo(a)pyrene. *Atmospheric Environment* 45, 7266-7275.

Grubbs F. E., 1950. Sample Criteria for Testing Outlying Observations, *Annals of Math. Statistics* 21 27-58.

Grubbs F. E., 1969. Procedures for Detecting Outlying Observations in Samples, *Technometrics* 11(1), 13-14.

Lee, J. H., Hopke, P. K., and Turner, J. R., 2006. Source identification of airborne PM2.5 at the St. Louis-Midwest Supersite. *Journal of Geophysical Research D: Atmospheres* 111.

Massey, F. J., 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46, 68-78.

Robinson, A. L., Subramanian, R., Donahue, N. M.m Rogge, W. F., 2006. Source Apportionment of Molecular Markers and Organic Aerosol1. Polycyclic Aromatic Hydrocarbons and Methodology for Data Visualization, *Environmental Science & Technology* 24, 7803-7810.

Royston, J. P., 1982. An Extension of Shapiro and Wilk's W Test for Normality to Large Samples. *Applied Statistics* 31(2), 115-124.

Schaap, M., Weijers, E.P., Mooibroek, D., Nguyen, L., Hoogerbrugge, R.,2010. Composition and origin of Particulate Matter in the Netherlands Results from the Dutch Research Programme on Particulate Matter. Report 500099007/2010. PBL

Shapiro S. S., Wilk. M. B., 1965. An Analysis of Variance Test for Normality (Complete Samples)." *Biometrika* 52(3/4), 591-611.

6. PRELIMINARY DATA QUALITY CHECKS

Missing values

In order to run a multivariate analysis, the entries in the dataset must comply with minimum requirements. This applies in particular for factor analysis. It is a common misconception that negative or zero concentrations are harmful for factor analyses. If a true value is zero or near zero, then there is a probability that the corresponding measured value will be negative. Such negative values should be kept in the dataset. They may be rejected only if their confidence interval does not include zero, which would obviously indicate a measurement error. If negative values are truncated to zero, then a modelling error is caused, and the data becomes biased. 'Least squares' (LS) methods are not appropriate for such kind of data. Unfortunately, some measurement techniques are not able to produce unbiased near-zero values. How to deal with such biased values is still an open question. The most promising approach in Multilinear Engine 2 (ME-2) seems to be to use error model code -16 (see box below).

Zero or negative *uncertainties* have no physical meaning and therefore should be excluded from the input file or replaced by reasonable values. Since it is not possible to perform the analysis when empty cells are present in the input data matrix, missing values should be handled in advance by the operator. The simplest choice is to cancel the row (sample) or the column (species) from the input matrix. However, this may cause the loss of important information. An alternative approach is to substitute missing values with estimated values, such as the mean, the median or the geometric mean of the measured concentrations of the species, in all the samples of that particular study site (Polissar et al., 1998). The procedure by Polissar et al. (1998) is often used without testing its validity for any given dataset. In the EPA Unmix 6.0 receptor model, for instance, there is an automated subroutine that substitutes missing values using the maximum and minimum ratios of the variable for which the value is missing (EPA Unmix 6.0 user manual). More sophisticated methods - such as multiple imputation, using statistical relations between data (Rubin, 1976 & 1977) - can also be used to substitute missing values (Hopke et al., 2001; Bacarelli et al., 2005, Ibrahim et al., 2005; Yang et al., 2005; Mooibroek et al., 2016).

It should be noted that for any receptor model, the more missing values are reconstituted, the greater the uncertainty of the source contribution estimates. As a rule of thumb, missing values substituted for a given species should not be more than half of the samples (Brown & Hafner, 2005). Scientists should find, for every dataset, the most suitable uncertainties of the substituted values to avoid distorting the model.

Values below the detection Limit

Values below the detection limit (BDL) of the analytical method should be used if they are available. If values are not provided by the laboratory they can be substituted either by zero (or by a value sufficiently close to zero), by the detection limit itself or by a fraction of the detection limit. The most common practice is to substitute BDL values with half of the detection limit (Polissar et al., 1998). Substitution of BDL values only makes sense if the number of values above the detection limit of that species in the dataset is sufficient to provide information about sources. It has been suggested that it is only worth including species that present more than 50% of BDL values in the data treatment if the signal-to-noise ratio is reasonable (see below) or the species is a tracer (Brown & Hafner, 2005).

Note: the official nomenclature (IUPAC, 1997-2006) defines ***detection limit*** (DL) as the minimum value that can be distinguished from the blanks, and ***limit of detection*** (LOD) as the blank value plus a multiple of the standard deviation of this measurement. In analytical chemistry, LOD is commonly considered to be the lowest analyte concentration at which detection is feasible; measurements below that value are reported as “<LOD”.

Advanced User Box

There are cases in which the substitution of many values below the detection limit in several trace species creates an artificial factor containing trace species with a characteristic pattern. This “ghost factor” is generated by the model to fit the substituted values in all those species in which they occur simultaneously. When using ME-2-based analysis tools, it is possible to avoid such distortions by using the special error model code -16 (Paatero, 2000) for all substituted data points. This code stipulates that all fitted values below the detection limit are to be considered a perfect fit, with Q contribution (see section 9) equal to zero. This alternative is not yet implemented in version 3.0 of the US - EPA Positive Matrix Factorization (PMF), but it may be used when controlling ME-2 using home-made scripts.

Signal-to-noise

The signal-to-noise ratio (S/N) is defined as the power ratio between a desired signal (S, meaningful information) and the background noise (N, unwanted signal).

In receptor model analysis this can be interpreted as the relationship between concentrations (x) and uncertainties (s) (Paatero and Hopke, 2003):

$$\left(\frac{S}{N}\right)_j = \sqrt{\frac{\sum_{I=1}^{n} x_{ij}^2}{\sum_{i=1}^{n} s_{ij}^2}} \qquad (6.1)$$

In the EPA PMF v3, the equation is even stricter and considers only the portion of the concentration that exceeds the uncertainty (EPA-PMF 3.0 User Guide):

$$\left(\frac{S}{N}\right)_j = \sqrt{\frac{\sum_{i=1}^{n} \left(x_{ij} - s_{ij}\right)^2}{\sum_{i=1}^{n} s_{ij}^2}} \qquad (6.2)$$

During the European RM intercomparison (Karagulian et al., 2012), it was discovered that both of the above-mentioned equations for S/N fail totally if a species contains strongly downweighted values or if different matrix rows contain different scaling factors, e.g. some in mg and others in µg. EPA PMF v5 contains an improved expression that should work well for all kinds of data, even those where different rows have different scaling factors or when values were downweighted (e.g. missing values). To determine the S/N, two calculations are performed where concentrations below uncertainty are determined to have no signal, and for concentrations above uncertainty, the difference between concentration (x_i) and uncertainty (s_i) is used as the signal (EPA-PMF 5.0 User Guide):

$$d_{ij} = \left(\frac{x_{ij}-s_{ij}}{s_{ij}}\right) \text{ if } x_{ij} > s_{ij}$$

$$d_{ij} = 0 \text{ if } x_{ij} \leq s_{ij} \qquad (6.3)$$

The S/N is then calculated using:

$$\left(\frac{S}{N}\right)_j = {}^{1}\!/_{n} \sum_{i=1}^{n} d_{ij} \qquad (6.4)$$

In this case species with concentrations always below their uncertainty have a S/N of 0. Species with concentrations twice their uncertainty have a S/N value of 1. When the S/N is greater than 1 this may indicate a species with a "good" signal, tough this still depends on the determination of the uncertainties. Using this approach will eliminate the contribution of negative concentration values to the S/N, as well as limiting the S/N for species with a handful of high concentration events. When using this new expression, the numerical limits for weak/bad/good discrimination must be changed from the customary values shown below.

The signal-to-noise ratio is useful for classifying variables according to the information they supply for the source identification analysis. According to Paatero and Hopke (2003), variables with signal-to-noise ratios below 0.2 (bad) are to be excluded from the analysis, while variables where the ratio falls between 0.2 and 2.0 (weak) are suitable for the analysis. However, it is recommended that such variables be downweighted by a factor of 1/2 or 1/3 by increasing the uncertainties outside the EPA-PMF programs.

Mass closure and ion balance

Preliminary tests exist to match the masses or the electric charges of species. In the first case, mass closure is accomplished by comparing the mass of particulate matter (PM) to the sum of the masses of the major chemical components. For this calculation, organic carbon (OC) is to be converted into organic matter (OM) using an empirical coefficient that normally ranges from 1.4 to 2.1 (e.g. Turpin and Lim, 2001). The mass of crustal fraction must also be estimated from elements, as these are frequently present as oxides or carbonates. Therefore, the mass of the missing oxygen and carbon atoms should be added. The following empirical equations have been proposed to estimate these kinds of materials, by accounting for unmeasured oxides in minerals (Watson et al., 2002; Malm & Hand, 2007):

$$\text{Geological} = 1.89\text{Al} + 2.14\text{Si} + 1.4\text{Ca} + 1.43\text{Fe} \qquad (6.5)$$

$$\text{Soil} = 2.2\text{Al} + 2.49\text{Si} + 1.94\text{Ti} + 1.63\text{Ca} + 2.42\text{Fe} \quad (6.6)$$

As Fe can be emitted by multiple sources, Marcazzan et al. (2001) proposed an equation which considers only the natural contribution of Fe in the soil making use of enrichment factors.

Commonly, the mass of PM, determined with the gravimetric method, is higher than the sum of the chemical components. This can be explained in different ways: a) not all the relevant chemical components have been determined; b) the mass measurement includes water adsorbed to particles that is not quantified in the chemical analyses; c) the selected coefficient for converting OC to OM is not optimal for the study area; d) the elements that have been assumed to be present as oxides and carbonate have not been taken into consideration.

By comparing the sum of anion equivalents with the sum of cation equivalents, it is possible to assess departure from neutrality, and plotting values in a graph helps to identify samples with an atypical ionic composition. The most common ionic species in PM are inorganic cations (ammonium, sodium, potassium, calcium, and magnesium) and inorganic anions (sulphate, nitrate, chloride, and carbonate). Among the organic acids, the most relevant anions are those deriving from oxalic, malonic, succinic, formic, and acetic acids (Chebbi & Carlier, 1996).

It is also possible to develop simplified mass closure models which provide an excellent check on the consistency of data from individual samples. An example is the Pragmatic Mass Closure Model (Harrison et al., 2003) which uses simple empirical parameterisations to account for the measured mass of particles in terms of a small number of analytical variables. Although such a model might be expected to be site-specific, it has proved to be transferable between sites, years and particle size fractions (Yin and Harrison, 2008). However, caution should be exercised in the application of the model to sites with entirely different pollution traits. Simple empirical corrections should be feasible in such cases.

Analysis of consistency in time and space

In order to populate a dataset with an appropriate number of samples it may be necessary to collect data for more than one year. However, species and other variables collected during different years may show different relationships. In order to check these patterns before running the analysis, scatter plots to look for edges (Henry, 2003) or time trend plots are useful. Changes in sampling methodologies or analytical techniques may create disruption in time series that must be duly taken into account during data elaboration. Comparing time series from different sites is helpful to detect anomalous patterns. Nevertheless, it must be considered that different monitoring networks may have different instrumentation (e.g. different inlets, different operation principles) or different data treatment protocols.

If several receptor sites have been operated near each other, e.g. within one city area, then it may be useful to soft-constrain regional factors (more details on constrained models in section 12), such as secondary sulphate, in order to have similar factors at all sites. In this way, a significant part of rotational uncertainty may be avoided. It should be kept in mind that the secondary sulphate G factor often has no rotation-limiting zero values and hence is prone to rotations if no constraints are applied.

References

Baccarelli, A., Pfeiffer, R., Consonni, D., Pesatori, A.C., Bonzini, M., Patterson Jr, D.G. 2005. Handling of dioxin measurement data in the presence of non-detectable values: overview of available methods and their application in the Seveso chloracne study. *Chemosphere*, 60, 898-906.

Brown S.G., Hafner H., 2005. Multivariate Receptor Modeling Workbook. Sonoma Technology Inc. Petaluma, CA. U.S.A. 236 pp. http://www.sonomatechdata.com/sti_workbooks/

Chebbi, A., Carlier, P., 1996. Carboxylic acids in the troposophere, occurrence, sources, and sinks: a review. *Atmospheric Environment* 30, 4233–4249.

Harrison, R.M., Jones, A.M. Lawrence, R.G., 2003. A Pragmatic Mass Closure Model for Airborne Particulate Matter at Urban Background and Roadside Sites. *Atmospheric Environment* 37, 4927-4933.

Henry RC., 2003. Multivariate receptor modeling by N-dimensional edge detection. *Chemometrics and Intelligent Laboratory Systems* 65, 179-89.

Hopke, P.K., Liu, C., Rubin, D.B. 2001. Multiple imputation for multivariate data with missing and below-threshold measurements: time-series concentrations of pollutants in the Arctic. *Biometrics*, 57, 22-33.

Ibrahim, J.G., Chen, M.H., Lipsitz, S.R., Herring, A.H. 2005. Missing-data methods for generalized linear models: a comparative review. *J. Am. Stat. Assoc.*, 100 (469), 332-346.

IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book"). Compiled by A. D. McNaught and A. Wilkinson. Blackwell Scientific Publications, Oxford ,1997. XML on-line corrected version: http://goldbook.iupac.org, 2006 created by M. Nic, J. Jirat, B. Kosata; updates compiled by A. Jenkins. ISBN 0-9678550-9-8. doi:10.1351/goldbook.

Karagulian F., Belis C. A., Borowiak A., 2012. Results of the European Intercomparison exercise for receptor models 2011-2012.Part I. JRC Report EUR 25727 EN.

Malm W., Hand J.L., 2007. An examination of the physical and optical properties of the aerosols collected in the IMPROVE program. *Atmospheric Environment* 41, 3407–3427.

Marcazzan G.M., Vaccaro S., Valli G., Vecchi R., 2001. Characterisation of PM10 and PM2.5 particulate matter in the ambient air of Milan (Italy), Atmospheric Environment, 35, 4639-4650

Mooibroek D., Staelens, J., Cordell, R., Panteliadis, P., Delaunay, T., Weijers, E., Vercauteren, J., Hoogerbrugge, R., Dijkema, M., Monk, P.S., Roekens, E. 2016. PM10 Source Apportionment in Five North Western European Cities – Outcome of the Joaquin Project. *Airborne Particulate Matter: Sources, Atmospheric Processes and Health,* ISBN: 978-1-78262-491-2

Norris G., Vedantham R., K. Wade, S. Brown, J. Prouty, C. Foley, 2008. EPA Positive Matrix Factorization (PMF 3.0) Fundamentals & User Guide. EPA 600/R-08/108.

Norris G, Vedantham R., Duvall R., Henry R. C., 2007. EPA Unmix 6.0 Fundamentals & User Guide. EPA/600/R-07/089.

Paatero P., Hopke P.K., 2003. Discarding or downweighting high-noise variables in factor analytic models. *Analytica Chimica Acta* 490 277–289.

Polissar, A. V., Hopke P. K., Paatero P., Malm W. C., and Sisler J. F., 1998. Atmospheric aerosol over Alaska 2. Elemental composition and sources, *Journal of Geophysical Research D: Atmospheres*, 103(D15), 19045-19057.

Rubin, D.B. 1976, Inference and missing data. *Biometrika*, 63, 581-592

Rubin, D.B. 1977, Formalizing subjective notions about the effect of nonrespondents in sample surveys. *J. Am. Stat. Assoc.*, 72, 538-543

Turpin, B.J., Lim, H-J., 2001. Species Contributions to $PM_{2.5}$ Mass Concentrations: Revisiting Common Assumptions for Estimating Organic Mass. *Aerosol Science and Technology* 35, 602–610.

Watson J. G., Zhu T., Chow J. C., Engelbrecht J., Fujita E. M., Wilson W. E., 2002. Receptor modeling application framework for particle source apportionment. *Chemosphere* 49, 1093–1136.

Yang, X., Belin, T.R., Boscardin, W.J. 2005. Imputation and variable selection in linear regression models with missing covariates. *Biometrics*, 61 (2), 498-506.

Yin, J., Harrison, R.M., 2008. Pragmatic Mass Closure Study for PM_{10}, $PM_{2.5}$ and PM_{10} at Roadside, Urban Background and Rural Sites. *Atmospheric Environment*. 42, 980-988.

7. INPUT DATA UNCERTAINTY CALCULATION

Uncertainty is the quantitative estimation of the quality of a measurement that makes it possible to compare results among themselves and with reference values (Joint Committee for Guides in Metrology (JCGM) 100:2008). Estimating the uncertainty of measurements is a common practice in analytical chemistry and physics that is performed routinely according to international criteria laid down in standards and implemented in reference methods. In analytical chemistry, uncertainty is evaluated both as the standard deviation of repeated observations and by comparison with reference materials.

In source apportionment, analytical uncertainty is important since the most commonly used models, like PMF and CMB, require the uncertainty of the species concentrations as input data in order to find the solution and the uncertainty of the output.

In PMF analysis, uncertainty estimation is particularly critical because every entry is weighted according to its uncertainty. Although analytical uncertainty estimation is an important step of receptor modelling, it must be noted that it is only one component of the overall input data uncertainty required by receptor models (Polissar et al., 1998). Other contributions to the overall uncertainty include flow rate uncertainty, between-sampler uncertainty and other unidentified noise.

Moreover, not all components of overall uncertainty behave equally. In PMF input, only the components of uncertainty that are capable of generating residuals, i.e. components that will increase the Q value of the fit (see section 9), should be included. Flow rate uncertainty is a prime example: flow rate variations influence all values on a matrix row by the same multiplier, hence causing no increase of residuals e_{ij} (equation I.1). Flow rate uncertainty, and other similar uncertainties, should bypass the PMF stage and be attached directly to the computed G factor elements (equation I.1). In addition to analytic uncertainty, modelling errors (e.g. variation of source profiles with time, chemical transformations during transport from source to receptor) also cause residuals in PMF modelling. Expected contributions from modelling errors must also be accounted for in the PMF input data uncertainties. There is no fixed rule for such contributions. To begin with, it is reasonable to include 10% of each data value as a provision for modelling errors. When experience is accumulated, this numerical coefficient may be adjusted. However, this additional uncertainty must always be reported in publications so that the work is reproducible.

The operator also needs to attribute an uncertainty to missing values and to values below the detection limit. That uncertainty is normally higher than that of measured values. Polissar et al. (1998) set the uncertainty of values below the detection limit to 5/6 of the detection limit, while the uncertainty of missing values is by convention set at four times the geometric mean. This convention has no general statistical basis. For some datasets, significantly larger uncertainty values are needed for missing values.

Sometimes the attribution of uncertainties may be achieved by a trial-and-error process that aims to obtain the best model fit which is evaluated using Q values (see section 9), scatterplots, distribution of residuals and results from multiple regressions (e.g. Polissar et al., 2001).

When dealing with databases in which single entry uncertainties are unavailable or are inappropriate for modelling purposes, the global input data uncertainties may be estimated using equation-based approaches, which rely on the species detection limit (DL), empirical constants (k), species concentration (C) and/or the coefficient of variation (CV) (Karagulian & Belis, 2012).

Analytical uncertainty can be estimated by the linear regression described in equation 7.1 where σ_a is the uncertainty of the analytical procedure, m is the mass of the analyte, and σ_0 and α are fitting parameters (Anttila et al., 1995):

$$\sigma_a^2 = \sigma_0^2 + (\alpha m)^2 \qquad (7.1)$$

In the estimated fractional uncertainties (EFU) method the error structures (s_{ij}) are (Kim and Hopke, 2005):

$$S_{ij} = [DL]/3 + kx \qquad (7.\ 2)$$

When no empirical constants are used other than the DL and coefficient of variation (CV), the analytical uncertainty is (Chow et al., 2007):

$$\sigma_{i,t}^a = \sqrt{DL_i^2 + (CV_i \times C_{i,t})^2} \qquad (7.3)$$

Sampling contributes to the uncertainty of measured values due to sampling volume uncertainty, selective effect and other artefacts caused by the sampler inlet, and losses due to sample transport and conservation. These contributions can be assessed with field tests (e.g. collocated measurements and comparison with reference instrumentation and techniques). In the case of destructive analysis of the filters where PM is collected, the procedure of subtracting blank filter (different from sampled ones) concentrations is an additional source of uncertainty. Sampling and blank subtraction uncertainties have been incorporated into the input data uncertainty by Amato et al. (2009):

$$\sigma_A^2 = \sigma_a^2 + \sigma_{BLK}^2 \qquad (7.4)$$

$$\sigma_{ij}^2 = \sqrt{\frac{\sigma_A^2}{V_i^2} + (\beta x_{ij})^2} \qquad (7.5)$$

where the standard deviation of species concentrations in blank filters (σ_{BLK}), the sampled volume (*V_i*) and a coefficient (β) are used to account for the additional uncertainty sources.

Input data uncertainties can also be estimated with the PMF2 software. This is a more complex procedure that uses three codes, C1, C2 and C3, the error model and the arrays T, U and V (Paatero, 2004).

CMB uses source profiles as input data with associated uncertainty estimation. When source profiles are too similar, CMB may be not able to find a solution (collinearity). In order to prevent problems related to collinearity, sources with similar chemical composition are either combined into source categories / composite profiles or only one profile is incorporated in the analysis while the other is dropped. The uncertainty of the composite is obtained by propagation of the uncertainty of the pooled single profiles (Watson, 2004). However, this may not fully account for the variety of similar sources in the study area and their variability over time.

In order to deal with the variability of source profiles, initial model runs often contain many profiles, and a sensibility test should be carried out to assess their influence on the precision and stability of the source contribution estimates.

A default value of zero with a standard deviation equal to the analytical detection limit may be assigned to a species of a source profile if that species is known to be absent from that source (Watson, 2004).

Metals are excellent receptor species given the assumption that such receptor species do not chemically react or physically repartition during transport from source to receptor. As such, metals have been used from the very beginning of receptor modelling activities (e.g. Hopke et al., 1991). However, in the search for specific receptor species for different combustion sources (also called molecular markers), the use of organic chemical compounds has grown popular in modern source apportionment studies (e.g. Srivastava et al., 2018a & 2018b), even though this class of compounds often comes into conflict with the above-mentioned assumption. An interesting utilisation of uncertainty data for the inclusion of semi-volatile and photo-chemically reactive species in CMB and PMF has been developed and adopted by Latella et al. (2005), Junninen et al. (2009), Belis et al. (2011) and Larsen et al. (2012). In these studies, methods are described for using information on volatility to account for the re-partitioning processes (PAHs; semi-volatile organic carbon fraction) and photochemical degradation (hydrocarbons, levoglucosan) from source to receptor. This information has served as error-input to CMB and PMF for the error weighting in the statistical procedures.

PMF input uncertainty for advanced users (by P. Paatero)

The definition of correct input uncertainties for PMF is not only based on measurement errors. Residuals are also caused by experimental problems (e.g. sample contamination) and by modelling problems, such as:

- source profiles being variable, in contrast to PMF   assumption of constant profiles.

- sporadic sources, too weak to get their own factor but strong enough to cause "too large" residuals

- incompatible or non-conforming variables that have nothing in common with other variables.

The basic idea of correctly weighted least squares (LS) fit is that assumed uncertainties s_ij should be specified so that expected scaled residuals will be, on the average, between -2 and +2 and, therefore, observed Q values approximate the theoretically expected Q values. This behaviour should be valid for each single species in the dataset. Uncertainties s_ij should be specified so that scaled residuals, for each variable, are of correct magnitudes even in the presence of those ill-known modelling problems.

A correct balance of different s_ij values is very important. The s_ij of different variables j must be in harmony. Equally important is that s_ij of large data values x_ij are in harmony with s_ij of small values x_ij for each variable j.

The uncertainty coefficients C1 and C3 were introduced in PMF2 in order to help users in adjusting the balance between small and large data values. For this purpose, C1 is specified small, corresponding to the uncertainty of small values. Then C3 should contribute the main part of uncertainty for large values.

Practical experience with many real data sets has shown that quite often, coefficients C1 and C3 lead to satisfactory uncertainties for both small, medium, and large data values. It has also been realized that in many practical cases, different values of C3 must be specified for different variables.

It is nowadays common that laboratories specify individual uncertainties for each measured value. These uncertainties do not take into account the other reasons for increased residuals. For this reason, it has been necessary to introduce in EPA PMF v5 the "extra modelling uncertainty" coefficient. Such coefficient does not, however, compensate for too large scaled residuals in small x_ij. If scaled residuals for small x_ij are too large, causing the overall Q to be too large, the situation cannot be corrected by only increasing the coefficient value(s).

The extra modelling uncertainty is to be used for model set up. For any final analyses, the user should not use the extra modelling uncertainty. Instead, the input file uncertainty calculation should be revised and documented.

The following procedure is proposed to check whether the input uncertainties are suitable for PMF analysis.

For each variable j, plot the scaled residuals r_ij vs. the fitted values y_ij (NOT vs. the measured values x_ij). Each figure should be similar, showing a band of randomly oscillating values, typically between -2 and +2. The band should be of same width for small, medium, and large values of variable j. The band should be approximately symmetrical with respect to zero.

If the figure(s) does not fulfil these criteria, then the users should correct their model.

For substituted values (e.g. substituted BDL values) the diagram may not respect the abovementioned criteria if the original BDL values are not used in the plot (sometimes the original are not available).

If the band of residuals is not right only for small values y_ij, then C1 values should be changed (if using PMF2). Similarly, for large values y_ij, adjust C3 (if using PMF2).

If the band of residuals is not symmetrical, then the problem may be that few uncertainties are much too small. This happens if the lab has specified proportional uncertainties for all values, so that a near-zero value gets an unrealistic small uncertainty that distorts the model. The EPA PMF v5 does not cope with this problem case, not even in robust mode. Users must correct such too small uncertainties manually.

All adjustments of uncertainties must be based on residuals of a number of values and must apply collectively to a number of values, e.g. to all small values of variable j.

There are a few exceptions to the procedure that was outlined above:

1. Variables fitted by unique factors. This case is discussed in some detail in the PMF handbook (Paatero, 2004). Even though the scaled residuals of such variables are much too small their uncertainties must not be reduced.

2. In some cases, it is not desirable to perform a correctly weighted LS fit with PMF. Downweighting weak variable(s) is the most usual case. Scaled residuals for a downweighted variable are by definition very small.

3. Suspect or non-conforming variables are sometimes downweighted by users. This is similar to the case of weak variables.

References

Amato, F., Pandolfi M., Escrig A., Querol X., Alastuey A., Pey J., Perez N., Hopke P. K., 2009. Quantifying road dust re-suspension in urban environment by Multilinear Engine: A comparison with PMF2, *Atmospheric Environment* 43(17), 2770-2780.

Anttila, P., Paatero P., Tappert U., Jarvinen O., 1995. Source identification of bulk wet deposition in Finland by positive matrix factorization, *Atmospheric Environment* 29(14), 1705-1718.

Belis, C. A., Cancelinha, J., Duane, M., Forcina, V., Pedroni, V., Passarella, R., Tanet, G., Douglas, K., Piazzalunga, A., Bolzacchini, E., Sangiorgi, G., Perrone, M. G., Ferrero, L., Fermo, P., and Larsen, B. R., 2011. Sources for PM air pollution in the Po Plain, Italy: I. Critical comparison of methods for estimating biomass burning contributions to benzo(a)pyrene. *Atmospheric Environment* 45, 7266-7275.

Chow, J. C., Watson J. G., Chen L. W. A., Chang M. C. O., Robinson N. F., Trimble D., Kohl S., 2007. The IMPROVE_A temperature protocol for thermal/optical carbon analysis: Maintaining consistency with a long-term database. *Journal of the Air and Waste Management Association* 57(9), 1014-1023.

Hopke P.K. (ed.), 1991. Receptor Modelling for air quality management. Data handling in sicnece and technology- volume 7. Elsevier. ISBN 0-444-88218-9

Joint Committee for Guides in Metrology, 2008. Evaluation of measurement data — Guide to the expression of uncertainty in measurement. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML.

Karagulian F., Belis C.A, 2012. Enhancing Source Apportionment with receptor models to Foster the Air Quality Directive Implementation. *International Journal of Environmental Pollution*. 50: 190-199.

Larsen, B.R., Gilardoni S., Stenström K., Niedzialek J., Jimenez J., Belis C. A., 2012. Sources for PM air pollution in the Po Plain, Italy: II. Probabilistic uncertainty characterization and sensitivity analysis of secondary and primary sources. *Atmospheric Environment* 50, 203-213.

Latella, A, Stani G., Cobelli L., Duane M., Junninen H., Astorga C., Larsen B.R., 2005. Semicontinuous GC analysis and receptor modelling for source apportionment of ozone precursor hydrocarbons in Bresso, Milan, 2003. *Journal of Chromatography A* 1071, 29–39.

Kim, Y. J., Hopke P. K., 2005. Estimation of Organic Carbon Blank Values and Error Structures of the Speciation Trends Network Data for Source Apportionment. *Journal of the Air and Waste Management Association* 55, 1190-1199.

Paatero P., 2004. User's guide for positive matrix factorization programs PMF2 and PMF3, Parts 1 and 2. University of Helsinki, Helsinki, Finland.

Polissar, A. V., Hopke P. K., Paatero P., Malm W. C., Sisler J. F., 1998. Atmospheric aerosol over Alaska 2. Elemental composition and sources. *Journal of Geophysical Research D: Atmospheres* 103(D15), 19045-19057.

Polissar A.V., Hopke P.K. Poirot R.L., 2001. Atmospheric aerosol over Vermont: chemical composition and sources. *Environmental Science and Technology.* 35, 4604-4621.

Srivastava D., Tomaz S., Favez O., Lanzafame G.M., Golly B., Besombes J.-L., Alleman L.Y., Jaffrezo J.-L., Jacob E., Perraudin E., Villenave E., Albinet A., 2018a. Speciation of organic fraction does matter for source apportionment. Part 1: a one-year campaign in Grenoble (France). *Science of the Total Environment*, 624, 1598-1611.

Srivastava D., Favez O., Bonnaire N., Lucarelli F., Haeffelin M., Perraudin E., Gros V., Villenave E., Albinet A., 2018b. Speciation of organic fraction does matter for source apportionment. Part 2: intensive short-term campaign in the Paris area (France). *Science of the Total Environment*, 634, 267-278.

Watson, J. G., 2004. Protocol for Applying and Validating the CMB model to PM2.5 and VOCs, US EPA, USA. Report EPA-451/R-04-001. 173 pp.

8. CHEMICAL MASS BALANCE MODELS

Chemical Mass Balance (CMB) is based on the mass conservation of individual chemical species: major aerosol components (such as elemental and organic carbon, sulphate, nitrate, ammonium, etc.), other inorganic and elements, as well as metallic or organic compounds which can be considered as source markers or tracers. In the mass conservation equations, deriving from the general equation I.1 (here the original notation was kept to facilitate consultation of the references and our notation is reported between parentheses), known concentrations C_{ik} (x_{ij}) of specific species at a receptor site *k* are written as the product of unknown source contributions s_{jk} (g_{ik}) and known source profiles a_{ij} (f_{kj}) (Cooper et al., 1984; Watson et al., 1998). a_{ij} (f_{kj}) are the fractional abundances of the species in the source emissions, commonly expressed by the ratios between the species and the $PM_{2.5}$ or organic carbon mass. The mass conservation equations for each species emitted from *m (p)* sources can be written as follows:

$$C_{ik} = \sum_{j=1}^{m} a_{ij} s_{jk} \tag{8.1}$$

In practice, the set of linear equations generated by equation 8.1 is solved with an effective variance-weighted least square method using the EPA-CMB8.2 software (USEPA, 2017). Note that although equation 8.1 is similar to equation A.1 (in this case a_{ij} (f_{kj}) are known values), the model is conceived for one sample per site and has no residual term.

Friedlander (1973) proposed a modified version of equation 8.1 that included a coefficient, α_{ij}, that accounted for changes in the profile values for specific species in transit. However, the current practice is to apportion the primary material that has not changed between source and receptor, so this coefficient is set to 1. The remaining quantities of reactive species such as ammonium, nitrate, sulphate, and organic carbon are then indirectly apportioned to secondary sources. Accordingly, the species used as fitting species are strictly of primary origin. They must be (i) stable during atmospheric transport (i.e. low volatility and moderately reactive), (ii) accurately determined at the receptor site and (iii) reported for all source profiles considered in the model. The number of fitting species has to exceed the number *m* of emission sources. The first attempts to solve the mass balance equation were based on tracer compounds (ideally one for each source, e.g. Miller, 1972). Since inorganic compounds rarely derive from a single source, this approach gave way to another that considers a higher number of species than sources. This latter approach was fully developed in the Chemical Mass Balance Model as described by Watson et al. (1997), among others. More recently, the identification of organic compounds that can be used as tracers for specific sources or types of sources (e.g. Schauer, 1999a and b) led to a combination of both approaches, i.e. containing more species than sources but including some organic species (tracer or markers) deriving from unique sources (e.g. Chow, 2007; Subramanian, 2006).

The main strength of the CMB model is that, unlike other statistical receptor models (e.g. PMF), it does not require a large dataset and, theoretically, equation 8.1 can be solved for an individual sample (see section 3). Moreover, unlike factor analysis techniques, the CMB output does not require additional identification of the contributing sources/factors, as the profiles are selected *a priori* for well-defined sources.

However, the most important issue generally encountered in CMB modelling is the selection of the source profiles that best represent the aerosol collected at the receptor site. This selection relies heavily on two implicit assumptions:

(i) The aggregate emissions from a given source class are well represented by an average source profile with well-known a_{ij} ratios.

(ii) All the major primary sources of the species are included in the model.

With most commonly measured species for particles (e.g. ions, elements, carbon and organic compounds) and common source types, approximately four to eight primary source classes are linearly independent and can thus be apportioned by the CMB. These conventionally comprise traffic emissions which are often separated between diesel and gasoline combustion engines, biomass burning, vegetative detritus, cooking emissions and dust (e.g. Zheng et al., 2006a and b; Sheesley et al., 2007; Docherty et al., 2008; Stone et al., 2008; Favez et al., 2010). Additional profiles can also be selected to specifically represent the area of study, including coal burning (Rutter et al., 2009), metal smelting (El Haddad et al., 2011), metallurgical coke production (Subramanian et al., 2007; El Haddad et al., 2011) and shipping/heavy fuel oil combustion (Minguillon et al., 2008; El Haddad et al., 2011).

Currently in the literature, there are a great number of profiles and composite profiles for the major primary sources (e.g. more than 50 profiles for traffic emissions and more than 40 profiles for biomass smoke). A comparison of these profiles reveals significant variations in emissions depending on the fuel type and combustion conditions, rendering the choice between these profiles very complex. Subramanian et al. (2007) show that library profiles may not always reflect the properties of a specific source in a given study area.

To achieve CMB analysis and validation, a number of steps must be followed.

First, for each source, several profiles and composite profiles have to be selected based on the specificity of the study area (e.g. harbour, industries, wood or coal burning, predominance of diesel cars, etc.) and the species concentrations at the receptor site. Examining diagnostic ratios between species can help to eliminate outlier profiles (Robinson et al., 2006a, b and c; El Haddad et al., 2011). Constructing composite profiles from available data (Sheesley et al., 2007; Favez et al., 2010) or developing new source profiles through real world measurements, for instance tunnel experiments (e.g. Phuleria et al., 2006; El Haddad et al., 2009), twin site studies (Pant et al., 2014) and open fires (e.g. Lee et al., 2005), are also common practices carried out to better represent the emissions in the study area. Comparison of contribution estimates for sources such as traffic, cooking, biomass burning, coal combustion and secondary organic aerosol can be made with simultaneously collected online and offline Aerosol Mass Spectrometer data (Huang et al., 2014; Yin et al., 2015).

Second, the model is run repeatedly, including different combinations of the selected profiles. Based on the quality of the CMB solutions, the best combinations can be selected. The sensitivity of the results to the choice of the profiles and the related uncertainties can be assessed. As a quality control check of the CMB calculation, statistical performance measures include the use of R-square (target 0.8–1.0), chi-square (target 0–4.0) and the species' calculated-to-measured ratios (target $0.5<C/M<2$), as indicators of the goodness of fit (Watson et al., 1998). If the CMB solutions do not meet these criteria, it would mean that one of the two aforementioned assumptions is transgressed (i.e. non-representative or missing profiles).

The CMB also provides the uncertainties of the source contribution estimates by propagating the uncertainty estimates of the receptor data and source profiles (entered as input by the operator) through the effective-variance least squares calculations. Their magnitudes are a function of the uncertainties in the input data and of the amount of collinearity (i.e. degree of similarity) among source profiles. Two or three times the standard error may be taken as an upper limit of the source contribution.

Third, CMB is often applied to the carbonaceous component of PM and, if the results are combined with those of other analytes using a simple mass closure approach, this can be a valuable check on data quality (e.g. Yin et al., 2010). A further useful check is whether the concentration of organic carbon unaccounted for in a CMB model and assumed to be secondary in origin can be compared with independent estimates of secondary organic carbon derived using the elemental carbon tracer method as reported by Yin et al. (2010).

When available, soluble potassium, water-soluble organic carbon, radiocarbon and/or Aethalometer measurements can also help corroborate the CMB outputs, especially in the case of high contributions from secondary organic aerosols and biomass burning organic aerosols (e.g. Docherty et al., 2008; Favez et al., 2010).

References

Cooper, J.A., Watson, J.G., Huntzicker, J.J., 1984. The Effective Variance Weighting for Least Squares Calculations Applied to the Mass Balance Receptor Model, *Atmospheric Environment* 18, 1347-1355.

Chow, J. C., Watson J. G., Lowenthal D. H., Chen L. W. A., Zielinska B., Mazzoleni L. R., Magliano K. L., 2007. Evaluation of Organic Markers for Chemical Mass Balance Source Apportionment at the Fresno Supersite. *Atmospheric Chemistry and Physics* 7 (7), 1741-1754.

Docherty, K. S., Stone, E. A., Ulbrich, I. M., DeCarlo, P. F., Snyder, D. C., Schauer, J. J., Peltier, R. E., Weber, R. J., Murphy, S. M., Seinfeld, J. H., Grover, B. D., Eatough, D. J., Jimenez, J. L., 2008. Apportionment of Primary and Secondary Organic Aerosols in Southern California during the 2005 Study of Organic Aerosols in Riverside (SOAR-1), *Environmental Science & Technology* 20, 7655-7662.

El Haddad, I., Marchand, N., Dron, J., Temime-Roussel, B., Quivet, E., Wortham, H., Jaffrezo, J. L., Baduel, C., Voisin, D., Besombes, J. L., Gille, G., 2009. Comprehensive primary particulate organic characterization of vehicular exhaust emissions in France, *Atmospheric Environment* 39, 6190-6198.

El Haddad, I., Marchand, N., Wortham, H., Piot, C., Besombes, J.-L., Cosic, J., Chauvel, C., Armengaud, A., Robin, D., Jaffrezo, J.-L., 2011. Primary sources of PM2.5 organic aerosol in an industrial Mediterranean city, Marseille, *Atmospheric Chemistry and Physics* 11, 2039–2058, doi:10.5194/acp-11-2039-2011.

Favez, O., Haddad, I. E., 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). *Atmospheric Chemistry and Physics discussion*, 559-613.

Friedlander, S.K., 1973. Chemical Element Balances and Identification of Air Pollution Sources. *Environmental Science and Technology* 7, 235-240.

Huang, R.-J. , Zhang, Y.L., Bozzetti, C., Ho, K.-F., Cao, J., Han, Y., Dällenbach, K., Slowik, J. G., Platt, S. M., Canonaco, F., Zotter, P., Wolf, R., Pieber, S.M., Bruns, E.A., Crippa, M., Ciarelli, G., Piazzalunga, A., Schwikowski, M., Abbaszade, G., Schnelle-Kreis, J., Zimmermann, R., An, Z., Szidat, S., Baltensperger, U., El Haddad, I., Prévôt, A.S.H., 2014. High contribution of secondary aerosol to particle pollution during haze events in China. Nature, 524, 218-222.

Lee, S., Baumann, K., Schauer, J. J., Sheesley, R. J., Naeher, L. P., Meinardi, S., Blake, D. R., Edgerton, E. S., Russell, A. G., Clements, M., 2005. Gaseous and Particulate Emissions from Prescribed Burning in Georgia, *Environmental Science & Technology* 23, 9049-9056.

Miller, M.S., Friedlander S.K., Hiddy G.M. 1972. A chemical element balance for the Pasadena aerosol. *Journal of Colloidal Interface Science* 39, 65-176.

Minguillon, M. C., Arhami, M., Schauer, J. J., Sioutas, C., 2008. Seasonal and spatial variations of sources of fine and quasi-ultrafine particulate matter in neighbourhoods near the Los Angeles-Long Beach harbor, *Atmospheric Environment* 42, 7317–7328.

Pant, P., Yin, J., Harrison, R. M., 2014, Sensitivity of a mass balance model to different molecular marker traffic source profiles. Atmospheric Environment 82, 238-249.

Phuleria, H. C., Geller, M. D., Fine, P. M., Sioutas, C., 2006. Size-resolved emissions of organic tracers from light-and heavy-duty vehicles measured in a California roadway tunnel. *Environmental Science & Technology* 13, 4109-4118.

Robinson, A. L., Subramanian, R., Donahue, N. M., Rogge, W. F., 2006a. Source Apportionment of Molecular Markers and Organic Aerosol1. Polycyclic Aromatic Hydrocarbons and Methodology for Data Visualization. *Environmental Science & Technology* 24, 7803-7810.

Robinson, A. L., Subramanian, R., Donahue, N. M., Bernardo-Bricker, A., Rogge, W. F., 2006b. Source Apportionment of Molecular Markers and Organic Aerosol. 2. Biomass Smoke. *Environmental Science & Technology* 24, 7811-7819,

Robinson, A. L., Subramanian, R., Donahue, N. M., Bernardo-Bricker, A., Rogge, W. F., 2006c. Source Apportionment of Molecular Markers and Organic Aerosol. 3. Food Cooking Emissions, *Environmental Science & Technology* 24, 7820-7827.

Rutter, A. P., Snyder, D. G., Schauer, J. J., Deminter, J., Shelton, B., 2009. Sensitivity and Bias of Molecular Marker-Based Aerosol Source Apportionment Models to Small Contributions of Coal Combustion Soot. *Environmental Science & Technology* 20, 7770-7777.

Schauer, J. J., Kleeman M. J., Cass G. R., and Simoneit B. R. T., 1999a. Measurement of Emissions from Air Pollution Sources. 1. C1 through C29 Organic Compounds from Meat Charbroiling." *Environmental Science and Technology* 33 (10), 1566-1577.

Schauer, J. J., M. J. Kleeman, G. R. Cass, and B. R. T. Simoneit. 1999b. Measurement of Emissions from Air Pollution Sources. 2. C 1 through C 30 Organic Compounds from Medium Duty Diesel Trucks. *Environmental Science and Technology* 33 (10), 1578-1587.

Sheesley, R. J., Schauer, J. J., Zheng, M., Wang, B., 2007. Sensitivity of molecular marker-based CMB models to biomass burning source profiles. *Atmospheric Environment* 39, 9050-9063.

Stone, E. A., Snyder, D. C., Sheesley, R. J., Sullivan, A. P., Weber, R. J., Schauer, J. J., 2008. Source apportionment of fine organic aerosol in Mexico City during the MILAGRO experiment 2006. *Atmospheric Chemistry and Physics* 5, 1249-1259.

Subramanian, R., Donahue, N. M., Bernardo-Bricker, A., Rogge, W. F., Robinson, A. L., 2006. Contribution of motor vehicle emissions to organic carbon and fine particle mass in Pittsburgh, Pennsylvania: Effects of varying source profiles and seasonal trends in ambient marker concentrations. *Atmospheric Environment* 40 8002–8019

Subramanian, R., Donahue, N. M., Bernardo-Bricker, A., Rogge, W. F., Robinson, A. L., 2007. Insights into the primary-secondary and regional-local contributions to organic aerosol and $PM_{2.5}$ mass in Pittsburgh, Pennsylvania. *Atmospheric Environment* 41, 7414-7433.

USEPA, 2017. https://www.epa.gov/scram/chemical-mass-balance-cmb-model.

Watson, J. G., Robinson, N. F., Fujita, E. M., Chow, J. C., Pace, T. G., Lewis, C., Coulter, T. 1998. CMB8 Protocol for Applying and Validatiing PM2.5 and VOCs, US EPA, USA.

Watson, J.G., Robinson, N.F., Lewis, C.W., Coulter, C.T., Chow, J.C., Fujita, E.M., Lowenthal, D.H., Conner, T.L., Henry, R.C., Willis, R.D., 1997. Chemical Mass Balance Receptor Model Version 8 (CMB) User's Manual. Reno. NV, US.

Yin, J., Harrison, R.M., Chen, Q., Rutter, A. and Schauer, J.J., 2010. Source Apportionment of Fine Particles at Urban Background and Rural Sites in the UK Atmosphere, Atmospheric Environment 44, 841-851.

Yin, J., Cumberland, S.A, Harrison, R.M., Allan, J., Young, D., Williams, D., Coe, H., 2015. Receptor modelling of fine particles in southern England using CMB including comparison with AMS-PMF factors. Atmospheric Chemistry and Physics, 2139-2158.

Zheng, M., Hagler, G. S. W., Ke, L., Bergin, M. H., Wang, F., Louie, P. K. K., Salmon, L., Sin, D. W. M., Yu, J. Z., Schauer, J. J., 2006a. Composition and sources of carbonaceous aerosols at three contrasting sites in Hong Kong. *Journal of Geophysical Research-Atmospheres*, D20.

Zheng, M., Ke, L., Edgerton, E. S., Schauer, J. J., Dong, M. Y., Russell, A. G., 2006b. Spatial distribution of carbonaceous aerosol in the southeastern United States using molecular markers and carbon isotope data. *Journal of Geophysical Research-Atmospheres*, D10.

9. FACTOR ANALYSIS I: SELECTION OF THE NUMBER OF FACTORS AND DEALING WITH ROTATIONAL AMBIGUITY (PMF)

The goal of Positive Matrix Factorization (PMF) - like any other multivariate receptor model (RM) - is to identify a number of factors p, the species profile f of each source, and the amount of mass g contributed by each factor to each individual sample (equation I.1).

PMF is an advanced factor analysis technique based on the work of Paatero and Tapper (1994); it uses realistic error estimates to weigh data values and imposes non-negativity constraints in the factor computational process. Briefly, it is a weighted least square fit, with weights based on the known standard uncertainties of the element concentrations in the data matrix. The factor model PMF can be written as:

$$\boldsymbol{X} = \boldsymbol{G} \cdot \boldsymbol{F} + \boldsymbol{E} \qquad (9.1)$$

where X is the known n by m matrix of the m-measured chemical species in n samples. G is an n by p matrix of source contributions to the samples (time variations of factors scores). F is a p by m matrix of source compositions (source profiles). G and F are factor matrices to be determined, and E is defined as a residual matrix i.e. the difference between the measurement X and the model Y = G · F as a function of G and F.

Two common programs solve the PMF problem described above: PMF2 (Paatero, 2010) and the multilinear engine (ME) platform (Paatero, 1999).

It is well known that factor analysis can give a number of possible solutions, all mathematically correct. The choice of the best solution in PMF analysis, e.g. the **number of factors** that best represent the real case under study, shall be supported by quantitative indicators (Hopke, 2000; Reff et al., 2007).

Examining the Q-value

The Q-value is a goodness of fit parameter, the evaluation of which may give useful indications when the data-point uncertainties are well determined.

The theoretical Q-value is approximately equal to the number of degrees of freedom or to the total number of good data points in the input data array minus the total number of fitted factor elements. If the errors are properly estimated, it can be seen that fitting each good (not weak) data point in such a way that the fitted value falls within the estimated error value, contributes a value of approximately 1 to the Q-value. Contributions resulting from fitting downweighted weak values are usually so small that they may be ignored.

The theoretical Q-value can be approximated by the user as $nm - p(n+m)$, where n is the number of species, m is the number of samples in the dataset, and p is the number of factors fitted by the model (Paatero and Hopke, 2009).

It is useful to look at the changes in the Q-value as additional factors are calculated. After an appropriate number of factors are included in the fit, additional factors will not result in further significant improvements in the Q-value.

It should be noted that the absolute level of Q-values depends strongly on the assumed uncertainties. Usually, it is not recommended to change uncertainties just to get closer to the theoretical Q-value (Brown and Hafner, 2005). If uncertainties have been adjusted to produce a reasonable Q-value, then the Q-value can no longer be considered a goodness of fit indicator (Paatero, 2010). However, the differences of Q-values obtained with different numbers of factors are useful indicators even with adjusted uncertainties. If introducing another factor lowers the Q-value only by the number of additional factor elements, then the introduced factor should be rejected.

Useful information can be retrieved by comparing the theoretical Q-value to Q(true) and Q(robust) values, which are output by each run of the EPA-PMF. Q(robust) is calculated by excluding outliers and the Q(true) includes all points. Solutions where Q(true) is 1.5 times greater than Q(robust) may indicate that the model is inconsistently modelling the data. Outliers may be causing this phenomenon and can be downweighted by the user so that they have less influence in the model (Brown and Hafner, 2005; Paatero, 2010). Weak variables (i.e. species with low S/N values as defined in section 6) may also be downweighted.

A good fit of the data is characterised by values for Q(robust) and Q(true) that are near to the theoretical Q-value calculated by the user (Brown and Hafner, 2005).

Examining the scaled residuals

The scaled residual is the ratio of the PMF-modelled residual e_{ij} to the input uncertainty σ_{ij} :

$$\frac{e_{ij}}{\sigma_{ij}} = \frac{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}{\sigma_{ij}} \qquad (9.2)$$

In PMF analysis, plotting the scaled residuals is also useful in choosing the final number of factors. These residuals should be symmetrically distributed within a range of -3 to +3 (and preferably less). If the scaled residuals are especially large (<<-3 or >>+3) for certain variables, then one may consider that perhaps the uncertainties specified for these variables are too small. If the scaled residuals are especially small (close to zero) for one variable, then either overly large uncertainties have been specified or this variable is explained by a unique factor. It may be acceptable to have a unique factor for a specific variable, but it must make physical/chemical sense for the problem under consideration. A spurious unique factor may arise if uncertainties that are too small are specified for a species. Too many very narrow distributions suggest the presence of too many factors such that the solution is fitting the errors rather than the concentration values. A strong skewness in the scaled residual plots suggests that the fit is not correct and that other solutions should be sought.

Examining the regression parameters

If in the original dataset there is a good mass closure (i.e. the sum of the mass of the single chemical components is close to the gravimetric mass), the “external mass” method - i.e. where the PM mass is not included in the data array analysed by PMF - can be applied. In this case, the measured mass is regressed against the estimated source contribution values. If the regression produces negative parameters, then too many factors have been included in the solution (Kim et al., 2003), or a strong source does not emit any of the measured species and hence is not represented in any factor but only in PM mass.

The regression parameters can be also used to obtain the scaled source/factor profiles. Once the source profiles are scaled, they can be summed together. It can then be determined whether the sum of a source/factor profile exceeds 100% (within a 20% tolerance level to account for errors). If this is the case, too few factors may have been chosen (Kim et al., 2003; Hopke, pers. comm.).

Examining the species/mass reconstruction

The appropriateness of the chosen solution can be also assessed by looking at the mass/species reconstruction, which should improve when approaching the best solution.

In the EPA-PMF, there is a regression analysis of the variable with its reconstructed values that provides some measure of the fit to the measurements. However, these regressions are unweighted and, thus, values that are below the detection limit or are missing have a large influence on the results and can produce degraded r^2 values (see section 5). To overcome this issue, regressions with weighted values should be calculated manually.

Examining the IM and IS parameters

The maximum individual column mean (IM) and the maximum individual column standard deviation (IS) parameters can be also used to identify the number of factors in a PMF. When the number of factors increases to a critical value, the IM and IS values will drop dramatically (Lee et al., 1999).

Examining multiple solutions

It is essential to perform the PMF analysis several times (typically ≥20) to be certain that the same solution is obtained. A test for the best selection of the number of factors is that one does not obtain multiple solutions or obtains at most one alternative solution. With greater or fewer factors than the optimum, multiple solutions are more often obtained.

In general, any bilinear factor analysis has **rotational ambiguity**. In other words, there is no unique solution even though there is a global minimum in the ‘least squares’ fitting process. The addition of constraints can reduce the rotational freedom in the system, but non-negativity alone does not generally result in a unique solution. One of the key features of PMF is that the rotations are part of the fitting process and are not applied after the extraction of the factors, as is done in eigenvector-based methods.

Controlling rotations by the FPEAK value

FPEAK is a parameter used to explore the rotational ambiguity of a PMF solution ‘a posteriori’. Assigning positive or negative FPEAK values produces rotations of which the suitability is assessed by observing the changes of the Q-value and the G and F factors. The mathematically optimum solution in PMF is FPEAK=0.0. Therefore, in the absence of any other consideration such as G-space plots (see below), and unless there is a substantial improvement in the interpretability of the profiles, the best fit is given by FPEAK = 0.0.

Controlling rotations by imposing external information

PMF rotations can also be controlled by imposing external information on the solution.

Fkey and Gkey constraints consist of binding individual elements of the F and G matrices, respectively, to zero. If specific values of profiles or time series are known to be zero, then it is possible to force the solution toward zero for those values through appropriate settings of the rotational tools Fkey and Gkey values in PMF2. Controls can be applied through modifying the script in Multilinear Engine-2 (ME-2) applications (see section 4).

In EPA-PMF v5 the Expression Builder is available to define additional constraints using three different methods (EPA-PMF 5.0 User Guide):

- **Ratio** – Predefine the ratio of two species in a specific factor (e.g. $Si_{soil}/Al_{soil}=3.2$)
- **Mass Balance** – Add a constraint using the mass balance between species across one or multiple factors (e.g. $Fe_{diesel}=2.5\ Fe_{gasoline}$)

Custom – Specify a constraint using a custom expression (e.g. $G_{smelter}=0$ from August to December 2008). The applied constraints are “soft pulling”, therefor a limit of change in the Q value must be specified. The default value for this limit is set to % dQ = 0.5 but can be updated if needed. See section 12 for more details.

Controlling rotations by examining G-space plots

G-space plots are source contribution scatter plots for pairs of factors (Paatero et al., 2005). When factors are plotted in this way, unrealistic rotations appear as oblique edges that define correlation between the factors. Edges are well-defined straight borders between regions that are densely populated with points and regions where no points occur. With a correct rotation, the limiting edges usually coincide with, or are parallel to, the axes.

Inspection of the plots helps choose a realistic rotation, but one must bear in mind the fact that specific physical situations might occur where there is an oblique edge even though a realistic rotation has been achieved. As an example, correlations could be induced by co-location of the sources coupled with meteorology (i.e. emissions from independent sources may have a certain degree of correlation due to the influence of meteorology).

It must be also emphasised that the presence of aligned edges in G-plots of factors does not necessarily guarantee that a unique rotation has been found. Although uniqueness will probably be the case in most practical situations, there may be special cases in which the rotational uniqueness does not hold.

References:

Brown S.G, Hafner H.R., 2005. Multivariate Receptor Modeling Workbook. http://www.sonomatechdata.com/

Hopke P.K., 2000. A guide to Positive Matrix Factorization, http://www.epa.gov/ttnamti1/files/ambient/pm25/workshop/laymen.pdf

Kim E., Larson T.V., Hopke P.K., Slaughter C., Sheppard L.E., Claiborn C., 2003. Source identification of PM2.5 in an arid Northwest U.S. city by positive matrix factorization. *Atmospheric Research* 66, 291-305.

Lee et al., 1999. Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong, *Atmospheric Environment* 33, 3201-3212

Paatero P., Tapper U., 1994. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics* 5, 111-126.

Paatero P., 1999. The multilinear engine - A table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *Journal of Graphical Statistics* 8, 854-888.

Paatero P., Hopke, P.K., Begum, B.A., Biswas, S.K., 2005. A graphical diagnostic method for assessing the rotation in factor analytical models of atmospheric pollution, *Atmospheric Environment* 39: 193-201

Paatero P. and Hopke P.K., 2009. Rotational tools for factor analytic models. *J. Chemometrics* 23, 91-100

Paatero P., 2010. User's guide for Potive Matrix Factorization programs PMF2 and PMF3, Part 1: tutorial. Available @ http://www.helsinki.fi/~paatero/PMF/pmf2.zip

Reff, A., Eberly, S.I., Bhave, P.V., 2007. Receptor modelling of ambient particulate matter data using Positive Matrix Factorization: review of existing methods. *Journal of Air & Waste Management Association* 57, 146-154.

10. FACTOR ANALYSIS II: EVALUATION OF SOURCE CONTRIBUTION ESTIMATION AND MODEL PERFORMANCE INDICATORS

Principal Component Analysis – Multilinear Regression

Different techniques exist to carry out source contribution estimations by performing multi-linear regression of the principal components versus the total PM mass: APCS (Absolute Principal Component Scores; Thurston and Spengler, 1985), APCA (Absolute Principal Component Analysis; e.g. Swietlicki and Krejci, 1996), and PCA-MLR (Principal Component Analysis – Multilinear Regression; e.g. Tauler et al., 2008).

In the following, these techniques are referred to as APCA.

This analysis may be carried out using numerous statistical software packages, many of them freely available. The computation of source contributions with APCA is characterised by:

- no specific software required
- fast source identification
- relatively time-consuming source contribution estimation.

However, this analysis suffers from three strong limitations:

1. Given that non-negativity constraints are not included in APCA, negative regression coefficients may be obtained. As a result, the output could show negative source contributions (in terms of mass). Two different approaches are generally used to solve this issue: including the resulting negative mass concentrations in the final result of APCA, even though this has no physical meaning, or eliminating the negative values by replacing them with zero or an empty cell. Evidently, the results obtained after the application of one or the other approach may vary largely. Thus, in the absence of consensus regarding the issue of negative regression coefficients, APCA solutions may be prone to high subjectivity and lack of comparability.

2. The number of sources resolved by APCA is generally lower than that obtained with other models (e.g. PMF or CMB), and quantification of source contributions may not be as precise (Table 10.1).

3. APCA has lower flexibility with respect to PMF and CMB, with fewer valid solutions being produced (Table 10.2).

Table 10.1. Sources and source contributions obtained during a receptor model inter-comparison (Viana et al., 2008).

APCA		**PMF**		**CMB**	
Sources	**% PM_{10}**	**Sources**	**% PM_{10}**	**Sources**	**% PM_{10}**
Individual sources					
Clay	31	Clay	16	Clay	41
Industrial#1	15	Industrial	16	Industrial#1	4
Industrial#2	2			Industrial#2	2
Vehicular	10	Vehicular	10	Vehicular	13
Regional+marine	34	Regional+marine	23	Regional	18
		Regional SO_4^{2-}	25	Marine	3
				Soil	12
Undetermined	8	Undetermined	10	Undetermined	7
Grouped sources					
Mining&Industry	**48**	**Mining&Industry**	**32**	**Mining&Industry**	**47**
Vehicular	**10**	**Vehicular**	**10**	**Vehicular**	**13**
Regional	**34**	**Regional**	**48**	**Regional**	**33**

Source contributions are divided into the sources obtained directly by the receptor models, and grouped into three main source categories to facilitate the intercomparison of results.

Several model performance indicators are available when applying APCA. These are relatively simple to use in order to assess:

a. Correlation between modelled and measured PM mass.

b. Chemical mass closure: sum of the estimated source contributions, and comparison with total measured PM mass.

c. Average absolute error (AAE): the average of the absolute percentage differences between the estimated and experimental PM mass data, when different numbers of sources or factors are considered (Chow et al., 2007, Table 10.2).

d. Correlation between modelled and measured known sources: the most commonly used source for this test is the marine source, calculated as the sum of the chemically determined Na and Cl in study areas with no other major sources of these elements.

Table 10.2. Number of solutions obtained during a receptor model intercomparison (Viana et al., 2008), and average absolute errors (AAEs) for each of the solutions computed using APCA, PMF and CMB.

	APCA		**PMF**		**CMB**	
Number of factors	**AAE (%)**	**AAE$_{max}$ (%)**	**AAE (%)**	**AAE$_{max}$ (%)**	**AAE (%)**	**AAE$_{max}$ (%)**
3	18	77	19	89	67	85
4	NS	NS	17	81	13	34
5	14	69	16	73	11	34
6	NS	NS	19	69	11-36*	25-68*
7	NS	NS	16	69	11	20
8	NS	NS	15	68	NS	NS

* range depending on which source profiles are excluded.

In addition, model performance may also be tested by:

- Understanding the data and the solution: Does it make sense? Is it reasonable?
- Combination with other tools (e.g. back-trajectory analysis)
- Combination with other models (PMF to refine the quantification of source contributions and potentially obtain uncertainty estimates)
- Correlation with EU guideline methods (SEC(2011) 208) for natural aerosol sources (e.g. African dust, sea-spray).

In conclusion:

a. APCA is an exploratory receptor modelling tool for urban air quality management, i.e. for the design of air pollution mitigation strategies, because of:
    i. Fast source identification
    ii. The size of datasets required could potentially be available from air quality monitoring networks, from at least one selected station/network
    iii. Easy to interpret model performance indicators, e.g. average absolute errors (AAE)
    iv. However, the quantification of source contributions is rather inaccurate, and source contributions (in terms of mass) are subject to uncertainty due to the absence of the non-negativity constraint.

b. In scientific applications, APCA should mainly be used to obtain a preliminary picture of the possible contribution sources, as a preparatory step for the use of more advanced models (PMF, CMB, etc.).

PMF (Positive Matrix Factorization)

In PMF, Q values indicate how well the model fits the input data. Q(robust) is calculated by excluding outliers while Q(true) is calculated including all data points. The expected (theoretical) Q is $nm - p(n+m)$, where n is the number of species, m is the number of samples in the dataset, and p is the number of factors fitted by the model (e.g., Norris et al., 2008).

An alternative estimation distinguishes weak from good species as follows (Brown and Hafner, 2005): Q = (# samples * # good species) + [(# samples * # weak species)/3] – (# samples * # factors being estimated)

In addition, several diagnostic tests are embedded in the EPA PMF v3 and v5 softwares to evaluate the runs: residual analysis, observed vs predicted scatter plot and time series, combined plots of profiles and contributions and box plots to summarise the distribution of the contributions, G-space plots (or G-plots) and factor pie charts.

Various elements can contribute to the uncertainty in the solutions modelled by PMF, including temporal variation of particulate matter (PM) source profiles, measurement error, sampling variability, and intrinsic limitations in the modelling process, such as rotational ambiguity and incorrectly specified number of factors (see section 9).

In PMF2, it is possible to estimate uncertainties in the F and G matrices (eq. 9.1) using the process originally described by Roscoe and Hopke (1981) and described in detail by Malinowski (1991). The errors in the elements of one matrix are estimated based on the errors in the ambient concentration values, assuming that the other matrix is error-free. Each matrix (F or G) is treated similarly in such a way that an uncertainty value is associated with each element of the matrix.

The standard deviation of the source contribution estimates (SCE) of every factor in all the samples can be used as an estimation of the uncertainty of the average SCEs.

Bootstrapping can be used to determine the precision of PMF profiles by calculating the standard deviation (assuming normality) or various percentiles of factor profiles (F-matrix values) from numerous bootstrap runs. Nevertheless, to obtain a better representation of the component of uncertainty associated with rotational ambiguity, an improved error estimation scheme has been proposed by Paatero et al. (2013), which is currently available in the EPA-PMF v5. The new scheme combines bootstrapping and a “displacement” technique based on the controlled perturbation of factor elements.

Advanced User Box: Q/Qexp ratio (by P. Paatero)

If the data matrix, error matrix, and analysis fulfil the PMF model assumptions the variability of Q values may be statistically estimated. This estimation of Q variation is based on the width of a chi-2 distribution and is very reliable. For large matrices, the variation is typically less than 10% of the expected Q. This holds also for the overall Q and for the column Q values of not-so-large matrices. For row Q values of small matrices (less than 20 good species) the variation may be up to a factor of 2. This statistically estimated variability is solely due to random variations in data noise, from row to row and from column to column.

If the observed Q values exceed significantly these estimated limits it is possible that not all model assumptions are fulfilled.

Examples of situations that may cause high Q values are:

- number of factors in PMF model is smaller than number of significant sources in the measured data,

- data uncertainties are incorrectly estimated for some species and/or for some samples,

- there are gross errors in some data values or in some uncertainty values (this happens surprisingly often, especially if values are typed manually in the file and no proofreading is performed),

- profiles of sources do not stay constant with time,

- there are sporadic sources, too weak to support dedicated factors but strong enough to disturb the model,

- there are incompatible species that do not correlate with any other species.

There are no fixed rules telling what to do when model assumptions are not fulfilled. Of course, such a situation must be openly discussed when the results are published.

For mass-spectral data, a limit for the overall Q/Qexp is reasonable if there are no modelling errors. In this case, the overall Q should not exceed the expected Q by more than a few tens of percent. Note that the expected Q for downweighted columns should be correctly evaluated; the expected Q contribution from those columns should be computed so that downweighting is taken into account.There are cases in which the substitution of many values below the detection limit in several trace species creates an artificial factor containing trace species with a characteristic pattern. This “ghost factor” is generated by the model to fit the substituted values in all those species in which they occur simultaneously. When using ME-2-based analysis tools, it is possible to avoid such distortions by using the special error model code -16 (Paatero, 2000) for all substituted data points. This code stipulates that all fitted values below the detection limit are to be considered a perfect fit, with Q contribution (see section 9) equal to zero. This alternative is not yet implemented in version 3.0 of the US - EPA Positive Matrix Factorization (PMF), but it may be used when controlling ME-2 using home-made scripts.

References:

Brown S.G, Hafner H.R., 2005. Multivariate Receptor Modeling Workbook. http://www.sonomatechdata.com/

Chow, J.C.; Watson, J.G.; Chen, L.-W.A.; Chang, M.-C.O.; Robinson, N.F.; Trimble, D.L.; and Kohl, S.D., 2007. The IMPROVE_A temperature protocol for thermal/ optical carbon analysis:

Maintaining consistency with a long-term data base. J*ournal of Air & Waste Management Association* 57:1014-1023.

Malinowski E.R., 1991. Factor Analysis in Chemistry, Wiley, New York, 2nd Ed.

Norris G., Vedantham R., Wade K., Brown S., Prouty J., Foley C., 2008. EPA Positive Matrix Factorization (PMF) 3.0 Fundamentals & User Guide. EPA 600/R-08/108.

Roscoe B.A., Hopke P.K., 1981. Error Estimation of Factor Loadings and Scores Obtained with Target Transformation Factor Analysis. *Analytica Chimica Acta* 132:89-97.

Paatero P., Eberly S., Brown S.G., Norris, G.A. 2013. Methods for estimating uncertainty in factor analytic solutions. *Atmospheric Measurement Techniques Discussions*, 6, 7593-7631.

SEC(2011) 2008. COMMISSION STAFF WORKING PAPER. Establishing guidelines for demonstration and subtraction of exceedances attributable to natural sources under the Directive 2008/50/EC on ambient air quality and cleaner air for Europe.

Swietlicki, E., Krejci, R., 1996. Source characterisation of the Central European atmospheric aerosol using multivariate statistical methods. Nuclear Instruments and Methods in Physics Research, Section B: Beam Interactions with Materials and Atoms 109-110, 519-525.

Tauler, R., Paatero, P., Henry, R.C., Spiegelman, C., Park, E.S., Poirot, R.L., Viana, M., Querol, X., Hopke, P.K., 2008. Identification, Resolution and Apportionment of Contamination Sources, vol. 3, pp. 269-284.

Thurston, G.D., Spengler, J.D., 1985. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. *Atmospheric Environment* 19, 9-25.

Viana, M., Pandolfi, M., Minguillón, M. C., Querol, X., Alastuey, A., Monfort, E., Celades, I., 2008. Inter-comparison of receptor models for PM source apportionment: Case study in an industrial area. *Atmospheric Environment* 42, 3820-3832.

11. FACTOR ANALYSIS III: CRITERIA FOR FACTOR ASSIGNMENT

The most subjective and least quantifiable step in applying PMF for source apportionment is the assignment of identities to the factors chosen as the final solution. It is important for the data analyst to know what types of sources are present in the study area. However, even in cases where there are good emission inventories, there can be situations where a source cannot be identified (Hwang and Hopke, 2006). In addition, atmospheric processes may result in multiple factors such as summer and winter secondary sulphate, or in producing sufficiently collinear sources that an irresolvable mixture of source profiles is obtained. Thus, profiles have to be interpreted with both knowledge of the study area and a background in atmospheric science. For that reason, any choice concerning the correspondence between source categories and factors must be supported by objective and quantitative tests.

High shares of a source marker in a factor profile may be used for a preliminary source attribution. However, further evidence is required for confirmation of this initial hypothesis.

Proposed steps to support factor assignment:

- Compare the obtained factor profiles with those reported in previously published PMF studies (the comparison can be performed either visually or numerically using, for instance, the Pearson coefficient);
- Search the literature for measured PM source profiles with characteristics similar to the factor profiles in the F-matrix;
- Search for measured PM source profiles in relevant databases (e.g. SPECIATE, SPECIEUROPE);
- Identify the source by comparing certain species ratios (also referred to as “enrichment factors”) in PMF source/factor profiles to the same ratios in measured PM source profiles (see also section 18);
- Perform local and/or regional source sampling along with the ambient PM sampling to develop source profiles needed to identify PMF profiles;
- Look at temporal patterns for “expected” behaviours (e.g. the largest contributions of a source believed to be residential wood burning should likely occur during winter months); plots of contributions over time can be inspected in order to look for daily, weekly, seasonal, and yearly oscillations of source contributions. Mean source contributions by season and by day of the week (weekend versus weekday) should also be examined;
- Check the correlations between factor time series and time series of external tracers.

It should be noted that when source profiles are not independent (i.e. there is severe collinearity) it is difficult to separate their contributions. In this case, additional chemical/physical information is needed to improve source segregation. Nevertheless, sources can clearly be separated for a sufficiently low level of collinearity and precision in the input data. In spectrochemical problems, good factors can be obtained despite quite severe collinearity. However, the collinearity inflates the uncertainties of the values (Cheng et al., 1988).

Advanced User Box

Auxiliary analyses can be used to aid in the identification of PMF factors: e.g. contribution of wind roses, conditional probability function, potential source contribution function, cluster analysis, and residence time analysis are some techniques for analysing wind or backward trajectories (see section 17).

References:

Cheng, M. -D, Hopke P. K., Jennings D. E., 1988. The Effects of Measurement Errors, Collinearity and their Interactions on Aerosol Source Apportionment Computations. *Chemometrics and Intelligent Laboratory Systems* 4 (3), 239-250.

Hwang I.J. Hopke P.K., 2006. Comparison of Source Apportionments of PM2.5 at Two San Jose Speciation Trends Network Sites. *Journal of Air & Waste Management Association* 56: 1287-1300

12. FACTOR ANALYSIS IV: CONSTRAINED AND EXPANDED MODELS

Common Positive Matrix Factorization (PMF) analysis fits the data into a limited bilinear model. However, PMF also permits the development of more complex mathematical models to account for physical and chemical information when fitting the data. “Constrained” and “Expanded” PMF models represent the advanced tools in receptor modelling, and efforts are being made to improve and increase their capabilities. Since this is a new field of research, the terminology is still evolving in the literature and the distinction between Constrained and Expanded models is not very well defined yet. In this document, Constrained models are considered a subcategory of Expanded (or Extended) models. More precisely, Constrained models are those in which additional constraints are introduced (in most cases after an initial run, the results of which are used as a starting point), while Expanded models are those in which the customary bilinear equation is augmented by another more complicated set of equations, depending on the aims of the study.

Constrained PMF

By definition, the Positive Matrix Factorization model is a weighted least squares analysis where the object function is minimised under the constraint that all or some of the elements of G and F are constrained to non-negative values (Paatero, 1997). Therefore, all PMF studies are constrained. Nevertheless, recent literature uses the term ‘constrained’ to refer to more complicated PMF models, where the constraint is not limited to non-negativity.

Different types of constraint can be implemented in PMF, but they must all derive from some a priori knowledge of the user about the system that is to be modelled. This knowledge can be of physical or chemical origin (Amato et al., 2009). Physical constraints can relate, for example, to the mass conservation principle (e.g. the sum of factor profiles cannot exceed unity; the lower the particle size, the lower the source contribution, etc.). On the other hand, chemical information is associated with source profiles. The relative abundance of some elements/compounds may already be known and can represent valuable information for the model in order to find a better solution, reducing the number of possible alternatives (the ‘rotational ambiguity’). Another example of a priori knowledge is the information about periods during which a specific point source is not operative. These data can be useful constraints to drive the model towards a more realistic solution by setting the emission of that source to zero.

The choice of the program to use in performing a Constrained PMF depends on the type of constraint to be used:

- PMF2 implements only the Fkey and Gkey constraints, which consist of binding individual elements of the F and G matrices, respectively, to zero. $Gkey_{ik}$ and $Fkey_{kj}$ are two matrices of the same shapes as G and F respectively. They are applied a posteriori (in a ‘continuation run’ that takes place after the base run) and each element of the matrix with a key value >1 is bound to zero, with an increased strength of the bond for higher key values. It is not possible to bind elements to non-zero values. Both constraints are imposed regardless of changes in the Q value, i.e. they are considered to be “hard” constraints.

- ME-2 (Multilinear Engine) is a special-purpose programming language, which allows for the incorporation of any additional constraints that are introduced by the user into the script (Paatero, 1999; Paatero and Hopke, 2009; Amato et al., 2009; Amato and Hopke, 2012). The constraints can be introduced in terms of pulling equations, upper/lower limits and fixed values. Pulling equations are weighted by uncertainties, which express the confidence of the user in the equation. A lower uncertainty corresponds to a harder pulling effect. Each pulling equation is converted into an auxiliary term of the object function to be minimised.
- As mentioned in chapter 9, EPA PMF v5 includes a user-friendly interface for introducing constraints in several ways:
    - Ratios of F constituents (e.g. $Si_{soil}/Al_{soil}=3.2$)
    - Mass Balance between F elements of the same or different factors (e.g. $Fe_{diesel}=2.5\ Fe_{gasoline}$)
    - Custom expression, where the user can build any kind of equation on F and G elements, based on a priori knowledge (e.g. $G_{smelter}=0$ from August to December 2008)

  When F and/or G elements are set to zero or confined to upper/lower limits, the constraints are “hard” or imposed without regard to the change in the Q value. Equations constraining variables towards a value, upwards or downwards, are classified as “soft” pulling, and their strength (based on the confidence of the user) is expressed by the limit of change allowed in the Q value. A higher dQ will determine a harder constraint.

Once the constraints are applied in a continuation run, the user should look at the deviations in the results between the two model runs and examine the impact of the constraints on:

- Achievement of the target values (within the uncertainty range in the case of ME-2)
- The increase of dQ
- Correlations between factor profiles and reference source profiles
- Changes in G-space plots
- Possible distortions in all factors and source contributions
- Possible factor swaps, so that identities of factors have changed. Such swaps cause the constraints to act on physical factor(s) that are different from those originally intended, so the constraints are meaningless. For details, see Paatero et al. (2013).

Sensitivity tests - carried out modifying the strength of pulling equations - can be useful for a comprehensive evaluation of the model output (Viana et al., 2009; Brown et al., 2012).

Expanded PMF

As already mentioned, the Multilinear Engine (ME-2) has been used to constrain PMF profiles and contributions. However, the flexible structure of ME-2 makes it suitable for solving any other complex problems such as expanded models. ME-2 has been applied to several datasets for multiple purposes.

In general, the expanded models were found to give similar source contributions and source profiles when compared with the original PMF analyses, but also provided information associated with the meteorological and temporal conditions. In some cases, the expanded model provided additional resolution of sources: Kim et al. (2003) were able to resolve diesel and gasoline emissions using the expanded model when they had been unable to do so with the basic bilinear factor analysis model. However, much more equivocal results were obtained for Washington, DC relative to the prior PMF analyses (Begum et al., 2005).

Zhao et al. (2004) developed a novel factor analysis model, in which the normal chemical mass balance model was augmented by a parallel equation that accounted for wind speed and direction, temperature, and weekend/weekday effects. The model was fitted with a multilinear engine (ME) to provide identification and apportionment of the VOC sources in Houston during the Texas Air Quality Study (TexAQS) 2000. The analysis determined the profiles and contributions of nine sources and the corresponding wind speed, wind direction, temperature, and weekend factors. The reasonableness of these results suggested the high resolving power of the expanded factor analysis model for source apportionment, but also provides novel and effective auxiliary information for more specific source identification. This study demonstrates the feasibility of the expanded model to identify sources in complex VOC systems.

Zhou et al. (2009) developed an expanded model to investigate the effect of wind direction, wind speed, seasons, and weekdays/weekends in the Cleveland (Ohio, USA) area. The expanded model and PMF2 produced essentially the same results with only minor differences being observed between the two sets of profiles and contributions. Thus, the addition of meteorological and temporal parameters to the model did not improve the source resolution. Zhao and Hopke (2006) followed a similar approach in Indianapolis and conclude that PMF coupled with a posteriori back-trajectory analysis (such as CPF, PSCF, seasonal variation analysis, and weekday/weekend variation analysis) yields comparable results to expanded factor analysis and is simpler to employ.

New monitoring technologies permit the measurement of a variety of chemical species with time resolution as high as 10 minutes to one hour. However, most species are still measured with longer integration periods such as several hours to a day. Traditional factor analysis techniques (PCA and customary PMF) are unable to analyse datasets consisting of different time scale measurements. Zhou et al. (2004) developed an expanded PMF model which can use each data value (of a mixed time-resolution dataset) in its original time schedule without averaging or interpolation. Averaging the high time-resolution data leads to a loss of valuable temporal information, while interpolating the low time-resolution data produces unreliable high-resolution series. The contribution series are smoothed by the regularisation of auxiliary equations especially for sources containing very little high-resolution species. Similar study designs were followed by Ogulei et al. (2005) and Zhao et al. (2004).

An expanded receptor model was applied to identify and apportion the $PM_{2.5}$ sources that were common to three different environments (personal, indoor and outdoor) and to which asthmatic children were exposed (Zhao et al., 2007). Two types of sources (factors), external and internal, were defined – the external sources were left free to contribute to all three environments while the internal sources were constrained to only contribute to the personal and indoor samples. The expanded receptor model was expressed as:

$$x_{ijdt} = \sum_{p=1}^{N} g_{ipdt} f_{jp} + \sum_{p=N+1}^{N+H} g_{ipdt} f_{jp} \qquad \text{(t=1 for personal; t=2 for indoor)}$$

$$x_{jdt} = \sum_{p=1}^{N} g_{pdt} f_{jp} \qquad \text{(t=3 for outdoor)}$$

where *i* is the individual index, *j* the species index, *d* the sampling date index, *t* the type index, *N* the number of external sources, and *H* the number of internal sources. x_{ijdt} denotes the concentration of species *j* in the sample of type *t* collected by subject *i* on date *d*, g_{ipdt} denotes the contribution of source *p* to the sample of type *t* collected by subject *i* on date *d*, and f_{jp} denotes the relative concentration of species *j* in source *p*. Further information is available in Hopke et al. (2003) and Zhao et al. (2006). This approach was able to resolve four external sources (sulphate, soil, nitrate and traffic) and three internal (chlorine-based cleaning, cooking, tobacco) sources. Strict bilinear (PMF2) and trilinear models (PMF3) were also applied to indoor-outdoor-personal samples (Larson et al., 2006).

Pere-Trepat et al. (2007) analysed data that combined particle size and composition data using an expanded PMF model to permit the extraction of maximal information from size-segregated aerosol composition data. This three-way model accounts for the variation in the composition of the source emissions in the different size ranges: three-stage DRUM impactor (Pere-Trepat et al., 2007), 12-stages cascade impactor (Bernardoni et al., 2017). The data are three-way in that their size and composition are measured over time. Three-way data have been also analysed by means of the stricter trilinear PARAFAC model (Yakovleva et al., 1999; Hopke et al., 2003), which does not offer the flexibility of ME-2.

With the ME-2 approach (Pere-Trepat et al., 2007), each profile is a matrix of *m*x*n* dimension where *m* is the number of measured variables and *n* the number of measured size fractions. The profiles are then a three-dimensional array of source by composition by size. For each source (factor), there is a vector of mass contributions, so combining them produces a matrix whose dimensions are defined by the number of sampling days by the number of sources (factors).

This model evolved from the Tucker 1 model (Tucker, 1966). This model is logically a two-way model, but is organised as a three-way array with data also in a three-way array, X. The main equation of the model is as follows:

$$\vec{X} = A \otimes \vec{B} + \vec{E}$$

where $\vec{X}(i,j,k)$ is the three-way array of observed data, $\otimes$ represents a Kronecker product of the source profile array $\vec{B}(i,j,k)$ with the contribution matrix $A(i, p,$ and $\vec{E}(i,j,k)$ is a three-way array of residuals (Pere-Trepat et al., 2007).

Developing new models using ME-2.

As already mentioned, new models can be developed by modifying existing ME-2 scripts, or by writing entirely new ones. Due to the difficulties normally encountered in debugging new scripts, practitioners are advised to use existing code as much as possible. To that end, developers are encouraged to obtain information on existing script material, preferably in the early stages of their work. In this way, they also contribute to guiding the future development of ME-2 in directions that are most useful for the further development of the source apportionment methods.

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. *Atmospheric Environment* 43, 2770–2780.

Amato, F., Hopke, P.K., 2012. Source Apportionment of the Ambient PM2.5 in East St. Louis using Constrained Positive Matrix Factorization. *Atmospheric Environment* 46, 329-337.

Begum, B.A., Hopke, P.K., Zhao, W., 2005. Source Identification of Fine Particles in Washington DC by Expanded Factor Analysis Modeling. *Environmental Science and Technology* 55, 227-240.

Bernardoni V., Elser M., Valli G., Valentini S., Bigi A., Fermo P., Piazzalunga A., Vecchi R., 2017. Size-segregated aerosol in a hot-spot pollution urban area: Chemical composition and three-way source apportionment. Environmental Pollution 231, 601-611.

Brown, S.G., Lee, T., Norris, G.A., Roberts, P.T., Collett, J.L., Paatero, P. Worsnop, D.R., 2012. Receptor modeling of near-roadway aerosol mass spectrometer data in Las Vegas, Nevada, with EPA PMF. *Atmospheric Chemistry and Physics* 12, 309-325.

Hopke, P.K., Ramadan, Z., Paatero, P., Norris, G., Landis, M., Williams, R., Lewis, C.W., 2003. Receptor modeling of ambient and personal exposure samples: 1998 Baltimore particulate matter epidemiology–exposure study. *Atmospheric Environment* 37, 3289–3302.

Kim, E., Hopke, P.K., Paatero, P., Edgeton, E.S., 2003. Incorporation of Parametric Factors into Multilinear Receptor Models Studies of Atlanta Aerosol. *Atmospheric Environment* 37, 5009-5021.

Larson, T., Gould, T., Simpson, C., Liu, L.-J.S., Claiborn, C., Lewtas, J., 2004. Source apportionment of indoor, outdoor, and personal PM2.5 in Seattle, Washington, using positive matrix factorization. *Journal of the Air and Waste Management Association* 54 (9), 1175-1187.

Ogulei, D., Hopke, P.K., Zhou, L., Paatero, P., Park, S.S., Ondov, J.M., 2005. Receptor modeling for multiple time resolved species: The Baltimore supersite. *Atmospheric Environment* 39 (20), 3751-3762.

Paatero, P., 1997. Least squares formulation of robust non-negativity factor analysis. *Chemometrics and Intelligent Laboratory Systems* 37, 23–35.

Paatero, P., 1999. The multilinear engine: a table-driven least squares program for solving multilinear problems, including the n-way parallel factor analysis model. *Journal of Computational and Graphical Statistics* 8, 854–888

Paatero, P., Hopke, P.K., 2009. Rotational tools for factor analytic models. *Journal of Chemometrics* 23 (2), 91–100.

Paatero P., Eberly S., Brown S.G., Norris, G.A. 2013. Methods for estimating uncertainty in factor analytic solutions. *Atmospheric Measurement Techniques Discussions*, 6, 7593-7631.

Pere-Trepat, E., Kim, E., Paatero, P., Hopke, P.K., 2007. Source apportionment of time and size resolved ambient particulate matter measured with a rotating DRUM impactor. *Atmospheric Environment* 41 (28), 5921-5933.

Tucker, L.R., 1966. Some mathematical notes on three-mode factor analysis. *Psychometrika* 31, 279–311.

Viana, M., Amato, F., Querol, X., Alastuey, A., Moreno, T., García, S., Herce D., Fernandez R., 2009. Chemical tracers of particulate emissions from commercial shipping. *Environmental Science and Technology* 43 (19), 7472–7477.

Yakovleva, E., Hopke, P.K., Wallace, L., 1999. Receptor modeling assessment of PTEAM data. *Environment Science and Technology* 33, 3645–3652.

Zhao, W., Hopke, P.K., Norris, G., Williams, R., Paatero, P., 2006. Source apportionment and analysis on ambient and personal exposure samples with a combined receptor model and an adaptive blank estimation strategy. *Atmospheric Environment* 40, 3788–3801.

Zhao, W., Hopke, P.K., Gelfand, E.W., Rabinovitch, N., 2007. Use of an expanded receptor model for personal exposure analysis in schoolchildren with asthma. *Atmospheric Environment* 41 (19), 4084-4096.

Zhao, W., Hopke, P.K., Karl, T., 2004. Source Identification of Volatile Organic Compounds in Houston, Texas. *Environmental Science and Technology* 38 (5), 1338-1347.

Zhao, W., Hopke, P., 2006. Source investigation for ambient PM 2.5 in Indianapolis, IN. *Aerosol Science and Technology* 40 (10), 898-909.

Zhou, L., Hopke, P.K., Zhao, W., 2009. Source apportionment of airborne particulate matter for the speciation trends network site in Cleveland, OH. *Journal of the Air and Waste Management Association* 59 (3), 321-331.

Zhou, L., Hopke, P.K., Paatero, P., Ondov, J.M., Pancras, J.P., Pekney, N.J., Davidson C.I., 2004. Advanced factor analysis for multiple time resolution aerosol composition data. *Atmospheric Environment* 38 (29), 4909-4920.

13. FACTOR ANALYSIS V: DATA PROCESSING OF IN SITU THERMO-DESORPTION ELECTRON IMPACT AEROSOL MASS SPECTROMETRY

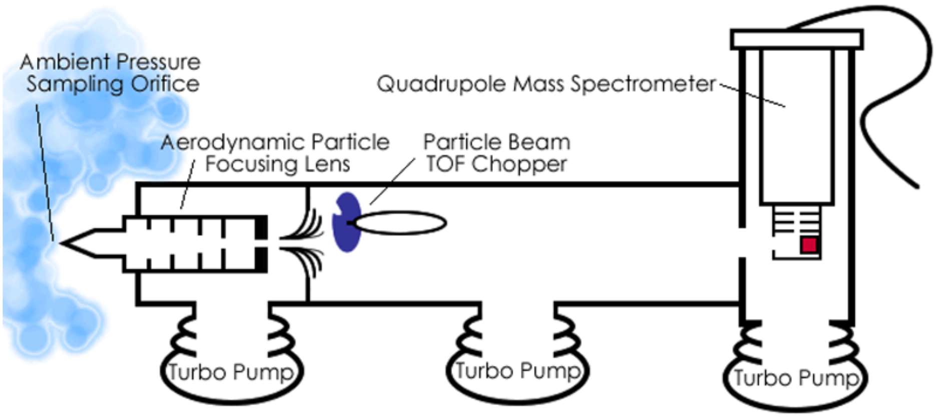
This section has been revised in collaboration with working group 2 (Source apportionment of organic aerosol) of the COLOSSAL COST action (www.costcolossal.eu/wg2/).

Despite environmental policy efforts and overall PM loading decreases, the relative contribution of organic aerosol (OA) to total fine aerosol mass is still high and relatively stable, or even shows increasing trends in Europe (e.g., Grigas et al., 2017). As a matter of fact, air quality measures implemented in Europe during the last decades succeeded in reducing inorganic aerosol concentration, whereas the limited knowledge of OA composition, formation pathway, and sources makes OA reduction a challenge (EMEP, 2016).

Analysis of speciated chemical composition of OA allows the identification of molecular markers linked to specific sources, which can be used for the quantification of OA sources (e.g. Srivastava et al., 2018a & 2018b). This approach is based on the assumption that such markers are chemically stable in the atmosphere, and that the emission factors, i.e. the molecular marker to OA ratio at emission, are known and constant. Both these assumptions are often not fully defendable under common atmospheric conditions. In addition, this approach cannot be used easily to quantify the contribution of secondary organic aerosol (SOA) - which accounts for the largest fraction of OA mass (Zhang et al., 2005) - as it requires a sound knowledge of the different formation pathways at stake, an area of knowledge generally regarded as incomplete, and detailed offline chemical analyses and/or sophisticated instrumentation (e.g., Nozière et al., 2015).

Meanwhile, the instrumental advances reached in the last two decades enabled alternative approaches based on total fine OA characterization and source's identification/quantification. Various types of instruments based on mass spectrometry have notably been developed by Aerodyne Research Inc. (ARI), Massachusetts, to provide real-time measurements of the major fine aerosol components, and quantitative information on size-resolved mass concentrations for volatile and semi-volatile components present in/on ambient aerosol particles (Jayne et al., 2000). The Aerosol Mass Spectrometer (AMS) is designed to provide quantitative composition information on ensembles of particles, with limited single particle information. The instrument combines standard vacuum, thermo-desorption, electron impact ionization, and mass spectrometric techniques with recently developed aerosol sampling techniques. A schematic representation of the AMS is shown in figure 13.1. The use of this instrument is now well-established and issues such as its composition-dependent collection efficiency are now characterised and can be taken into account (Middlebrook et al., 2012) and some features of the AMS are still evolving, also allowing for ionisation techniques other than electron impact (e.g., Onasch et al., 2012). However, the AMS remains very labour intensive to run over long time periods, while the Aerosol Chemical Speciation Monitor is specifically designed for long-term monitoring (Ng et al., 2011). A first dedicated network of ACSM instruments in Europe could then be established since 2012 (http://www.psi.ch/acsm-stations/acsm-and-emep-stations). These instruments have traditionally been limited to the study of PM_1 due to limitations of the aerodynamic lens inlet technology, however recently versions of the instruments capable of studying $PM_{2.5}$ have become available (Zhang et al., 2017).

Figure 13.1. Schematic description of the AMS



(Source : http://cires.colorado.edu/~jjose/ams.html#Aerodyne-AMS)

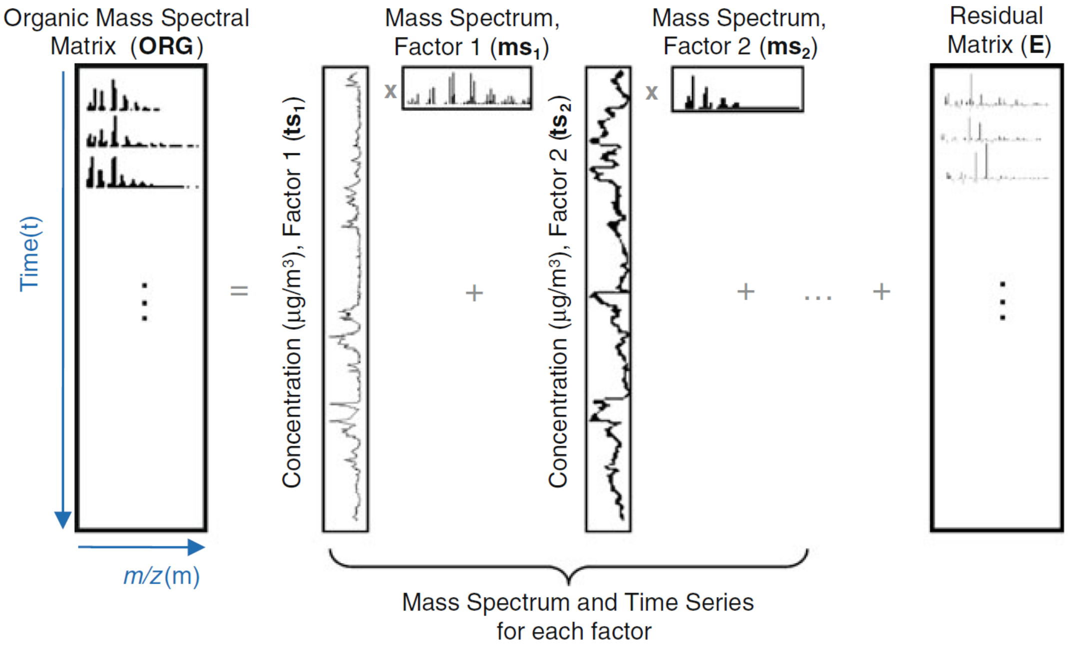
AMS and ACSM characterize OA through volatilization and ionization by electron impact (IE 70eV) of non-refractory components (i.e., aerosols that can be volatilized under flash vapaorisation at about 600°C), and analysis of m/z fragments with time of flight analyzer (e.g., HR-ToF-AMS, C-ToF-AMS, ToF-ACSM) or quadrupole analyzer (Q-ACSM). This results in time resolved mass spectra of relatively high-energy fragmented OA (Jayne et al., 2000; DeCarlo et al., 2006; Ng et al., 2011). Multivariate statistical analysis of the measured aerosol mass spectra ensemble allows the identification of OA component, which can be linked to primary sources or atmospheric processing. Accordingly, the different OA fractions are divided into hydrocarbon-like organic aerosol (often mostly from traffic), oxygenated organic aerosol (mostly secondary organic aerosol), cooking organic aerosol, biomass burning/domestic wood burning aerosol and other components (Table 13.1).

Table 13.1. Commonly reported factors using AMS or ACSM systems (adapted from Zhang et al., 2011, and Sun et al., 2017)

Factor name	Description
1. MO-OOA: more oxidized oxygenated organic aerosol (Also previously named LV-OOA due to its expected low volatility or OOA1 in early publications)	High oxygen-to-carbon atomic ratio (O/C of about 0.6), often correlates with sulphate in summer
2. LO-OOA: less oxidized oxygenated organic aerosol (Also previously named SV-OOA due to its expected semi-volatile nature or OOA2 in early publications).	Relatively low O/C ratio (approx. 0.4)
3.HOA: hydrocarbon-like organic aerosol, mainly deriving from fossil fuel combustion	Very low O/C ratio (approx. 0.05) and good correlation with primary hydrocarbon combustion emission species, for example NO_x and elemental carbon
4. BBOA (or WBOA): particulate organic aerosol from biomass/wood burning	Spectral features similar to those from wood burning emissions (with significant m/z 60 signal), often with high night-time contributions in winter. High correlation with levoglucosan or BC_{wb} from the Aethalometer model
5. COA: cooking-related organic aerosol	Spectral features similar to those of particulate organic aerosol from cooking emissions and a distinctive diurnal pattern peaking during lunch and dinner times in some urban environments.

Multivariate factor analysis is applied to deconvolute the observed organic (ORG) matrix into unique factors. As illustrated in Figure 13.2, factor analysis of the data matrices from quantitative instruments usually involves solving a mass conservation model expressed as a two-dimensional bilinear equation. In the past, the solution to the equation was found using different methods, including the custom principal component analysis (CPCA) method and multiple component analysis (MCA) (Zhang et al., 2005). Nowadays, most of the OA source apportionment studies in literature refer to Ulbrich et al., (2009) developing a Positive Matrix Factorization (PMF) Evaluation Tool (PET), or to Canonaco et al. (2013) presenting the Source Finder tool (SoFi) to support source apportionment with a-priori information. Both toolkits have been implemented into Igor Pro interfaces.

Figure 13.2. Schematic representation of an ORG matrix (from Zhang et al., 2011 adapted from Ulbrich et al., 2009)



The PET is taking advantage of the PMF2 algorithm (Paatero et al., 2007), to facilitate PMF output evaluation and related statistics. PMF analysis is usually performed in robust mode, in order to reduce the influence of observation close to detection limit, or outliers. In such a mode, data-points characterized by a large residual-to-error ratio are down-weighted, and the Q value is redefined using the reduced weights. In order to explore the solution space and potential multiple Q local minima, the algorithm can be run several times using different starting initial seeds. In addition, rotational ambiguity of the solution is investigated with *fpeak* parameters: investigating a subset of *fpeak* values allows exploring the effect of factor rotation on the final solution. Commonly used procedures for selecting the best solutions from unconstrained PMF analysis are reported in Table 13.2. They are adjusted here for AMS spectra, but can be useful for other source apportionment studies too. It should also be reminded that nothing is worth more than user’s brainpower, and knolwedge of the ambient conditions and local sources is required to better adapt these recommendations to the actual conditions of the study.

Table 13.2 Steps commonly used for preparing and choosing the best solution from unconstrained PMF analysis of AMS datasets (adapted from Zhang et al., 2011)

1. Calculate data and error matrices.
2. Further data and error treatment.
2a. Apply minimum error.
2b. Remove anomalous spikes, if desired.
2c. Smooth data, if desired.
2d. Downweight data with low signal-to-noise ratios.
2e. Downweight repeated information (m/z = 44 and related m/z values).
3. Run PMF for a range of factors (P) and random starts (seeds). Examine the ratio between the observed and expected Q (Q/Qexp) vs P. A steep change in slope indicates the minimum P to be considered.
3a. Examine results from different random starts for each P. Sort results by Q/Qexp values and compare the factors in each. Are there multiple types of results (representing local minima in the solution space)? If not, proceed to use SEED 0. If yes, can any of the result types be excluded because the factors are not physically meaningful? Proceed, exploring seeds that have results with physically meaningful factors.
3b. Try to determine the optimum number of factors by examining multiple criteria: • Check for correlations between factor time series and time series of external tracers (Table 13.3). • Examine factor mass spectra for tracer ions and fragmentation patterns. Also check for correalation between obtained mass spectra and those obtained in previous relevant studies, when available. • Investigate expected correlations between factor time series and time series of individual m/z values or ions, when appropriate (i.e., when the latter ones can unambiguously be attributed to a given factor). • Consider factor diurnal profiles, meteorological data (when appropriate), etc. • Look for signs of “split” factors**, considering the correlation of mass spectra and time series of factors in the same solution. After identifying factors that may have split, explore solutions with more factors to check for new, physically meaningful factors.
3c. Examine solution Q contributions and residuals. Do the residuals and Q values summed to form time series or mass spectra show periods or m/z values that do not fit well? Is this because the solution needs more factors, because the data do not fit the model of constant spectra for a given component, or because of instrumental drift, etc.? Are the distributions of the scaled residuals (xij/σij) for each m/z approximately Gaussian, centred around 0, with a reasonable standard deviation?
4. For the best solution chosen from step 3, run PMF for a range of FPEAKs (Paatero, 2004; Norris, et al., 2008) such that the range of Q/Qexp values is at least 3% above the minimum Q/Qexp or the algorithm fails to converge.
4a. Exclude from further consideration solutions from FPEAK analysis that have unrealistic mass spectra and/or time series.
4b. Does changing the FPEAK change the solution in a way that would change the interpretation of

the factors from step 3, or do these solutions just represent rotational ambiguity in the solution? If the interpretation changes, choose the most representative solution and support this choice. If the differences represent rotational ambiguity, choose the solution at FPEAK = 0.
5. Conduct bootstrapping analysis on the final solution from step 4 to check for the stability within the solutions. *
6. Make and examine key diagnostic plots.
6a. Q/Qexp vs varying P.
6b. Q/Qexp vs FPEAK for the best P.
6c. Fractions of OA factors vs FPEAK for the best P.
6d. Correlations among PMF factors for the best P.
6e. The box and whisker plots of scaled residuals as a function of m/z for the best P.
6f. The time series of the measured OA concentration and the reconstructed organic mass (= sum of all factors) for the best P.
6g. The variations of the residual (= measured − reconstructed) of the fit as a function of time.
6h. The time series and mass spectra of total residuals and Q contribution for the best P solution.
6i. Comparisons of the P − 1, P, and P + 1 solutions for the acceptable FPEAK, where P is the best solution.

* Uncertainty related to the instability of each factor element should be obtained as the larger of the two values: variation caused by FPEAK variation, and variation in bootstrapped results (see chapter 18).

** Co-variation of factors with similar mass spectra may not be excluded. In such a case, constrained PMF analysis can help resolving these factors.

The analysis of ACSM datasets, with lower resolution mass spectra than AMS, and/or of long-term datasets (e.g., multi-seasonal time period) highlighted the relevance of rotational ambiguity problem. The use of the multilinear engine (ME-2, see section 12) is allowing for a deeper investigation of the solution space and for achieving constrained PMF analysis (Lanz et al., 2008). The Source Finder (SoFi) toolkit, implementing ME-2, see section 12) into Igor Pro (Canonaco et al., 2013), is nowadays commonly used for this purpose. . It notably allows using a-priori information on factor/source profiles and time series. Constraining the factor/source profiles, by setting the variability range per each factor (a-value), the algorithm performs source apportionment analysis using a hybrid chemical mass balance-positive matrix factorization approach.

To identify the best set of a-values per each factor, a multidimensional sensitivity analysis is suggested. The solutions obtained from all possible combination of a-values are explored based on pre-defined acceptance criteria, such as temporal correlation with external tracers, average factor to external tracer ratios, and diurnal profile analysis (Crippa et al., 2014; Froehlich et al., 2015). To explore the solution space and rotation ambiguity, several bootstrap runs are performed using optimal combination of a-values. The source apportionment solution is then identified as the mean of bootstrap runs, while the standard deviation of the mean quantifies the solution uncertainty (Brown et al., 2015).

Table 13.3 Examples of factors with their relevant ions and external tracers commonly observed in PMF AMS/ACSM studies

	Factor	Relevant ions	External tracers	Reference
POA	HOA (Hydrocarbon like OA)	55, 57, 67, 69, 71, 79, 81, 83	BC, BC_{ff}, EC, NO_x, benzene, hopanes	Crippa et al., 2013, Mohr et al., 2012, Docherty et al., 2011, Ng et al., 2011
	BBOA (biomass burning OA) or SFOA (solid fuel OA)	44, 60, 73	Levoglucosan, non sea salt K^+, Benzo[a]Pyrene, CO, BC_{wb}, PAHs	Lin et al., 2017 ; Bruns et al., 2015; He et al., 2010 Allan et al., 2010; Alfarra et al., 2007,
	COA (Cooking OA)	41, 43,	Fatty acids, cholesterol,	Yin et al., 2015 ; Mohr et al., 2012; He et al. 2010 ; Allan et al., 2010
	CCOA (coal combustion OA)	91, 115, and other high m/z values	PAHs, benzene, toluene, selenium	Dall'Osto et al., 2013; Elser et al., 2016
SOA	MO-OOA (more oxidized oxygenated OA)	44	SO_4^{2-} (site dependent)	Ng et al., 2010; Ng et al., 2011b ; Sun et al., 2017
	LO-OOA (less oxidized oxygenated OA)	43 (as well as 53 and 55, in most cases)	NO_3^- (site / season dependent)	Ng et al., 2010; Ng et al., 2011b, Sun et al., 2017
	IEPOX-SOA (isoprene epoxydiol SOA)	82	SO_4^{2-} (possibly in relation with formation pathways and uptake mechnisms)	Hu et al., 2015; Xu et al., 2015
	Aq-OOA (aqueous oxidized OA)	44, 29	NO_3^-, particle Liquid Water Content (LWC)	Gilardoni et al, 2016; Sun et al., 2016

Different options are suggested to select reference profiles: several studies use single spectra or spectra average from literature. Although hydrocarbon like OA spectra are quite similar across sites, biomass burning and cooking OA are characterized by a larger variability (Reyes-Villegas et al, 2016, Elser et al., 2016). When available, emission spectra and ambient spectra from the same geographical area are employed (Lin et al., 2017, Reyes-Villegas et al., 2016,). Finally, Fröhlich et al., (2015) suggests using factor profiles from unconstrained PMF analysis with higher number of factors to anchor constrained ME-2 analysis. Definition of best practices for the use of the ME-2 solver in constrained PMF analysis is currently being developed in the framework of the COLOSSAL COST Action, soon leading to a guidance document that will be available at https://www.costcolossal.eu/.

As for unconstrained PMF analysis, the factor identification is based on the correlation of profile time series with external tracers, and mass spectra profiles with reference profiles, from ambient, laboratory, and emission measurements. Source identification is further supported by analysis of diurnal profiles and factor elemental composition. Table 13.3 gives examples of factors commonly identified, together with their relevant mass fragments and their corresponding external tracers used for source identification (in the case of primary sources) or further interpretation (in the case of secondary sources). It should be noted that multilinear regression analyses can be conducted when a given external tracer is expected to be associated with several factors. It should also be noted that most datasets will yield a subset of these, depending on the environment and the season.

The family of mass spectrometric techniques for the analysis of aerosols has evolved swiftly in the past decade, reaching a degree of specialisation and diversification that makes it suitable for long-term monitoring and for investigating the wide variety of atmospheric processes. In the framework of the ACTRIS (www.actris.eu), the availability of long-term OA database opens new challenges to receptor model source apportionment which should reflect the seasonality of some factor profiles (e.g., MO-OOA and LO-OOA, Canonaco et al., 2015; Bressi et al., 2016). Moreover, the emergence of off-line AMS analysis is notably allowing for the long-term investigation of primary biogenic OA and other coarse mode OA for (Bozzetti et al., 2017; Daellenbach et al., 2017).

Finally, in some cases, combining different datasets from several measurement systems may allow refining the source apportionment of OA, as well as other/larger PM fraction. Slowik et al. (2010) were the first ones to combine the AMS and PTR-MS data measurements, highlighting the capability of PMF to resolve more OOA factors and improving the interpretations of their sources and photochemical processes. This kind of approach has been explored in other studies, with the combination of AMS or ACSM data with other measurements, such as ambient and thermally denuded OA spectra (TD-PMF-AMS) (Docherty et al., 2011), by merging high resolution mass spectra of organic and inorganic aerosols from AMS measurements (Sun et al., 2012; McGuire et al., 2014; Freney et al., 2018) or combing offline AMS data and organic markers or ^{14}C measurements (Huang et al., 2014; Minguillon et al., 2016; Vlachou et al., 2017). In addition, the combination of PMF-ACSM outputs with inorganic species and black carbon (BC) measurements (Petit et al., 2014), as well as ACSM mass spectra with BC and/or metal concentrations (Jeong et al., 2016; Sofowote et al., 2018), allowed the source apportionment of PM rather than only OA.

References

Alfarra, M.R., Prevot, A.S.H., Szidat, S., Sandradewi, J., Weimer, S., Lanz, V.A., Schreiber, D., Mohr, M. and, Baltensperger, U., 2007. Identification of the Mass Spectral Signature of Organic Aerosols from Wood Burning Emissions, Environmental Science & Technology 41 (16), 5770-5777.

Allan, J. D., Williams, P. I., Morgan, W. T., Martin, C. L., Flynn, M. J., Lee, J., Nemitz, E., Phillips, G. J., Gallagher, M. W., and Coe, H., 2010. Contributions from transport, solid fuel burning and cooking to primary organic aerosols in two UK cities, Atmos. Chem. Phys., 10, 647-668.

Bozzetti, C., El Haddad, I., Salameh, D., et al., 2017. Organic aerosol source apportionment by offline-AMS over a full year in Marseille, Atmos. Chem.Phys., 17, 8247-8268.

Bressi, M., Cavalli, F., Belis, C. A., Putaud, J.-P., Fröhlich, R., Martins dos Santos, S., Petralia, E., Prévôt, A. S. H., Berico, M., Malaguti, A., and Canonaco, F., 2016. Variations in the chemical composition of the submicron aerosol and in the sources of the organic fraction at a regional background site of the Po Valley (Italy), Atmos. Chem. Phys., 16, 12875-12896, doi.org/10.5194/acp-16-12875-2016.

Brown, S. G., Eberly, S., Paatero, P., and Norris, G. A., 2015. Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results, Sci. Total Environ., 518–519, 626–635.

Bruns, E. A., Krapf, M., Orasche, J., Huang, Y., Zimmermann, R., Drinovec, L., Mocnik, G., El-Haddad, I., Slowik, J. G., Dommen, J., Baltensperger, U., and Prévôt, A. S. H., 2015. Characterization of primary and secondary wood combustion products generated under different burner loads, Atmos. Chem. Phys., 15, 2825–2841.

Canonaco F., Crippa M., Slowik J. G., Baltensperger U., Prévôt A. S. H. 2013, SoFi, an Igor based interface for the efficient use of the generalized multilinear engine (ME-2) for source apportionment: application to aerosol mass spectrometer data. *Atmospheric Measurement Techniques*, 6, 6409-6443.

Canonaco, F., Slowik, J. G., Baltensperger, U., and Prévôt, A. S. H., 2015. Seasonal differences in oxygenated organic aerosol composition: implications for emissions sources and factor analysis, Atmos. Chem. Phys., 15, 6993-7002, https://doi.org/10.5194/acp-15-6993-2015.

Crippa, M., Canonaco, F., Slowik, J. G., et al., 2013. Primary and secondary organic aerosol origin by combined gas-particle phase source apportionment, Atmos. Chem. Phys., 13, 8411-8426, 10.5194/acp-13-8411-2013.

Crippa, M., Canonaco, F., Lanz, V. A., et al., 2014. Organic aerosol components derived from 25 AMS data sets across Europe using a consistent ME-2 based source apportionment approach, Atmos. Chem. Phys., 14, 6159-6179.

Daellenbach, K. R., Stefenelli, G., Bozzetti, C., et al., 2017. Long-term chemical analysis and organic aerosol source apportionment at 9 sites in Central Europe: Source identification and uncertainty assessment, Atmos. Chem. Phys. Discuss., doi: 10.5194/acp-2017-124.

Dall'Osto, M., Ovadnevaite, J., Ceburnis, D., Martin, D., Healy, R. M., O'Connor, I. P., Kourtchev, I., Sodeau, J. R., Wenger, J. C., and O'Dowd, C., 2013. Characterization of urban aerosol in Cork city (Ireland) using aerosol mass spectrometry, Atmos. Chem. Phys., 13, 4997-5015.

DeCarlo, P.F., J.R. Kimmel, A. Trimborn, et al., 2006. A Field-Deployable High-Resolution Time-of-Flight Aerosol Mass Spectrometer, Anal. Chem., 78, 8281-8289.

Docherty, K. S., Aiken, A. C., Huffman, J. A., et al., 2011. The 2005 Study of Organic Aerosols at Riverside (SOAR-1): instrumental intercomparisons and fine particle composition, Atmos. Chem. Phys., 11, 12387-12420, 10.5194/acp-11-12387-2011.

EMEP, 2016, Air pollution trends in the EMEP region between 1990 and 2012, Colette et al., European Monitoring and Evaluation Programme EMEP/CCC-Report 1/2016, Norway.

Elser, M., HuangR. J., Wolf, R., et al., 2016. New insights into PM2:5 chemical composition and sources in two major cities in China during extreme haze events using aerosol mass spectrometry, Atmos. Chem. Phys., 16, 3207-3225.

Freney, E., Sellegri, K., Chrit, M. et al., 2018. Aerosol composition and the contribution of SOA formation over Mediterrenean forests, Atmospheric Chemistry and Physics, 18(10), 7041-7056.

Fröhlich, R., Crenn, V., Setyan, A., et al., 2015. ACTRIS ACSM intercomparison – Part 2: Intercomparison of ME-2 organic source apportionment results from 15 individual, co-located aerosol mass spectrometers, Atmos, Meas. Tech., 8, 2555-2576.

Gilardoni, S., Massoli, P., Paglione, M., et al., 2016. Direct observation of aqueous secondary organic aerosol from biomass-burning emissions, Proceedings of the National Academy of Sciences, vol. 113, issue 36, pp.10013-10018.

Grigas, T., Ovadnevaite, J., Ceburnis, D., Moran, E., McGovern, F.M., Jennings, S.G., O'Dowd C., 2017. Sophisticated cleanaor strategies required to mitigate against particulate organic pollution, Sientific Reports, 7 (44737), doi: 10.1038/srep44737.

He, L.-Y., Lin, Y., Huang, X.-F., Guo, S., Xue, L., Su, Q., Hu, M., Luan, S.-J., and Zhang, Y.-H., 2010. Characterization of highresolution aerosol mass spectra of primary organic aerosol emissions from Chinese cooking and biomass burning, Atmos. Chem. Phys., 10, 11535–11543.

Hu, W. W., Campuzano-Jost, P., Palm, B. B., et al., 2015. Characterization of a real-time tracer for isoprene epoxydiols-derived secondary organic aerosol (IEPOX-SOA) from aerosol mass spectrometer measurements, Atmos. Chem. Phys., 15, 11807-11833.

Huang, R.-J., Zhang, Y., Bozzetti, et al., 2014. High secondary aerosol contribution to part iculate pollution during haze events in China, Nature, 514, 218-222, 10.1038/nature13774

Jayne, J.T., D.C. Leard, X. Zhang, P. Davidovits, K.A. Smith, C.E. Kolb, D.R. Worsnop, 2000. Development of an Aerosol Mass Spectrometer for Size and Composition. Analysis of Submicron Particles. *Aerosol Science Technology* 33 49-70.

Jeong, C.H., Wang, J.M., Evans, G.J., 2016. Source apportionment of urban particulate matter using hourly resolved trace metals, organics, and inorganic aerosol components, Atmospheric Chemistry and Physics Discussions, doi.org/10.5194/acp-2016-189.

Jimenez, J.L., M.R. Canagaratna, N.M. Donahue, et al., 2009. Evolution of Organic Aerosols in the Atmosphere. *Science* 326, 1525-1529.

Lanz, V. A., M. R. Alfarra, U. Baltensperger, B. Buchmann, C. Hueglin, A. S. H. Prévôt, 2007. Source apportionment of submicron organic aerosols at an urban site by factor analytical modelling of aerosol mass spectra. *Atmospheric Chemistry and Physics* 7:1503–1522.

Lanz, V.A., M.R. Alfarra, U. Baltensperger, B. Buchmann, C. Hueglin, S. Szidat, M.N. Wehrli, L. Wacker, S. Weimer, A. Caseiro, H. Puxbaum, A.S.H. Prevot, 2008. Source Attribution of

Submicron Organic Aerosols During Wintertime Inversions by Advanced Factor Analysis of Aerosol Mass Spectra. *Environmental Science & Technology* 42(1), 214-220.

Lin, C., Ceburnis, D., Hellebust, S., Buckley, P., Wenger, J., Canonaco, F., Pre´voˆt, A.S.H., Huang, R.J., O'Dowd, C., and Ovadnevaite, J., 2017. Characterization of Primary Organic Aerosol from Domestic Wood, Peat, and Coal Burning in Ireland, Environ. Sci. Technol., 51, 10624−10632.

McGuire, M., Chang, R.-W., Slowik, J., Jeong, C.-H., Healy, R., Lu, G., Mihele, C., Abbatt, J., Brook, J., and Evans, G., 2014. Enhancing non-refractory aerosol apportionment from an urban industrial site through receptor modeling of complete high time-resolution aerosol mass spectra, Atmos. Chem. Phys., 14, 8017-8042, 2014

Middlebrook A.M., R. Bahreini, J.L. Jimenez, M.R. Canagaratna, 2012. Evaluation of Composition - Dependent Collection Efficiencies for the Aerodyne Aerosol Mass Spectrometer using Field Data. *Aerosol Science and Technology* 46, 258–271.

Minguillon, M.C., Perez, N., Marchand, N., et al., 2016. Secondary organic aerosol origins in an urban environment: influence of biogenic and fuel combustion precursors, Faraday Discussions, 189, 337-389, doi: 10.1039/c5fd00182.

Mohr, C., DeCarlo, P. F., Heringa, M. F., et al., 2012. Identification and quantification of organic aerosol from cooking and other sources in Barcelona using aerosol mass spectrometer data, Atmos. Chem. Phys., 12, 1649-1665, doi: 10.5194/acp-12-1649-2012.

Ng, N. L., Canagaratna, M. R., Zhang, Q., et al., 2010. Organic aerosol components observed in Northern Hemispheric datasets from Aerosol Mass Spectrometry, Atmos. Chem. Phys., 10.

Ng, N.L., S.C. Herndon, A. Trimborn, et al., 2011. An Aerosol Chemical Speciation Monitor (ACSM) for Routine Monitoring of the Composition and Mass Concentrations of Ambient Aerosol. *Aerosol Science and Technology* 45(7), 780-794.

Ng, N. L., Canagaratna, M. R., Jimenez, J. L., Chhabra, P. S., Seinfeld, J. H., and Worsnop, D. R., 2011b. Changes in organic aerosol composition with aging inferred from aerosol mass spectra, Atmos. Chem. Phys., 11, 6465-6474.

Norris G., Vedantham R., K. Wade, S. Brown, J. Prouty, C. Foley, 2008. EPA Positive Matrix Factorization (PMF 3.0) Fundamentals & User Guide. EPA 600/R-08/108.

Nozière, B., Kalberer, M., Claeys, M., et al., 2015. The molecular identification of organic compounds in the atmosphere: start of the art and challenges, Chemical Reviews, 115 (10), 3919-3983.

Onasch, T., Trimborn, A., Fortner, E.C., Jayne, J.T., Kok, G.L., Williams, L.R., Davidovits, P., Worsnop, D.R., 2012. Soot particle aerosol mass spectrometer: development, validation, and initial application, Aerosol Science and Technology, 10, 804-817.

Paatero P., 2004. User's guide for positive matrix factorization programs PMF2 and PMF3, Parts 1 and 2. University of Helsinki, Helsinki, Finland.

Petit, J. E., Favez, O., Sciare, J., Canonaco, F., Croteau, P., Močnik, G., Jayne, J., Worsnop, D., and Leoz-Garziandia, E., 2014. Submicron aerosol source apportionment of wintertime pollution in Paris, France by double positive matrix factorization (PMF2) using an aerosol chemical speciation monitor (ACSM) and a multi-wavelength Aethalometer, Atmos. Chem. Phys., 14, 13773-13787, 10.5194/acp-14-13773-2014.

Sofowote, U. M., Healy, R. M., Su, Y., Debosz, J., Noble, M., Munoz, A., Jeong, C. H., Wang, J. M., Hilker, N., Evans, G. J., and Hopke, P. K., 2018. Understanding the PM2.5 imbalance between a far and near-road location: Results of high temporal frequency source apportionment and parameterization of black carbon, Atmos. Environ., 173, 277-288, doi.org/10.1016/j.atmosenv.2017.10.063.

Reyes-Villegas, E., Green, D.C., Priestman, M., Canonaco, F., Coe, H., Prévôt, A.S.H., Allan, J.D., 2016. Organic aerosol source apportionment in London 2013 with ME-2: exploring the solution space with annual and seasonal analysis, Atmos. Chem. Phys, 16, 15545-15559.

Slowik, J.G., Vlasenko, A., McGuire, M., Evans, G.J., and Abbatt, J.P.D., 2010. Simultaneous factor analysis of organic particle and gas mass spectra: AMS and PTR-MS measurements at an urban site. Atmos. Chem. Phys., 10, 1969–1988.

Srivastava D., Tomaz S., Favez O., Lanzafame G.M., Golly B., Besombes J.-L., Alleman L.Y., Jaffrezo J.-L., Jacob E., Perraudin E., Villenave E., Albinet A., 2018a. Speciation of organic fraction does matter for source apportionment. Part 1: a one-year campaign in Grenoble (France). *Science of the Total Environment*, 624, 1598-1611.

Srivastava D., Favez O., Bonnaire N., Lucarelli F., Haeffelin M., Perraudin E., Gros V., Villenave E., Albinet A., 2018b. Speciation of organic fraction does matter for source apportionment. Part 2: intensive short-term campaign in the Paris area (France). *Science of the Total Environment*, 634, 267-278.

Sun, Y. L., Zhang, Q., Anastasio, C., and Sun, J., 2010. Insights into secondary organic aerosol formed via aqueous-phase reactions of phenolic compounds based on high resolution mass spectrometry, Atmos. Chem. Phys., 10, 4809-4822, 2010.

Sun, Y., Zhang, Q., Schwab, J., Yang, T., Ng, N., Demerjian, K., 2012. Factor analysis of combined organic and inorganic aerosol mass spectra from high resolution aerosol mass spectrometer measurements, Atmos. Chem. Phys., 12, 8537-8551.

Sun, Y., Xu, W., Zhang, Q., Jiang, Q., Canonaco, F., Prévôt, A.S.H., Fu, P., Li, J., Jayne, J., Worsnop, D.R. and Zifa Wang, Z., 2017. Source apportionment of organic aerosol from two-year highly time-resolved measurements by an aerosol chemical speciation monitor in Beijing, China. Atmos. Chem. Phys. Discuss., doi: 10.5194/acp-2017-1195.

Ulbrich I.M., M.R. Canagaratna, Q. Zhang, D.R. Worsnop, J.L. Jimenez., 2009. Interpretation of Organic Components from Positive Matrix Factorization of Aerosol Mass Spectrometric Data. *Atmospheric Chemistry and Physics* 9(9), 2891-2918.

Vlachou, A., Daellenbach, K. R., Bozzetti, C., Chazeau, B., Salazar, G. A., Szidat, S., Jaffrezo, J. L., Hueglin, C., Baltensperger, U., El Haddad, I., and Prévôt, A. S. H., 2017. Advanced source apportionment of carbonaceous aerosols by coupling offline AMS and radiocarbon size segregated measurements over a nearly two-year period, Atmos. Chem. Phys. Discuss., 2017, 1-25, 10.5194/acp-2017-1102.

Xu, L., Suresh, S., Guo, H., Weber, R. J., and Ng, N. L., 2015. Aerosol characterization over the southeastern United States using high-resolution aerosol mass spectrometry: spatial and seasonal variation of aerosol composition and sources with a focus on organic nitrates, Atmos. Chem. Phys., 15, 7307-7336.

Zhang, Q., M.R. Alfarra, D.R. Worsnop, J.D. Allan, H. Coe, M.R. Canagaratna, J.L. Jimenez, 2005. Deconvolution and Quantification of Hydrocarbon-like and Oxygenated Organic Aerosols Based on Aerosol Mass Spectrometry. *Environmental Science & Technology* 39, 4938-4952.

Zhang, Q., Jimenez, J. L., Canagaratna, M. R., Ulbrich, I. M., Ng, N. L., Worsnop, D. R., Sun, Y., 2011. Understanding atmospheric organic aerosols via factor analysis of aerosol mass spectrometry: a review. *Analytical and Bioanalytical Chemistry* 401(10), 3045–3067.

Zhang, Y., Tang, L., Croteau, P.L., et al., 2017. Field characterization of the $PM_{2.5}$ Aerosol Chemical Speciation Monitor: insights into the composition, sources and processes of fine particles in Eastern China, *Atmospheric Chemistry and Physics*, 17, 14501-14517, doi: 10.5194/acp-17-14501-2017.

14. AETHALOMETER MODELS

This section has been revised in collaboration with working group 3 (refractory fine aerosol fraction) of the COLOSSAL COST action (www.costcolossal.eu/wg3/).

The Aaethalometer instrument was originally developed to quantify light absorption by carbonaceous aerosols in general and black carbon (EC) in particular (Hansen et al., 1984). BC is known as the predominant light-absorbing aerosol species at visible wavelengths. Furthermore, several studies pointed out that organic carbon significantly absorbs light in the ultraviolet wavelengths and less significantly going into the visible (e.g. Kirchstetter et al., 2004). This fraction, known as brown carbon for its light brownish colour, includes tar materials from smouldering fires or solid fuel combustion, pyrolysis products from biomass burning and humic-like substances from soil or biogenic emissions (Feng et al., 2013).

Light absorption by aerosols is usually parameterised as proportional to $\lambda^{-\alpha}$, where λ is the light wavelength and α represents the Ångstrom absorption exponent. While the spectral dependence of elemental carbon light absorption is low (α ~ 1, Bond and Bergstrom, 2006), brown carbon exhibits a much higher Ångstrom absorption exponent (up to 7, see e.g. Hoffer et al., 2006). Based on these differences in optical properties, a growing number of studies recently used multi-wavelength Aethalometers to detect and/or apportion wood burning carbonaceous aerosols in ambient air (e.g. Jeong et al., 2004; Sandradewi et al., 2008a, 2008b; Yang et al., 2009; Favez et al., 2009). In particular, Sandradewi et al. (2008b) introduced a methodology to quantitatively assess the contribution of wood burning and fossil fuel to the absorption coefficients (b_{abs}) measured by a multi-wavelength Aethalometer. Assuming that the source specific values of Ångstrom absorption exponent for pure fossil fuel (α_{ff}) and biomass burning (α_{wb}) are known, the wavelength dependence of these two components of absorbing aerosol can be expressed as:

$$\frac{b_{abs}(\lambda_{UV})_{ff}}{b_{abs}(\lambda_{IR})_{ff}} = (\frac{\lambda_{UV}}{\lambda_{IR}})^{-a_{ff}} \qquad (14.1)$$

$$\frac{b_{abs}(\lambda_{UV})_{wb}}{b_{abs}(\lambda_{IR})_{wb}} = (\frac{\lambda_{UV}}{\lambda_{IR}})^{-a_{wb}} \qquad (14.2)$$

where $b_{abs}(\lambda_{UV})_{wb}$ & $b_{abs}(\lambda_{IR})_{wb}$ and $b_{abs}(\lambda_{UV})_{ff}$ & $b_{abs}(\lambda_{IR})_{ff}$ are absorption coefficients at the UV and IR wavelengths that are related to wood burning (wb) and fossil fuel combustion (ff), respectively. The overall absorption measured at a given wavelength ($b_{abs}(\lambda)$) is then assumed to be the sum of the contributions from these two sources (Figure 14.1):

$$b_{abs}(\lambda) = b_{abs}(\lambda)_{ff} + b_{abs}(\lambda)_{wb} \qquad (14.3)$$

By solving equations (14.1)-(14.3) using a wavelength pair, $b_{abs}(\lambda_{UV})_{wb}$, $b_{abs}(\lambda_{IR})_{wb}$, $b_{abs}(\lambda_{UV})_{ff}$, and $b_{abs}(\lambda_{IR})_{ff}$ can be calculated through the following formulas (Kalogridis et al., 2017):

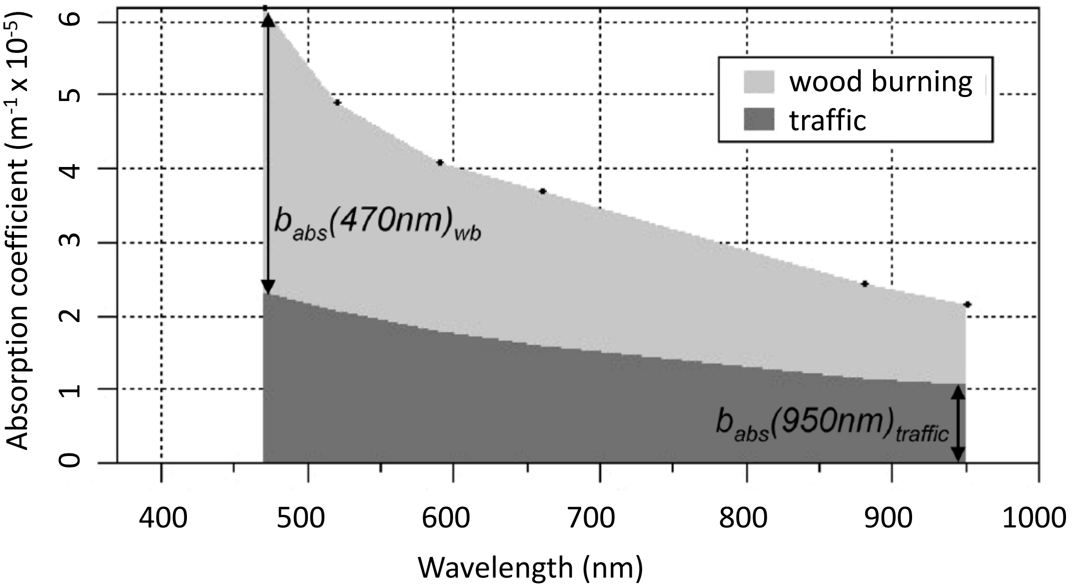
$$b_{abs}(\lambda_{UV})_{wb} = \frac{1}{1-(\frac{\lambda_{UV}}{\lambda_{IR}})^{-a_{ff}} \cdot (\frac{\lambda_{UV}}{\lambda_{IR}})^{a_{wb}}} \cdot \left[b_{abs}(\lambda_{UV}) - \left(\frac{\lambda_{UV}}{\lambda_{IR}}\right)^{-a_{ff}} \cdot b_{abs}(\lambda_{IR})\right] \quad (14.4)$$

$$b_{abs}(\lambda_{IR})_{wb} = (\frac{\lambda_{UV}}{\lambda_{IR}})^{a_{wb}} \cdot b_{abs}(\lambda_{UV})_{wb} \quad (14.5)$$

$$b_{abs}(\lambda_{UV})_{ff} = b_{abs}(\lambda_{UV}) - b_{abs}(\lambda_{UV})_{wb} \quad (14.6)$$

$$b_{abs}(\lambda_{IR})_{ff} = b_{abs}(\lambda_{IR}) - b_{abs}(\lambda_{IR})_{wb} \quad (14.7)$$

Figure 14.1 Graphical representation of the apportionment of light absorption between wood burning and traffic sources (from Sandradewi et al., 2008b).



This methodology has then been used to assess the fossil fuel and wood burning contributions to BC, assuming a single mass absorption cross-section for both sources and performing the conversion of the absorption coefficient to mass equivalent BC concentration in the infrared part of the spectrum (IR), where the absorption of BC dominates relative to organic compounds (Favez et al., 2010; Zotter et al., 2017). This model has also been directly implemented in the new Aethalometer instruments (i.e., AE33), as described in Drinovec et al. (2015) and has also been used in a source-component model by Massabò et al. (2015) and Bernardoni et al. (2017).

The choice of the wavelength pair used to calculate the contributions is not arbitrary: the 2 wavelengths should cover a large interval and should be source specific. The longer wavelength in the IR is usually 950 nm. Due to possible variable contribution of secondary organic aerosol to absorption at the extreme end of the measurement spectrum and/or sampling artefacts, the choice is usually made to use 470nm as the shorter wavelength (Sandradewi et al., 2008b).

The estimated contributions are very sensitive to the selection of the α_{ff} and α_{wb}. For this reason, users usually perform (and give results of) sensitivity tests with wide ranges for these values (see e.g. Sciare et al., 2011), and suggest that the results of these sensitivity tests be considered as the total uncertainties of the model outputs. The number of negative values of source specific BC concentrations should be minimal for the optimal selection of the α_{ff} and α_{wb} pair (there will still be some negative concentrations due to noise). The use of markers typical for biomass burning (such as levoglucosan, potassium and OC/EC ratios) and for fossil fuel (such as NOx) may provide further support of the validity of the estimated contributions from wood burning and fossil fuel (Diapouli et al., 2017, Zotter et al., 2017). Recent studies have also proposed to directly combine levoglucosan and aethalometer datasets in order to decrease the uncertainty associated with the model (Fuller et al., 2014; Diapouli et al., 2017; Titos et al., 2017). Based on equation (14.4), it is evident that the selection of α_{wb} only affects the magnitude of the estimated $b_{abs}(\lambda)_{wb}$, while α_{ff} has an impact on its temporal variability as well. By selecting a typical value α_{wb}, $b_{abs}(\lambda)_{wb}$ may be calculated for a range of α_{ff} values. The estimated $b_{abs}(\lambda)_{wb}$ are then assessed based on levoglucosan data. Since levoglucosan is only emitted during wood burning processes, no levoglucosan is expected to be measured during periods with zero contribution from wood burning. Therefore, the regression of levoglucosan concentrations over the estimated $b_{abs}(\lambda)_{wb}$ is expected to yield a zero intercept (assuming levoglucosan is stable in the atmosphere). This analysis allows for the identification of an optimal α_{ff}. The identification of a suitable value for the Ångström exponent for wood burning (α_{wb}) is a more complex procedure, given also that α_{wb} may be affected by the type of wood burnt and the combustion regime (the stove and its operation) and by the generation of absorbing secondary organic aerosol. This implies that representative α_{wb} values for summer (when wood burning mostly relates to forest fires and/or agricultural fires) and winter conditions (when residential heating may contribute significantly to BB aerosol) can differ. Recently, Zotter et al. (2017) proposed the use of the ratio ECfossil fuel/EC, derived from ^{14}C measurements to define the best combination of both α_{ff} and α_{wb} values, and determining the uncertainties using probability density functions. The same analysis has also shown that not all combinations of the α_{ff} and α_{wb} values, while equally “good” according to the criteria described above, are not conformant with the

^{14}C results, and that the mass absorption cross-sections of BC_{ff} and BC_{wb} determined in the IR are not different (Zotter et al., 2017).

Most commonly used values for the Ångstrom absorption exponent are in the ranges 0.9-1.1 and 1.7-2.2 for α_{ff} and α_{wb}, respectively.

It should be noted that BC_{ff} comprises traffic emissions as well as carbonaceous aerosols originating from fuel oil and natural gas combustion but excludes coal-burning organic aerosol. Indeed, the latter was shown to significantly absorb light at near UV wavelengths (e.g. Yang et al., 2009) and may thus interfere with $b_{abs,wb,470nm}$. Another limitation of these approaches might be the presence of mineral dust particles (notably containing iron oxides), that also absorb light at near UV wavelengths (Fialho et al., 2006) and should thus be considered carefully.

The light absorbing properties of carbonaceous aerosols, when measured by filter photometers, need a careful analysis of the interaction of the particle sample and the filter matrix due to the non-linear relationship between the mass loading of the sample spot in the instrument and the optical measurement (Weingartner et al., 2003; Collaud Coen et al., 2010; Drinovec et al., 2015; Drinovec et al., 2017). Different parametrizations of this interaction usually separate the enhancement of probability, that a photon is absorbed by the sample in the light scattering filter matrix, and the loading effects due to the accumulation of the sample. The multiple scattering of the filter matrix may to a degree depend on the wavelength (Segura et al., 2010), but this dependence has been in other cases insignificant (Corbin et al., 2018). On the other hand, the non-linear loading effects depend heavily on the wavelength and the coating of BC cores (Drinovec et al., 2017). The loading effects strongly influence the determination of aerosol optical absorption and hence the source apportionment, based on the dependence of absorption on the wavelength. The loading effects need to be compensated off-line or on-line with one of the compensation algorithms (Weingartner et al., 2003; Collaud Coen et al., 2010; Drinovec et al., 2015) and the compensation needs to be verified – the easiest self-contained verification method is plotting the absorption coefficient (at each relevant wavelength) as a function of the attenuation of light through the sample loaded filter (as the parameter describing the loading of the spot with the absorbing sample; Drinovec et al., 2015). The absorption coefficient should depend only on the time derivative of attenuation, not the attenuation itself in the absence of non-linearities. This approach makes no assumptions and any dependence of the absorption coefficient on attenuation indicates a suboptimal compensation of the loading effect. The remaining slope can be compensated by postprocessing the data and the "goodness" of the compensation at the wavelengths used in the source apportionment model can be used to assess the model uncertainty in addition to the choice of the source specific Ångstrom absorption exponents.

As initially proposed by Sandradewi et al. (2008c), total carbonaceous material (CM_{total}) could also be primarily considered as the sum of brown-carbon-containing carbonaceous material (i.e. CM_{wb} here), non-brown-carbon-containing carbonaceous material originating from fossil fuel combustion (CM_{ff}), and non-absorbing organic aerosol from other sources (CM_{other}), as follows:

$$CM_{total} = CM_{ff} + CM_{wb} + CM_{other} = C_1 \times b_{abs,ff,950nm} + C_2 \times b_{abs,wb,470nm} + C_3 \quad (14.8)$$

where $b_{abs,ff,950nm}$ represents the absorption coefficient of CM_{ff} at 950nm, $b_{abs,wb,470nm}$ represents the absorption coefficient of CM_{wb} at 470 nm, C_1 and C_2 relate the light absorption to the particulate mass of both sources (parameters are inverse mass absorption cross-sections), and C_3 corresponds to the amount of other organic aerosol (assumed here to have a negligible light absorption capacity: to a degree, this is then the model residuum). This approach is still currently subject of debate and might need further evaluation as it has been shown to be barely applicable for long-term datasets at some sites (Herich et al., 2011). Its main disadvantage is the assumption of constant CM composition by using constant mass absorption cross-sections (i.e. parameters C_1 and C_2), disregarding any diurnal variation of the secondary organic aerosol production. Care is however recommended in the interpretation of data from the Aethalometer because of the potentially non-specific nature of the absorption at lower wavelengths (Harrison et al., 2013).

In another methodology, Wang et al. (2012a and b) included Delta-C (the difference in Aethalometer BC measured at 370 nm and that measured at 880 nm) in their PMF analyses of data from Rochester, NY. With the typical collection of elements, ions, organic carbon and elemental carbon, the addition of DeltaC provided a clear resolution of biomass burning from traffic sources (Wang et al., 2012a). In an analysis including molecular markers, the Delta-C was observed primarily in the biomass burning factor along with levoglucosan (Wang et al., 2012b).

It should also be noted that an EMEP/ACTRIS/COLOSSAL initiative has been launched recently to extensively apply positive matrix factorization (PMF) analyses to multi-wavelength Aethalometer datasets obtained all over Europe during intensive measurement campaigns in winter 2017-2018 (EMEP status report, 2018).

The considerable increase in measurements carried out using Aethalometers associated with the interest in the potential impacts of light absorpbing carbonaceous aerosol on climate and on health, makes this technique an interesting resource for improving the understanding of aerosol sources, with particular reference to biomass burning. Due to the methodology used by Aethalometers (filter-based measurement), absorption coefficients directly obtained from these instruments are affected by various sampling and analytical artefacts (mostly referred to as multiple scattering and loading effects) which need to be carefully taken into account prior to any data treatment (Collaud Coen et al., 2010; Drinovec et al., 2015; and references therein).

References

Bond, T.C., Bergstrom, R.W., 2006. Light absorption by carbonaceous particles: an investigative review. *Aerosol Science Technology* 40, 27-67

Bernardoni V., Pileci R.E., Caponi L., Massabò D., 2017. The Multi-Wavelength Absorption Analyzer (MWAA) Model as a Tool for Source and Component Apportionment Based on Aerosol Absorption Properties: Application to Samples Collected in Different Environments. Atmosphere, 8, 218.

Collaud Coen, M., Weingartner, E., Apituley, A., Ceburnis, D., Fierz-Schmidhauser, R., Flentje, H., Henzing J.S., Jennings, S.G., Moerman, M., Petzold, A., Schmid, O. Baltensperger, U., 2010. Minimizing light absorption measurement artifacts of the Aethalometer: evaluation of five correction algorithms. *Atmospheric Measurement Techniques* 3, 457–474.

Corbin, J. C., Pieber, S. M., Czech, H., Zanatta, M., Jakobi, G., Massabò, D., Orasche J., El Haddad, I., Mensah, A. A., Stengel, B., Drinovec, L., Močnik, G., Zimmermann, R., Prévôt, A. S. H., Gysel, M., 2018. Brown and black carbon emitted by a marine engine operated on heavy fuel oil and distillate fuels: Optical properties, size distributions, and emission factors. Journal of Geophysical Research: Atmospheres, 123, 6175–6195.

Diapouli, E., Kalogridis, A.-C., Markantonaki, C., Vratolis, S., Fetfatzis, P., Colombi, C., Eleftheriadis, K., 2017. Annual variability of black carbon concentrations originating from biomass and fossil fuel combustion for the suburban aerosol in Athens, Greece. Atmosphere (accepted for publication)

Drinovec, L., Močnik, G., Zotter, P., Prévôt, A. S. H., Ruckstuhl, C., Coz, E., Rupakheti, M., Sciare, J., Müller, T., Wiedensohler, A., and Hansen, A. D. A., 2015. The "dual-spot" Aethalometer: an improved measurement of aerosol black carbon with real-time loading compensation. *Atmos. Meas. Tech.* **8**, 1965-1979.

Drinovec, L., Gregorič, A., Zotter, P., Wolf, R., Bruns, E. A., Prévôt, A. S. H., Petit, J.-E., Favez, O., Sciare, J., Arnold, I. J., Chakrabarty, R. K., Moosmüller, H., Filep, A., and Močnik, G.,2017. The filter-loading effect by ambient aerosols in filter absorption photometers depends on the coating of the sampled particles, Atmos. Meas. Tech., 10, 1043-1059.

EMEP status report, 2018. http://emep.int/publ/reports/2018/EMEP_Status_Report_1_2018.pdf.

Favez, O., Cachier, H., Sciare, J., Sarda-Estève, R., Martinon, L., 2009. Evidence for a significant contribution of wood burning aerosols to PM2.5 during the winter season in Paris, France. *Atmospheric Environment* 43, 3640-3644.

Favez, O., El-Haddad, I., Piot, C., Boreave, A., Abidi, E., Marchand, N., Jaffrezo, J.-L., Besombes, J.-L., Personnaz, M.-B., Sciare, J., Wortham, H. 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). *Atmospheric Physics and Chemistry* 10, 5295-5314.

Feng Y., Ramanathan V., Kotamarthi V. R., 2013. Brown carbon: a significant atmospheric absorber of solar radiation? *Atmospheric Chemistry and Physics*, 13, 8607–8621.

Fialho, P., Freitas, M.C., Barata, F., Viera, B., Hansen, A.D.A., Honrath, R.E., 2006. The Aethalometer calibration and determination of iron concentration in dust aerosols, *Journal of Aerosol Science* 37, 1497-1506.

Fuller, G.W., Tremper, A.H., Baker, T.D., Yttri, K.E., Butterfield, D., 2014. Contribution of wood burning to PM10 in London. Atmospheric Environment 87, 87-94

Hansen, A.D.A., Rosen, H., Novakov, T., 1984. The Aethalometer – An instrument for the real-time measurement of optical absorption by aerosol particles. *Science of the Total Environment* 36, 191-196.

Harrison R.M., Beddows, D.C.S., Jones, A.M., Calvo, A., Alves, C. and Pio, C.A., 2013. An evaluation of some issues regarding the use of Aethalometers to measure woodsmoke concentrations. Atmospheric Environment 80, 540-548.

Herich, H., Hueglin, C., and Buchmann, B., 2011. A 2.5 year's source apportionment study of black carbon from wood burning and fossil fuel combustion at urban and rural sites in Switzerland, Atmos. Meas. Tech., 4, 1409–1420, doi:10.5194/amt-4-1409-2011.

Hoffer, A., Gelencsér, A., Guyon, P., Kiss, G., Schmid, O., Frank, G.P., Artaxo, P., Andreae, M.O., 2006. Optical properties of humic-like substances (HULIS) in biomass-burning aerosols. *Atmospheric Chemistry and Physics* 6, 3563-3570.

Jeong, C.-H., Hopke, P.K., Kim, E., Lee, D.-W., 2004. The comparison between thermal-optical transmittance elemental carbon and Aethalometer black carbon measured at multiple monitoring sites. *Atmospheric Environment* 38, 5193-5204.

Kalogridis, A.-C., Vratolis, S., Liakakou, E., Gerasopoulos, E., Mihalopoulos, N., and Eleftheriadis, K., 2018. Assessment of wood burning versus fossil fuel contribution to wintertime black carbon and carbon monoxide concentrations in Athens, Greece, *Atmospheric Chemistry and Physics*, 18, 10219-10236.

Kirchstetter, T.W., Novakok, T., Hobbs, P.V., 2004. Evidence that the spectral dependence of light absorption by aerosols is affected by organic carbon. *Journal of. Geophysical. Research* 109, D21208.

Massabò D., Caponi L., Bernardoni V., Bove M.C., Brotto P., Calzolai G., Cassola F., Chiari M., Fedi M., Fermo P., Giannoni M., Lucarelli F., Nava S., Piazzalunga A., Valli G., Vecchi R., Prati P., 2015. Multi-wavelength optical determination of Black and Brown Carbon in atmospheric aerosols. Atmospheric Environment 108, 1-12.

Sandradewi, J., Prévôt, A.S.H., Weingartner, E., Schmidhauser, R., Gysel M., Baltensperger, U., 2008a. A study of wood burning and traffic aerosols in an Alpine valley using a multi-wavelength, Aethalometer. *Atmospheric Environment* 42, 101-112.

Sandradewi, J., Prévôt, A.S.H., Szidat, S., Perron, N., Alfarra, M.R., Lanz, V.A., Weingartner, E., and Baltensperger, U., 2008b. Using aerosol light absorption measurements for the quantitative determination of wood burning and traffic emission contributions to particulate matter. *Environmental Science and Technology* 42, 3316-3323.

Sandradewi, J., Prévôt, A. S. H., Alfarra, M. R., Szidat, S., Wehrli, M. N., Ruff, M., Weimer, S., Lanz, V. A., Weingartner, E., Perron, N., Caseiro, A., Kasper-Giebl, A., Puxbaum, H., Wacker, L., and Baltensperger, U., 2008c. Comparison of several wood smoke markers and source apportionment methods for wood burning particulate mass, Atmos. Phys. Chem. Discuss., 8, 8091-8118.

Segura, S., Estellés, V., Titos, G., Lyamani, H., Utrillas, M. P., Zotter, P., Prévôt, A. S. H., Močnik, G., Alados-Arboledas, L., and Martínez-Lozano, J. A., 2014. Determination and analysis of in situ spectral aerosol optical properties by a multi-instrumental approach, Atmos. Meas. Tech., 7, 2373-2387.

Sciare, J., d'Argouges, O., Sarda-Esteve, R., Gaimoz, C., Dolgorouky, C., Bonnaire, N., Favez, O., Bonsang, B., and Gros, V., 2011. Large contribution of water-insoluble secondary organic aerosols in the Paris region (France). *Journal of Geophysical Research* 116, D22203.

Titos, G., del Águila, A., Cazorla, A., Lyamani, H., Casquero-Vera, J.A., Colombi, C., Cuccia, E., Gianelle, V., Močnik, G., Alastuey, A., Olmo F.J., Alados-Arboledas, L., 2017. Spatial and temporal variability of carbonaceous aerosols: Assessing the impact of biomass burning in the urban environment. Science of the Total Environment 578, 613–625.

Weingartner, E., Saathoff, H., Schnaiter, M., Streit, N., Bitnar, B., Baltensperger, U., 2003. Absorption of light by soot particles: determination of the absorption coefficient by means of Aethalometers. *Journal of Aerosol Science* 34, 1445-1463.

Yang, M., Howell, S.G., Zhuang, J., Huebert, B.J., 2009. Attribution of aerosol light absorption to black carbon, brown carbon and dust in China - interpretations of atmospheric measurements during EAST-AIRE. *Atmospheric Chemistry and Physics* 9, 2035-2050.

Wang, Y., Hopke, P.K., Xia, X., Rattigan, O., Chalupa, D.C., Utell, M.J., 2012a. Multiple year black carbon measurements and source apportionment using Delta-C in Rochester, NY. *Journal of Air and Waste Management Association* 62, 880-887.

Wang, Y., Hopke, P.K., Xia, X., Rattigan, O., Chalupa, D.C., Utell, M.J., 2012b. Source apportionment of airborne particulate matter using inorganic and organic species as tracers. *Atmospheric Environment* 55: 525-532.

Zotter, P., Herich, H., Gysel, M., El-Haddad, I., Zhang, Y., Močnik, G., Hüglin, C., Baltensperger, U., Szidat, S., Prévôt, A.S.H., 2017. Evaluation of the absorption Ångström exponents for traffic and wood burning in the Aethalometer-based source apportionment using radiocarbon measurements of ambient aerosol. Atmospheric Chemistry and Physics 17, 4229–4249.

15. RADIOCARBON AND TRACER ANALYSIS FOR THE APPORTIONMENT OF THE PM CARBONACEOUS FRACTION

The carbonaceous fraction is one of the main components of particulate matter (PM). The study of carbonaceous aerosol is important because of its adverse effects on health (Highwood & Kinnersley, 2006; Mauderly & Chow, 2008), air quality (Putaud et al., 2004; Turpin & Huntzicker, 1995; Vecchi et al., 2008; among others), visibility (Watson, 2002), cultural heritage (Bonazza et al., 2005), and the Earth's radiation balance (IPCC, 2007).

Total carbon (TC) in atmospheric aerosols consists of two main fractions: elemental (EC) and organic (OC) carbon. EC is produced by the incomplete combustion of fossil and biomass fuels in an oxygen-poor environment (Chow et al., 2001). It is the most refractory carbon fraction and the most efficient solar-light absorber. OC is contained in organic matter which is composed of thousands of chemical constituents belonging to many compound classes, for which complete characterisation is extremely difficult. Carbonatic carbon (CC), that is the carbon contained in carbonates, may also be present, however its contribution to total carbon may be considered negligible in most European areas, with few exceptions (Perrone et al., 2011; Cuccia et al., 2011; Yubero et al., 2011).

While EC is exclusively produced by direct combustion emissions, OC may derive from primary sources (primary organic carbon, POC), such as fossil-fuel combustion, biomass burning and bioaerosol emissions, as well as from the atmospheric gas-to-particle conversion of other pollutants through condensation processes (driven by temperature and dilution effects) and oxidation processes (secondary organic carbon, SOC). Since most of the emitted POC is semivolatile and some gas-to-particle processes take place shortly after emission, some authors consider the distinction between POC and SOC to be obsolete. In this document, this terminology is used for coherence with the reviewed literature and to emphasise the difference between sources and processes, which is relevant for the development of abatement measures. The lack of direct chemical analysis methods for the determination of either POC or SOC led to the development of different indirect approaches, of which the most widespread is the method based on variations of measured OC/EC ratios. In this approach, elemental carbon is assumed to be a conservative tracer for primary combustion-generated OC emissions, and SOC simply appears as an increase in the OC/EC ratio relative to that of the primary OC/EC ratio (Turpin and Huntzicker, 1995).

Large uncertainties still affect emission inventories of carbonaceous particles. Monks et al. (2009) reviewed global emission estimates: uncertainties up to factors 3.4 and 80 are reported for primary and secondary carbonaceous particles, respectively. The highest uncertainties still concern natural emissions.

The reactivity, volatility and to some extent the hygroscopicity of compounds in the OC fraction, also including main source tracers (such as levoglucosan), may compromise the basic assumptions for receptor models and strongly increase difficulties and uncertainties in source apportionment.

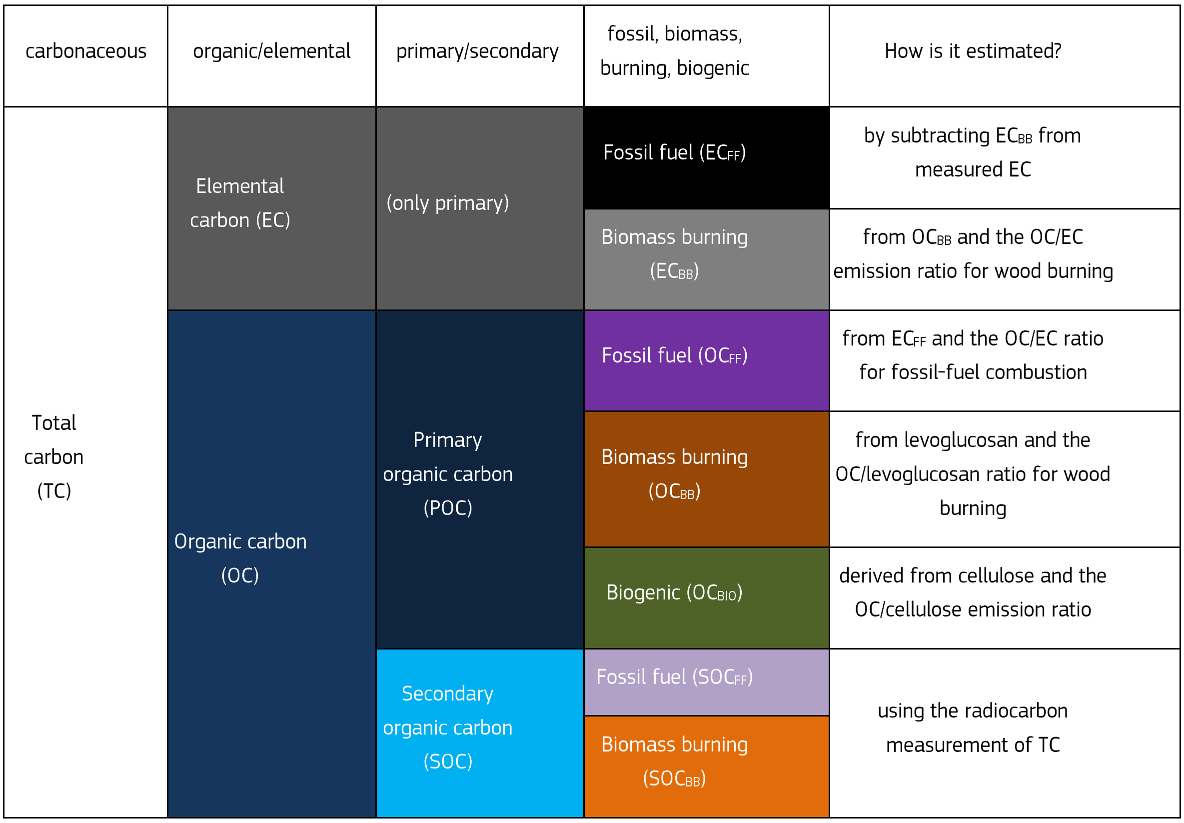
In this context, the use of “inert” tracers, such as the $^{14}C/^{12}C$ isotopic ratio, may be of great help. Radiocarbon measurement of TC is a good tool for fossil/non-fossil source separation (Currie, 2000 and the literature cited therein; Hildemann et al., 1994). The main principle may be briefly explained as follows. "Modern" carbon from biomass contains a constant proportion of radioactive ^{14}C, giving a $^{14}C/^{12}C$ isotopic ratio of 1. ^{14}C decays with a radiocarbon half-life of 5 730 years, which means that none is left in fossil fuels, which are millions of years old. Therefore, as the fraction of modern carbon (f_m) is zero for fossil fuels, and as f_m should be 1 for modern materials, it is possible to estimate the proportion of fossil and non-fossil fuels that led to a particular level of total carbon in the atmosphere by looking at the value of the $^{14}C/^{12}C$ isotopic ratio. Actually, nuclear tests in the 1950s increased the $^{14}C/^{12}C$ ratio in the atmosphere by up to a factor of 2 in the early 1960s. Values have been slowly decreasing since then and f_m is now approaching 1: the trend of the ^{14}C content in the atmosphere can be found in Levin et al. (2010). The excess of ^{14}C in the atmosphere led to the increase of $^{14}C/^{12}C$ ratio in biological material and must be taken into account when apportioning modern sources.

However, the sole use of radiocarbon measurements on total carbon only allows for a separation between modern and fossil contributions. This simple division is not enough to apportion natural and anthropogenic sources since modern carbon could result from natural emissions as well as from wood/biomass burning and other anthropogenic activities (such as cooking). To overcome this limitation, Szidat et al. (2004; 2006) proposed performing radiocarbon measurements of OC and EC separately. In this way, EC may be directly apportioned between fossil-fuel combustion and biomass burning, and the fossil-fuel combustion contribution to OC may be also directly obtained; provided that the OC/EC emission ratio for wood/biomass burning is known, the two primary modern contributions to OC may be estimated. This model is limited by the uncertainty regarding the knowledge of the OC/EC emission ratio for wood/biomass burning and by the difficulty in the assessment of the secondary contribution of this source (Szidat et al., 2009), as the OC/EC emission ratio measured at the source cannot correctly account for secondary aerosol formation. Moreover, this method requires an effective physical isolation of the two carbonaceous fractions (Andersson et al., 2011; Bernardoni et al., 2013; Calzolai et al., 2011; Heal et al., 2011; Szidat et al., 2004, 2009; Zencak et al., 2007; Zhang et al., 2012; Minguillon et al., 2016), which are operationally defined quantities. Indeed, the analytical separation of OC from EC using thermal protocols is ambiguous because part of the OC can pyrolyse, especially in an oxygen-poor atmosphere, and some of this EC can evolve in the presence of oxygen (Watson et al., 2005). It is also noteworthy that water-soluble organic carbon (WSOC) is particularly prone to pyrolysis and that soluble inorganic compounds can catalyse EC pre-combustion (Chow et al., 2001; Novakov and Corrigan, 1995; Wang et al., 2010; Yu et al., 2002).

Recent literature has attempted a natural/anthropogenic source apportionment, coupling ^{14}C measurements of TC with the analysis of other micro and macro tracers (Gelencsér et al., 2007; May et al., 2009; Gilardoni et al., 2011; Holden et al., 2011; Yttri et al., 2011a, 2011b). A number of tracers and emission factors have been employed in these studies: levoglucosan as tracer for biomass combustion together with OC/levoglucosan and OC/EC emission ratios; carbon monoxide as tracer for primary fossil-fuel combustion together with the OC/EC emission ratio; cellulose for plant debris together with the OC/cellulose ratio; arabitol and mannitol saccharide concentrations as tracers of fungal spores. In these papers, marker concentrations, emission ratios and their uncertainties were used to estimate possible ranges of source contributions identified by modelling techniques.

In the work by Gelencsér et al. (2007), measurements of EC, OC, levoglucosan, cellulose and f_m(total carbon) are used for TC apportionment in the following basic classes (Table 15.1): EC from fossil fuel combustion (EC_{FF}) and from biomass burning (EC_{BB}), POC from fossil fuel combustion (OC_{FF}), biomass burning (OC_{BB}) and biogenic sources (OC_{BIO}), and SOC from precursors emitted by fossil and non-fossil sources. Briefly, OC_{BIO} is derived from cellulose and the OC/cellulose emission ratio; OC_{BB} from levoglucosan and the OC/levoglucosan ratio for wood burning; EC_{BB} from OC_{BB} and the OC/EC emission ratio for wood burning; EC_{FF} by subtracting EC_{BB} from measured EC; OC_{FF} from EC_{FF} and the OC/EC ratio for fossil-fuel combustion. SOC is grouped as fossil and non-fossil using the radiocarbon measurement of TC.

Table 15.1. Schematic representation of the carbonaceous fractions derived by combining radiocarbon measurements with organic markers in Gelencsér et al. (2007).



carbonaceous	organic/elemental	primary/secondary	fossil, biomass, burning, biogenic	How is it estimated?
Total carbon (TC)	Elemental carbon (EC)	(only primary)	Fossil fuel (EC_{FF})	by subtracting EC_{BB} from measured EC
			Biomass burning (EC_{BB})	from OC_{BB} and the OC/EC emission ratio for wood burning
	Organic carbon (OC)	Primary organic carbon (POC)	Fossil fuel (OC_{FF})	from EC_{FF} and the OC/EC ratio for fossil-fuel combustion
			Biomass burning (OC_{BB})	from levoglucosan and the OC/levoglucosan ratio for wood burning
			Biogenic (OC_{BIO})	derived from cellulose and the OC/cellulose emission ratio
		Secondary organic carbon (SOC)	Fossil fuel (SOC_{FF})	using the radiocarbon measurement of TC
			Biomass burning (SOC_{BB})	

This method involves many steps, each of which has substantial uncertainty, mainly due to the high variability of emission ratios: to tackle the multitude of possible combinations of these uncertainty parameters, a statistical approach, the Latin-hypercube sampling method, was used. A very similar approach is used by Gilardoni et al. (2011). In this case, however, the estimation of the primary biogenic OC contribution is based on the measurements of arabitol and mannitol saccharides, assuming that fungal spores dominate primary biological aerosol particles in $PM_{2.5}$. A Quasi-Montecarlo approach is used to determine the most probable contributions of the different source categories and the associated uncertainties.

In other works (e.g., Aiken et. al., 2010; Minguillon et al., 2011; Heal et al., 2011; Bernardoni et al., 2013), source apportionment based on f_m measurements of OC and EC fractions is reinforced by comparison with other approaches, such as the levoglucosan-tracer method, PMF-aerosol mass spectrometry, PMF-offline filter dataset and Chemical Mass Balance of Organic Molecular Markers. In Bernardoni et al. (2013) and Minguillon et al. (2011), the fossil OC is also apportioned into primary and secondary contributions; moreover, a further apportionment of the non-fossil OC is performed in order to take into account modern contributions to OC other than wood burning and biogenic aerosols (e.g. cooking, tyre wear). For instance, in Bernardoni et al. (2013) the fossil fuel combustion contribution to OC is split into primary and secondary components using the EC-tracer method (Turpin and Huntzicher, 1995) which uses PMF-derived emission ratios, while the non-fossil OC is grouped into primary OC_{BB}, secondary OC_{BB}, OC_{BIO} and OC from other urban modern sources, according to emission ratios and literature approaches.

Minguillon et al. (2016) recently proposed advanced combination with aerosol mass spectrometry data (see section 13).

Finally, it should be noted that these recent probabilistic uncertainty characterisations have demonstrated that results obtained with such trace-based methods may have high uncertainties (Larsen et al, 2012). It is also worth mentioning that radiocarbon analyses are extremely time-consuming and expensive, due to the procedures for sample preparation and to the need for an Accelerator Mass Spectrometry system to determine the radiocarbon concentration. Such features limit the number of samples that can be characterised and, therefore, the representativeness of the obtained data. Nevertheless, some steps to overcome such limitations have been recently taken by coupling a commercial OC/EC analyser with an Accelerator Mass Spectrometry system (Perron et al., 2010).

References

Aiken A. C., de Foy B., Wiedinmyer C., DeCarlo P. F., Ulbrich I. M., Wehrli M. N., Szidat S., Prevot A. S. H., Noda J., Wacker L., Volkamer R., Fortner E., Wang J., Laskin A., Shutthananda V., Zheng J., Zhang R., Paredes-Miranda G., Arnott W. P., Molina L. T., Sosa G., Querol X., Jimenez J. L., 2010. Mexico-city aerosol analysis during MILAGRO using high resolution aerosol mass spectrometry at the urban supersite (T0) – Part 2: Analysis of the biomass burning contribution and the non-fossil carbon fraction. *Atmospheric Chemistry and Physics* 10, 5315–5341.

Andersson, A., Sheesley, R.J., Krusa, M., Johansson, C., Gustafsson, O., 2011. 14C-based source assessment of soot aerosols in Stockholm and the Swedish EMEP-Aspvreten regional background site. *Atmospheric Environment* 45, 215–222.

Bernardoni V., Calzolai G., Chiari M., Fedi M., Lucarelli F., Nava S., Piazzalunga A., Riccobono F., Taccetti F., Valli G. & Vecchi R., 2013. Radiocarbon analysis on organic and elemental carbon in aerosol samples and source apportionment at an urban site in Northern Italy. *Journal of Aerosol Science* 56, 88–99.

Bonazza, A., Sabbioni, C., & Ghedini, N., 2005. Quantitative data on carbon fractions in interpretation of black crusts and soiling on European built heritage. *Atmospheric Environment* 39, 2607–2618.

Calzolai G., Bernardoni V., Chiari M., Fedi M., Lucarelli F., Nava S., Riccobono F., Taccetti F., Valli G., Vecchi R., 2011. The new sample preparation line for radiocarbon measurements on atmospheric aerosol at LABEC. *Nuclear Instruments and Methods B* 269, 203-208.

Chow, J.C., Watson, J.G., Crow, D., Lowental, D.H., & Merrifield, T., 2001. Comparison of IMPROVE and NIOSH carbon measurements. *Aerosol Science and Technology* 34, 23–34.

Cuccia, E., Piazzalunga, A., Bernardoni, V., Brambilla, L., Fermo, P., Massabò o, D., Molteni, U., Prati, P., Valli, G., & Vecchi, R., 2011. Carbonate measurements in PM10 near the marble quarries of Carrara (Italy) by infrared spectroscopy (FT-IR) and source apportionment by positive matrix factorization (PMF). *Atmospheric Environment* 45, 6481–6487.

Currie, L.A., 2000. Evolution and multidisciplinary frontiers of 14C aerosol science. *Radiocarbon* 42, 115–126.

Gelencsér, A., May, B., Simpson, D., Sánchez-Ochoa, A., Kasper-Giebl, A., Puxbaum, H., Caseiro, A., Pio, C., & Legrand, M., 2007. Source apportionment of PM2.5 organic aerosol over Europe: primary/secondary, natural/anthropogenic, and fossil/biogenic origin. *Journal of Geophysical Research* 112, D23S04.

Gilardoni, S., Vignati, E., Cavalli, F., Putaud, J.P., Larsen B.R., Karl, M., Stenstrom, K., Gemberg, J., Henne, S., & Dentener, F., 2011. Better constraints on sources of carbonaceous aerosols using a combined 14C-macro tracer analysis in a European rural background site. *Atmospheric Chemistry and Physics* 11, 5685-5700.

Heal, M.R., Naysmith, P., Cook, G.T., Xu, S., Ravento´s Duran, T., & Harrison, R.M., 2011. Application of 14C analyses to source apportionment of carbonaceous PM2.5 in the UK. *Atmospheric Environment* 45, 2341–2348.

Holden, A.S., Sullivan, A.P., Munchak, L.A., Kreidenweis, S.M., Schichtel, B.A., Malm, W.C., & Collet, J.L., Jr., 2011. Determining contributions of biomass burning and other sources to fine particle contemporary carbon in the western United States. *Atmospheric Environment* 45, 1986–1993.

Hildemann, L.M., Klinedinst, D.B., Klouda, G.A., Currie, L.A., & Cass, G.R., 1994. Sources of urban contemporary carbon aerosol. *Environmental Science and Technology* 28, 1565–1575.

Highwood, E.J., & Kinnersley, R.P., 2006. When smoke gets in our eyes: the multiple impacts of atmospheric black carbon on climate, air quality and health. *Environment International* 32, 560–566.

IPCC 2007. In: S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, & H.L. Miller (Eds.), Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press: Cambridge, UK and New York, NY, USA.

Larsen B.R., Gilardoni S., Stenstro¨m K., Niedzialek J., Jimenez J., Belis C.A., 2012. Sources for PM air pollution in the Po Plain, Italy: II. Probabilistic uncertainty characterization and sensitivity analysis of secondary and primary sources. *Atmospheric Environment* 50, 203-213.

Levin, I., Naegler, T., Kromer, B., Diehl, M., Francey, R.J., Gomez-Pelaez, A.J., Steele, P., Wagenbach, D., Weller, R., & Worthy, D.E., 2010. Observations and modelling of the global distribution and long-term trend of atmospheric 14CO2. Tellus B 62B, 26–46.

Mauderly, J.L., Chow, J.C., 2008. Health effects of organic aerosols. *Inhalation Toxicology* 20, 257–288.

May, B., Wagenbach, D., Hammer, S., Steier, P., Puxbaum, H., & Pio, C., 2009. The anthropogenic influence on carbonaceous aerosol in the European background. *Tellus* 61B, 464–472.

Minguillon, M.C., Perron, N., Querol, X., Szidat, S., Fahrni, S.M., Alastuey, A., Jimenez, J.L., Mohr, C., Ortega, A.M., Day, D.A., Lanz, V.A., Wacker, L., Reche, C., Cusack, M., Amato, F., Kiss, G., Hoffer, A., Decesari, S., Moretti, F., Hillamo, R., Teinila, K., Seco, R., Penuelas, J., Metzger, A., Schallhart, S., Muller, M., Hansel, A., Burkhart, J.F., Baltensperger, U., & Prévot, A.S.H., 2011. Fossil versus contemporary sources of fine elemental and organic carbonaceous particulate matter during the DAURE campaign in Northeast Spain. *Atmospheric Chemistry and Physics* 11, 12067–12084.

Minguillon, M.C., Perez, N., Marchand, N. et al., 2016. Secondary organic aerosol origins in an urban environment: influence of biogenic and fuel combustion precursors. *Faraday Discussions*, 189, 337-359, doi: 10.1039/c5fd00182.

Monks, P.S., Granier, C., Fuzzi, S., Stohl, A., Williams, M.L., Akimoto, H., Amanni, M., Baklanov, A., Baltensperger, U., Bey, I., Blake, N., Blake, R.S., Carslawn, K., Cooper, O.R., Dentener, F., Fowler, D., Fragkou, E., Frost, G.J., Generoso, S., Ginoux, P., Grewet, V., Guenther, A., Hansson, H.C., Hennew, S., Hjorth, J., Hofzumahaus, A., Huntrieser, H., Isaksen, I.S.A., Jenkin, M.E., Kaiser, J., Kanakidou, M., Klimont, Z., Kulmala, M., Laj, P., Lawrence, M.G., Lee, J.D., Liousse, C., Maione, M.,McFiggans, G.,Metzger, A., Mieville, A., Moussiopoulos, N., Orlando, J.J., O'Dowd, C.D., Palmer, P.I., Parrish, D.D., Petzold, A., Platt, U., Poschl, U., Prévot, A.S.H., Reeves, C.E., Reimann, S., Rudich, Y., Sellegri, K., Steinbrecher, R., Simpson, D., ten Brink, H., Theloke, J., van der Werf, G.R., Vautard, R., Vestreng, R., Vlachokostas, C., von Glasow, R., 2009. Atmospheric composition change—global and regional air quality. *Atmospheric Environment* 43, 5268–5350.

Novakov, T., & Corrigan, C.E., 1995. Thermal characterization of biomass smoke particles. *Mikrochimica Acta* 119, 157–166.

Perron N., Szidat S., Fahrni S., Ruff M., Wacker L., Prévot A.S.H, Baltensperger U., 2010. Towards on-line 14C analysis of carbonaceous aerosol fractions. *Radiocarbon* 52, 761–768.

Perrone, M.R., Piazzalunga, A., Prato, M., Carofalo, I., 2011. Composition of fine and coarse particles in a coastal site of the Central Mediterranean: carbonaceous specie contributions. *Atmospheric Environment* 45, 7470–7477.

Piazzalunga, A., Belis, C., Bernardoni, V., Cazzuli, O., Fermo, P., Valli, G., Vecchi, R., 2011. Estimates of wood burning contribution to PM by the macrotracer method using tailored emission factors. *Atmospheric Environment* 45, 6642–6649.

Putaud, J.P., Raes, F., Van Dingenen, R., Brüggenmann, E., Facchini, M.C., Decesari, S., Fuzzi, S., Gehrig, R., Hüglin, C., Laj, P., Lorbeer, G., Maenhaut, W., Mihalopulos, N., Müller, K., Querol, X., Rodriguez, S., Schneider, J., Spindler, G., ten Brink, H., Tørseth, K., Wiedensholer, A., 2004. A European aerosol phenomenology-2: chemical characteristics of particulate matter at kerbside, urban, rural and background sites in Europe. *Atmospheric Environment* 38, 2579–2595.

Szidat, S., Jenk, T.M., Gaggeler, H.W., Synal, H.-A., Hajdas, I., Bonani, G., Saurer, M., 2004. THEODORE, a two-step heating system for the EC/OC determination of radiocarbon (14C) in the environment. *Nuclear Instruments and Methods in Physics Research Section B*: Beam Interactions with Materials and Atoms, 223–224, 829–836.

Szidat, S., Jenk, T.M., Synal, H.-A., Kalberer, M., Wacker, L., Hajdas, I., Kasper-Giebl, A., Baltensperger, U., (2006). Contributions of fossil fuel, biomassburning, and biogenic emissions to carbonaceous aerosols in Zurich as traced by 14C. *Journal of Geophysical Research* 111, D07206.

Szidat, S., Ruff, M., Perron, N., Wacker, L., Synal, H.-A., Hallquist, M., Shannigrahi, A.S., Yttri, K.E., Dye, C., Simpson, D., 2009. Fossil and non-fossil sources of organic carbon (OC) and elemental carbon (EC) in Goteborg, Sweden. *Atmospheric Chemistry and Physics* 9, 1521–1535.

Turpin, B.J., Huntzicker, J.J., 1995. Identification of secondary organic aerosol episodes and quantitation of primary and secondary organic aerosol concentrations during SCAQS. *Atmospheric Environment* 29, 3527–3544.

Vecchi R., Chiari M., D'Alessandro A., Fermo P., Lucarelli F., Mazzei F., Nava S., Piazzalunga A., Prati P., Silvani F., Valli G., 2008. A mass closure and PMF source apportionment study on the sub-micron sized aerosol fraction at urban sites in Italy. *Atmospheric Environment* 42, 2240–2253.

Wang, Y., Chung, A., & Paulson, S.E., 2010. The effect of metal salts on quantification of elemental and organic carbon in diesel exhaust particles using thermal-optical evolved gas analysis. *Atmospheric Chemistry and Physics* 10, 11447–11457.

Watson, J.G., 2002. Visibility: science and regulation. *Journal of Air & Waste Management Association* 52, 628–713.

Watson, J.G., Chow, J.C., & Chen, L.-W.A., 2005. Summary of organic and elemental carbon/black carbon analysis methods and intercomparisons. A*erosol and Air Quality Research* 5, 65–102.

Yttri, K.E., Simpson, D., Nøjgaard, J.K., Kristensen, K., Genberg, J., Stenstr ¨om, K., Swietlicki, E., Hillamo, R., Aurela, M., Bauer, H., Offenberg, J.H., Jaoui, M., Dye, C., Eckhardt, S., Burkhart, J.F., Stohl, A., & Glasius, M., 2011a. Source apportionment of the summer time carbonaceous aerosol at Nordic rural background sites. *Atmospheric Chemistry and Physics* 11, 13339–13357.

Yttri, K.E., Simpson, D., Stenstrom, K., Puxbaum, H., & Svendby, T., 2011b. Source apportionment of the carbonaceous aerosol in Norway—quantitative estimates based on 14C, thermal-optical and organic tracer analysis. *Atmospheric Chemistry and Physics* 11, 9375–9394.

Yu, J.Z., Xu, J.H., & Yang, H., 2002. Charring characteristics of atmospheric organic particulate matter in thermal analysis. *Environmental Science and Technology* 36, 754–761.

Yubero, E., Carratalá, A., Crespo, J., Nicolás, J., Santacatalina, M., Nava, S., Lucarelli, F., & Chiari, M., 2011. PM10 source apportionment in the surroundings of the San Vicente del Raspeig cement plant complex in southeastern Spain. *Environmental Science and Pollution Research* 18, 64–74.

Zencak, Z., Elmquist, M., & Gustafsson, O., 2007. Quantification and radiocarbon source apportionment of black carbon in atmospheric aerosols using the CTO-375 method. *Atmospheric Environment* 41, 7895–7906.

Zhang, Y.L., Perron, N., Ciobanu, V. G., Zotter P., Minguillon M. C., Wacker, L., Prevot A.S.H., Baltensperger U., & Szidat, S., 2012. On the isolation of OC and EC and the optimal strategy of radiocarbon-based source apportionment of carbonaceous aerosols, *Atmospheric Chemistry and Physics* 12, 10841–10856.

16. PROTON-NUCLEAR MAGNETIC RESONANCE SPECTROSCOPY FOR THE SOURCE APPORTIONMENT OF WATER SOLUBLE ORGANIC CARBON

Proton-Nuclear Magnetic Resonance (1H-NMR) spectroscopy is a quasi-universal analytical technique for organic compounds analysis. First applications to atmospheric aerosol samples date back the early 2000's (Decesari et al., 2000; Suzuki et al. 2001). The technique is suitable for off-line analysis of organic compounds extractable into solvents. As impurities in the solvent carry interferences in the spectra, low-concentration analyses have been performed mostly with the cleanest solvent commercially available, which is deuterium oxide (D2O), although a sparse literature exists also for aerosol organic compounds soluble in methanol and other organic solvents (Facchini et al., 2008a; Schmitt-Kopplin 2010, Matos et al., 2017). It is a matter of fact, however, that all existing source apportionment studies relying on 1H-NMR spectroscopy provide information solely on aerosol water-soluble organic carbon (WSOC). 1H-NMR spectroscopy in protic solvents mainly provides the speciation of hydrogen atoms bound to carbon atoms (C-H bonds). The 1H-NMR resonances can be attributed to specific C-H containing functionalities. On the basis of the range of frequency shifts (the so-called "chemical shift", ppm) (Decesari et al., 2000; Decesari et al., 2007). The concentrations of individual organic compounds including amines, carboxylic and sulphonic acids can also be determined (e.g., Facchini et al., 2008b). First attempts of WSOC source attribution based on 1H-NMR functional group compositions are provided by Decesari et al. (2007). The approach, relying on combined metrics derived from 1H-NMR functional group molar ratios, was used to provide a coarse categorization of WSOC sources ("biomass burning", "secondary organic aerosols (SOA)", "marine organic aerosols"). The method was subsequently applied in a range of environments (Ziemba et al., 2011; Cleveland et al., 2012; Shakya et al. 2012; Lopes et al., 2015), and was shown to only partly capture the site-specific characteristics of the 1H-NMR aerosol compositions.

Starting from 2010, atmospheric aerosol 1H-NMR spectral datasets were elaborated using chemometric techniques, including principal component analysis (PCA) and factor analysis, with the aim of source identification and source apportionment. The approach is completely analogous to the existing applications developed for mass spectrometric datasets. The 1H-NMR spectra are normally binned to some hundred data points, but still retaining much more information with respect to methods relying solely on major functional group compositions such as in Decesari et al. (2007). Examples of PCA approaches are provided by Decesari et al. (2011) and Wozniak et al. (2015), both focusing on marine organic aerosols. In particular, 1H-NMR spectroscopy provides a clear discrimination between aromatic and aliphatic structures, which can be challenging to achieve by means such as aerosol mass spectrometry but can be critical for the discrimination between biogenic and anthropogenic SOA.

Factor analysis was also employed for 1H-NMR spectra deconvolution and organic source apportionment. Five distinct methods of non-negative factor analysis were tested and inter-compared as explained by Paglione et al. (2014a and 2014b). Specifically, “Positive Matrix Factorization” (PMF) (Paatero and Tapper, 1994; Paatero, 1999) was applied exploiting the US EPA open-source software EPA-PMF (v3 and v5), while two methods of “Non-negative Matrix Factorization” (N-NMF) and other two techniques for “Multivariate Curve Resolution” (MCR) (Karakach et al., 2009) were also employed (Table 1). On the basis of these first findings, the performance of simple factor analysis methods such as MCR and NMF in deconvolving 1H-NMR spectra was not inferior with respect to methods accounting for a better description of the uncertainty of the measurements, such as PMF. Nevertheless, more efforts in constraining the sources of uncertainty in the 1H-NMR spectra of atmospheric WSOC would certainly lead to an improvement and possibly to a standardization of PMF-NMR methods, analogously to what has been achieved for PMF-AMS. The interpretation of the spectral profiles relies on the comparison with libraries of 1H-NMR spectra of SOA and primary organic aerosol (POA) generated in laboratory conditions, or on the presence of NMR peaks identifying specific molecular tracers (e.g., methanesulphonic acid, pinonic acid, dimethylamine, levoglucosan, etc.). The existing libraries of 1H-NMR spectra of SOA and POA are largely incomplete if compared to the vast database obtained by the AMS user community, posing a major limitation to the source attribution of many 1H-NMR spectral profiles if not supported by the information derived from molecular markers. As a consequence, spectral profiles of so-called “humic-like substances” (HULIS) are often extracted in 1H-NMR factor analysis but not attributable to any specific biogenic or anthropogenic source (Paglione et al. 2014a). However, some hints on the possible source types or source regions of unassigned 1H-NMR factors can be acquired by inspecting their time trends, comparing with the concentration trends of additional markers, and taking into account the air mass origin and history.

The literature on 1H-NMR factor analysis refers to intensive field campaigns performed in Europe, North America and in the North Atlantic. The factor analysis techniques allowed the identification of recurrent source contributions to aerosol WSOC in the Northern Hemisphere:

1. Marine SOA, containing methane-sulphonate (MSA) and found at coastal Atlantic stations and in Mediterranean countries (Decesari et al., 2011, Paglione et al., 2014a; Wozniak et al. 2015);
2. Aliphatic amines from terrestrial sources (soil respiration, agricultural activities, waste management, etc.) which were found northern countries and in the Po Valley, Italy (Paglione et al., 2014b; Finessi et al., 2012);
3. Biomass burning POA, containing tracers of cellulose and lignine pyrolysis (Paglione et al., 2014b; Willoughby et al. 2016; Chalbot et al., 2016);
4. Biogenic SOA from terpene oxidation, with spectral features matching those of SOA generated in smog chamber experiments starting from monoterpene ozonolysis along with photooxidation, and those occurring at forest sites in Scandinavia and central Europe (Finessi et al., 2012).
5. Other factors, possibly including contributions from Primary Biological Aerosol Particles (PBAPs) found at forest sites (Chalbot et al., 2016), and by cooking aerosol (Decesari et al., 2014).

Additional factors characteristic of HULIS represent the most common 1H-NMR factor type found in the background European atmosphere (Decesari et al., 2011; Finessi et al., 2012; Paglione et al., 2014a & 2014b; Decesari et al., 2014; Gilardoni et al., 2016), but, as mentioned above, are not specific for any distinct biogenic or anthropogenic source. HULIS are considered a product of atmospheric ageing. Recent research findings in the Po Valley (Italy) support the hypothesis that HULIS can form through atmospheric ageing of biomass burning compounds in fog droplets and deliquesced particles. Overall, 1H-NMR -based source apportionment techniques can be used for research purposes for the identification of sources of anthropogenic and biogenic SOA and of polar POA fractions. Being still largely non-standardized, the technique is not aimed for implementation in monitoring networks.

References

Chalbot, M.C.G., P. Chitranshi, G.G. da Costa, E. Pollock, I.G. Kavouras, 2016. Characterization of water-soluble organic matter in urban aerosol by 1H-NMR spectroscopy, *Atmos. Environ.*, 128, 235–245.

Cleveland, M.J., L.D. Ziemba, R.J. Griffin, J.E. Dibb, C.H. Anderson, B. Lefer, B. Rappenglück, 2012. Characterization of urban aerosol using aerosol mass spectrometry and proton nuclear magnetic resonance spectroscopy, *Atmospheric Environment*, 54, 511-518.

Decesari, S., Facchini, M. C., Fuzzi, S., and Tagliavini, E., 2000. Characterization of water-soluble organic compounds in atmospheric aerosol: a new approach, *J. Geophys. Res.*, 105, 1481–1489.

Decesari, S., Mircea, M., Cavalli, F., Fuzzi, S., Moretti, F., Tagliavini, E., and Facchini, M. C., 2007. Source attribution of water-soluble organic aerosol by nuclear magnetic resonance spectroscopy, *Environ. Sci. Technol.*, 41, 2479–2484.

Decesari, S., Finessi, E., Rinaldi, M., Paglione, M., Fuzzi, S., Stephanou, E. G., Tziaras, T., Spyros, A., Ceburnis, D., O'Dowd, C., Dall'Osto, M., Harrison, R. M., Allan, J., Coe, H., Facchini, M. C., 2011. Primary and secondary marine organic aerosols over the North Atlantic Ocean during the MAP experiment, *J. Geophys. Res.*, 116, D22210, doi:10.1029/2011JD016204.

Decesari, S., Allan, J., Plass-Duelmer, C., Williams, B. J., Paglione, M., Facchini, M. C., O'Dowd, C., Harrison, R. M., Gietl, J. K., Coe, H., Giulianelli, L., Gobbi, G. P., Lanconelli, C., Carbone, C., Worsnop, D., Lambe, A. T., Ahern, A. T., Moretti, F., Tagliavini, E., Elste, T., Gilge, S., Zhang, Y., and Dall'Osto, M., 2014. Measurements of the aerosol chemical composition and mixing state in the Po Valley using multiple spectroscopic techniques, *Atmos. Chem. Phys.*, 14, 12109-12132, doi:10.5194/acp-14-12109-2014.

Facchini, M. C., et al., 2008a. Primary submicron marine aerosol dominated by insoluble organic colloids and aggregates, Geophys. Res. Lett., DOI: 10.1029/2008GL034210.

Facchini, M. C., et al., 2008b. Important source of marine secondary organic aerosol from biogenic amines, *Environ. Sci. Technol.*, 42(24), 9116–9121, doi:10.1021/es8018385.

Finessi, E., Decesari, S., Paglione, M., Giulianelli, L., Carbone, C., Gilardoni, S., Fuzzi, S., Saarikoski, S., Raatikainen, T., Hillamo, R., Allan, J., Mentel, Th. F., Tiitta, P., Laaksonen, A., Petäjä, T., Kulmala, M., Worsnop, D. R., and Facchini, M. C., 2012. Determination of the biogenic secondary organic aerosol fraction in the boreal forest by NMR spectroscopy, *Atmos. Chem. Phys.*, 12, 941–959, doi:10.5194/acp-12-941-2012.

Gilardoni, S., Massoli, P., Paglione, M., Giulianelli, L., Carbone, C., Rinaldi, M., Decesari, S., Sandrini, S., Costabile, F., Gobbi, G. P., Pietrogrande, M. C., Visentin, M., Scotto, F., Fuzzi, S., and Facchini, M. C., 2016. Direct observation of aqueous secondary organic aerosol from biomass burning emissions, *P. Natl. Acad. Sci.* USA, 113, 10013–10018.

Jaumot, J., Gargallo, R., de Juan, A., and Tauler, R., 2005. A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB, *Chem. Int. Lab. Syst.*, 76, 101– 110.

Karakach, T. K., Knight, R., Lenz, E. M., Viant, M. R., and Walter, J. A., 2009. Analysis of time course 1H NMR metabolomics data by multivariate curve resolution, Magn. Reson. Chem., 47 S105-117.

Lee, D. and Seung, H., 2001. Algorithms for non-negative matrix factorization, *Adv. Neural Inform. Process. Syst.*, 13, 556–562.

Lin, C.-J., 2007. Projected Gradient methods for Non-negative matrix factorization, *Neural Comp.*, 19, 2756–2779.

Lopes, S. P., J. T.V. Matos, A.M.S. Silva, A. C. Duarte, R.M.B.O. Duarte, 2015. 1H-NMR studies of water- and alkaline-soluble organic matter from fine urban atmospheric aerosols, *Atmospheric Environment*, 119, 374-380.

Matos J.T.V., Duarte R.M.B.O., Lopes S.P., Silva A.M.S., Duarte A.C., 2017. Persistence of urban organic aerosols composition: Decoding their structural complexity and seasonal variability, *Environ Pollut.*, 11;231(Pt 1):281-290

Paatero, P. and Tapper, U., 1994. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values, Environmetrics, 5, 111–126, doi:10.1002/env.3170050203.

Paatero, P., 1999. The Multilinear Engine: A Table-Driven, Least Squares Program for Solving Multilinear Problems, including the n-Way Parallel Factor Analysis Model, *J. Comp. Graph. Stat.*, 8, 854–888.

Paglione, M., Kiendler-Scharr, A., Mensah, A. A., Finessi, E., Giulianelli, L., Sandrini, S., Facchini, M. C., Fuzzi, S., Schlag, P., Piazzalunga, A., Tagliavini, E., Henzing, J. S., and Decesari, S., 2014a. Identification of humic-like substances (HULIS) in oxygenated organic aerosols using NMR and AMS factor analyses and liquid chromatographic techniques, *Atmos. Chem. Phys.*, 14, 25-45, doi:10.5194/acp-14-25-2014.

Paglione, M., Saarikoski, S., Carbone, S., Hillamo, R., Facchini, M.C., Finessi, E., Giulianelli, L., Carbone, C., Fuzzi, S., Moretti, F., Tagliavini, E., Swietlicki, E., Eriksson Stenström, K., Prévôt, A. S. H., Massoli, P., Canaragatna, M., Worsnop, D., and Decesari, S., 2014b. Primary and secondary biomass burning aerosols determined by proton nuclear magnetic resonance (1H-NMR) spectroscopy during the 2008 EUCAARI campaign in the Po Valley (Italy), *Atmos. Chem. Phys.*, 14, 5089–5110, doi:10.5194/acp-14-5089-2014.

Schmitt-Kopplin et al., 2010. Analysis of the Unresolved Organic Fraction in Atmospheric Aerosols with Ultrahigh-Resolution Mass Spectrometry and Nuclear Magnetic Resonance Spectroscopy: Organosulfates As Photochemical Smog Constituents, *Anal. Chem.*, 82, 8017–8026.

Suzuki, Y., Kawakami, M., Akasaka, K., 2001. 1H-NMR Application for Characterizing Water-Soluble Organic Compounds in Urban Atmospheric Particles, *Environ. Sci. Technol.*, 35, 2656–2664.

Tauler, R., 1995. Multivariate Curve Resolution applied to second order data, *Chem. Int. Lab. Syst.*, 30, 133–146.

Shakya, K.M., Place Jr., P.F., Griffin, R.J., Talbot, R.W., 2012. Carbonaceous content and water-soluble organic functionality of atmospheric aerosols at a semi-rural New England location, *J. Geophys. Res.*, 117, D03301.

Wentzell, P. D., Karakach, T. K., Roy, S., Martinez, M. J., Allen, C. P., and Werner-Washburne, M., 2006. Multivariate curve resolution of time course microarray data, *BMC Bioinformatics*, 7, 343–362.

Willoughby, A. S., A. S. Wozniak, P. G. Hatcher, 2016. Detailed Source-Specific Molecular Composition of Ambient Aerosol Organic Matter Using Ultrahigh Resolution Mass Spectrometry and 1H-NMR, *Atmosphere*, 7, 79; doi:10.3390/atmos7060079.

Wozniak A.S., R. U. Shelley, S. D. McElhenie, W. M. Landing, P. G. Hatcher, 2015. Aerosol water soluble organic matter characteristics over the North Atlantic Ocean: Implications for iron-binding ligands and iron solubility, *Mar. Chem.*, 173, 162-172.

Ziemba L.D., R.J. Grif, S. Whitlow, R.W. Talbot, 2011. Characterization of water-soluble organic aerosol in coastal New England: Implications of variations in size distribution, *Atmos. Environ.*, 45, 7319-7329.

17. SOURCE APPORTIONMENT BY FOURIER-TRANSFORM INFRARED (FTIR) ANALYSIS

FTIR measures the absorption of mid-infrared radiation due to change in dipole moments of vibrational modes of molecular bonds present (Anderson and Seyfried, 1948; Griffiths and Haseth, 2007; Hastings et al., 1952) in the aerosol. Many major constituents of PM are represented in the infrared spectrum, including ammonium sulfate, ammonium nitrate, dust, and a multitude of organic compounds. Due to the complexity of the mixture and the diversity of molecules that cannot be individually-resolved, the organic fraction is typically characterized by the abundance of its constituent functional groups using this technique (Allen et al., 1994; Blando et al., 1998; Russell, 2003). Functional groups thus far quantitatively characterized include aliphatic C-H, alcohol O-H, carboxylic COOH, non-carboxylic carbonyl C=O, and primary amine NH_2. Organonitrate $RONO_2$ and organosulfate $ROSO_3$ have additionally been quantified but not included in the PMF analysis. Source apportionment of ambient PM with FTIR has thus far focused on the organic fraction.

While there are many ways FTIR spectra can be obtained from atmospheric PM, the most common method — and the method by which all published source apportionment using this method has been performed thus far — is to record the infrared absorbance of PM collected on a Polytetrafluoroethylene (PTFE) filter by transmission-mode analysis (Maria et al., 2003). This analysis technique is non-destructive and therefore the same filter has been used for subsequent XRF analysis to obtain additional source-related characteristics for interpretation. The logistics of filter sampling together with the contribution of filter blanks in the analysis has led to FTIR measurements of PM typically on the order of a few hours to 24 hours.

The magnitude of infrared absorption is proportional to the abundance of the bond (Brouguer-Beer-Lambert law), which forms the basis for linear RM. However, the recorded absorbance comprises contributions from absorption and scattering by both the particles and PTFE. The data matrix is therefore prepared such that the variations to be explained by the RM are largely due to absorption by molecules present in the aerosol. Scattering by PTFE and larger particles is removed by a baseline model that interpolates between regions with little or no absorption (Takahama et al., 2013b). The PTFE absorption is many times greater than peaks of the aerosol and is removed by excluding this region of the spectra. Therefore, regions only above 1500 cm^{-1} have been used for PMF analysis thus far, though, in principle, lower frequencies beyond the PTFE peak include absorption from peroxides, organonitrates, and organosulfates that could additionally be included. Additional variations such as carbon dioxide and water vapor present in the analysis chamber that is not attributed to aerosol is removed by interpolation and wavelet denoising, respectively (Takahama et al., 2013b). The uncertainty matrix used for FTIR PMF only accounts for the blank variability across wavenumbers and the heteroscedastic contribution to the error has thus far not been included in published instances of this technique

(Russell et al., 2009). Therefore, the reduced chi-square of the selected solution is expected to be greater than unity.

Table 1. FTIR factors for organic PM1 aerosol and their associated functional groups, tracers, and studies. Categories are taken from Russell et al. (2011). Tracers are not comprehensive but is meant to be indicative of what was used in these studies. C-H: alkane hydrocarbon, COOH: carboxylic group, O-H: alcohol hydroxyl, C=O: non-carboxylic carbonyl, and NH4: inorganic ammonium (which is not quantified by present in the spectrum).

Categories		Functional groups	Tracers	Refs.
Fossil-fuel combusion	low O/C	C-H, COOH (, NH$_4$)	S, V, Ni	1–3, 5, 7, 8, 10–14
	high O/C	C-H, COOH (, NH4)	S, V, Ni	2–4, 6, 9, 12, 13
Terrestrial vegetation	non-burning	O-H, C=O, C-H		1, 6, 9–1, 14
	burning	C-H, C=O	K	2, 3, 5, 9, 11, 12
Marine biogenic	high O/C	O-H	Na, Cl	1, 4, 5, 7, 8, 12, 13
	low O/C	O-H, C-H	Na, Cl	3, 5

[1] Bahadur et al. (2010)
[2] Liu et al. (2009)
[3] Russell et al. (2009)
[4] Russell et al. (2010)
[5] Hawkins and Russell (2010)
[6] Hawkins et al. (2010)
[7] Schwartz et al. (2010)
[8] Shaw et al. (2010)
[9] Takahama et al. (2011)
[10] Liu et al. (2012)
[11] Corrigan et al. (2013)
[12] Takahama et al. (2013a)
[13] Liu et al. (2018a)
[14] Liu et al. (2018b)

Identifying underlying constituents solely from variations in spectra is more broadly addressed under the umbrella of multivariate curve resolution (de Juan and Tauler, 2006), in which constraints such as smoothness and non-negativity may be additionally imposed on the inverse solution. However, following the wide adoption in the environmental science community, PMF has been used for this task. Early work explored the number of factors and rotational parameter (Russell et al., 2009), and later analyses added pseudo-random seed values. Some spectral overlap (i.e., correlations) among factor profiles is expected, but solutions are found in which correlation among factor strengths that are minimized. As a preliminary analysis, the complexity of specific spectral features is examined by an evolving window factor analysis (Keller and Massart, 1992), in which principal component analysis is applied to a sliding window to estimate the cumulative fraction of variance explained as a function wavenumber and number of components. Together with this analysis, the number of factors is determined by examining the change in chi-square (*Q*-value) and fraction of variance explained by PMF. While each component corresponds to a single molecule in the ideal case, actual PMF factors are interpreted to be molecular mixtures statistically-resolvable from the number of samples and the range of variations captured by them. The factors are then associated with sources by correlation to concentrations of auxiliary particle measurements (elemental and molecular markers), source regions identified by back trajectory analyses (by PSCF), or similarity to prior PMF factors or source spectra. Two to five factors are typically chosen in a typical analysis.

FTIR PMF has been largely applied to samples collected in North America, but also in diverse regions around the world such as marine environments near Europe and South America, and Antarctica. A comprehensive synthesis of field studies until 2010 is provided by Russell et al. (2014; 2011), and their major factor categories are retained in the current summary in Table 17.1. The past studies have focused on PM_1 aerosol, but source apportionment for $PM_{2.5}$ was also performed by Liu et al. (2012) (Bakersfield, CA) and Liu et al. (2018b) (Southeast US). Similar conclusions regarding sources were reached except that some factors were more oxygenated in the $PM_{2.5}$. In Europe, fossil fuel combustion and marine factors were found in the North Atlantic (Russell et al., 2010) and fossil fuel combustion, biogenic, and biomass burning in Finland (Corrigan et al., 2013).

The number of factors selected and its impact on interpretation has been explored in a limited number of cases. For instance, Takahama et al. (2011) found that in Whistler Canada, a three-factor solution yielded factors associated with urban, non-burning biogenic, and burning biogenic sources, while a two-factor solution resulted in the latter two factors combined as a single one. Factor recombination was explored by Hawkins and Russell (2010) as a three-factor solution had high OA reconstruction error, while a four-factor solution yielded two factors which individually had low explained variation and interpretability.

Because the FTIR spectrum comprises primary aerosol that may have undergone atmospheric processing or secondary aerosol which originated in the gas-phase, we interpret the source profiles obtained by PMF as the combined contributions from emission and transformation over a distinct set of variations (quasi-stationary assumption, Zhou et al. (2005)). In principle, for the same emission source this can lead to multiple profiles to the extent that they are statistically distinguishable from another, which results in multiple fossil fuel combustion factors with different O/C, and marine aerosol that is atmospherically transformed by mixing with polluted air masses. Guzman-Morales et al. (2014) used organic functional group estimates (particularly the alkane C-H fraction) to bound estimates of primary and secondary OA from their mixed contribution in the same PMF factor.

PMF analysis is aided by high time resolution and chemical specificity of the measurements. AMS benefits from the former, while FTIR particularly benefits from the latter; the two provide complementary information. For instance, PMF analysis of FTIR and AMS spectra at Bakersfield, CA, yielded good agreement in factor profiles and their contributions to OA, the factor profiles of AMS PMF were nearly identical and source association was largely due to tracer correlations (Liu et al., 2012). FTIR factor profiles, on the other hand, were more distinctive and provided additional support to the tracer correlations (Liu et al., 2012). Corrigan et al. (2013) also report that FTIR show higher discrimination between biogenic and biomass burning, particularly due to differences between proportions of carboxylic COOH and alcohol O-H groups. Furthermore, Hawkins et al. (2010) reported that AMS underestimated OA relative to when associated with refractory substances — i.e., crustal sources — potentially due to lower collection efficiency. Dust sources are often grouped into the biogenic category on account of plant materials comprising the organic fraction (Liu et al., 2012).

Given the large solution space of PMF, agglomerative cluster analysis (Ward Jr., 1963) has been used to provide additional confirmation of PMF results. The task of clustering can be interpreted as finding groups of spectra for which factor proportions are relatively similar. Cluster analysis of spectra has been used for grouping backtrajectories for PSCF analysis, and cluster analysis and PMF analysis have typically yield consistent interpretations (references in Table 1). Several source spectra with FTIR have been studied: biogenic SOA and biomass burning (Corrigan et al., 2013; Russell et al., 2011), marine aerosol (Frossard et al., 2014), and combustion of fossil fuel or alkane (Liu et al., 2018b; Price et al., 2017; Russell et al., 2011). However, CMB-type regression has been used so far by Takahama et al. (2011), who used previously-derived PMF spectra of fossil fuel combustion, biogenic, and burning sources and found excellent agreement between the explained variation by components to the identified cluster types and backtrajectory analyses in Whistler, Canada.

References

Allen, D.T., Palen, E.J., Haimov, M.I., Hering, S.V., Young, J.R., 1994. Fourier-transform Infrared-spectroscopy of aerosol collected in a low-pressure Impactor (LPI/FTIR) - Method development and field calibration, *Aerosol Science and Technology*, 21, 325-342.

Anderson, J.A., Seyfried, W.D., 1948. Determination of Oxygenated and Olefin Compound Types by Infrared Spectroscopy, *Analytical Chemistry*, 20, 998–1006.

Bahadur, R., Uplinger, T., Russell, L.M., Sive, B.C., Cliff, S.S., Millet, D.B., Goldstein, A., Bates, T.S., 2010. Phenol Groups in Northeastern US Submicrometer Aerosol Particles Produced from Seawater Sources, *Environ. Sci. Technol.*, 44, 2542–2548. https://doi.org/10.1021/es9032277

Blando, J.D., Porcja, R.J., Li, T.H., Bowman, D., Lioy, P.J., Turpin, B.J., 1998. Secondary formation and the Smoky Mountain organic aerosol: An examination of aerosol polarity and functional group composition during SEAVS RID F-6148-2011. *Environ. Sci. Technol.,* 32, 604–613. https://doi.org/10.1021/es970405s

Corrigan, A.L., Russell, L.M., Takahama, S., Äijälä, M., Ehn, M., Junninen, H., Rinne, J., Petäjä, T., Kulmala, M., Vogel, A.L., Hoffmann, T., Ebben, C.J., Geiger, F.M., Chhabra, P., Seinfeld, J.H., Worsnop, D.R., Song, W., Auld, J., Williams, J., 2013. Biogenic and biomass burning organic aerosol in a boreal forest at Hyytiälä, Finland, during HUMPPA-COPEC 2010. *Atmospheric Chemistry and Physics*, 13, 12233–12256. https://doi.org/10.5194/acp-13-12233-2013

de Juan, A., Tauler, R., 2006. Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications. *Critical Reviews in Analytical Chemistry*, 36, 163–176. https://doi.org/10.1080/10408340600970005

Frossard, A.A., Russell, L.M., Burrows, S.M., Elliott, S.M., Bates, T.S., Quinn, P.K., 2014. Sources and composition of submicron organic mass in marine aerosol particles, *Journal of Geophysical Research-atmospheres*, 119, 12977-13003. https://doi.org/10.1002/2014JD021913

Griffiths, P., Haseth, J.A.D., 2007. Fourier Transform Infrared Spectrometry, 2nd ed. John Wiley & Sons, In.

Guzman-Morales, J., Frossard, A.A., Corrigan, A.L., Russell, L.M., Liu, S., Takahama, S., Taylor, J.W., Allan, J., Coe, H., Zhao, Y., Goldstein, A.H., 2014. Estimated contributions of primary and secondary organic aerosol from fossil fuel combustion during the CalNex and Cal-Mex campaigns, *Atmos. Environ.*, 88, 330–340. https://doi.org/10.1016/j.atmosenv.2013.08.047

Hastings, S.H., Watson, A.T., Williams, R.B., Anderson, J.A., 1952. Determination of Hydrocarbon Functional Groups by Infrared Spectroscopy, *Analytical Chemistry*, 24, 612–618. https://doi.org/10.1021/ac60064a002

Hawkins, L.N., Russell, L.M., 2010. Oxidation of ketone groups in transported biomass burning aerosol from the 2008 Northern California Lightning Series fires, *Atmospheric Environment*, 44, 4142–4154. https://doi.org/10.1016/j.atmosenv.2010.07.036

Hawkins, L.N., Russell, L.M., Covert, D.S., Quinn, P.K., Bates, T.S., 2010. Carboxylic acids, sulfates, and organosulfates in processed continental organic aerosol over the southeast Pacific Ocean during VOCALS-REx 2008, *Journal of Geophysical Research-atmospheres*, 115. https://doi.org/10.1029/2009JD013276

Keller, H.R., Massart, D.L., 1992. Evolving Factor-analysis. *Chemometrics and Intelligent Laboratory Systems*, 12, 209–224. https://doi.org/10.1016/0169-7439(92)80002-L

Liu, J., Dedrick, J., Russell, L.M., Senum, G.I., Uin, J., Kuang, C., Springston, S.R., Leaitch, W.R., Aiken, A.C., Lubin, D., 2018a. High summertime aerosol organic functional group concentrations from marine and seabird sources at Ross Island, Antarctica, during AWARE. *Atmospheric Chemistry and Physics*, 18, 8571–8587. https://doi.org/10.5194/acp-18-8571-2018

Liu, J., Russell, L.M., Ruggeri, G., Takahama, S., Claflin, M.S., Ziemann, P.J., Pye, H.O.T., Murphy, B.N., Xu, L., Ng, N.L., McKinney, K.A., Budisulistiorini, S.H., Bertram, T.H., Nenes, A., Surratt, J.D., 2018b. Regional Similarities and NOx-Related Increases in Biogenic Secondary Organic Aerosol in Summertime Southeastern United States. *Journal of Geophysical Research-atmospheres*. https://doi.org/10.1029/2018JD028491

Liu, S., Ahlm, L., Day, D.A., Russell, L.M., Zhao, Y., Gentner, D.R., Weber, R.J., Goldstein, A.H., Jaoui, M., Offenberg, J.H., Kleindienst, T.E., Rubitschun, C., Surratt, J.D., Sheesley, R.J., Scheller, S., 2012. Secondary organic aerosol formation from fossil fuel sources contribute majority of summertime organic mass at Bakersfield. *Journal of Geophysical Research-atmospheres*, 117. https://doi.org/10.1029/2012JD018170

Liu, S., Takahama, S., Russell, L.M., Gilardoni, S., Baumgardner, D., 2009. Oxygenated organic functional groups and their sources in single and submicron organic particles in MILAGRO 2006 campaign. *Atmos. Chem. Phys.*, 9, 6849–6863. https://doi.org/10.5194/acp-9-6849-2009

Maria, S.F., Russell, L.M., Turpin, B.J., Porcja, R.J., Campos, T.L., Weber, R.J., Huebert, B.J., 2003. Source signatures of carbon monoxide and organic functional groups in Asian Pacific Regional Aerosol Characterization Experiment (ACE-Asia) submicron aerosol types. *Journal of Geophysical Research-atmospheres*, 108. https://doi.org/10.1029/2003JD003703

Price, D.J., Chen, C.-L., Russell, L.M., Lamjiri, M.A., Betha, R., Sanchez, K., Liu, J., Lee, A.K.Y., Cocker, D.R., 2017. More unsaturated, cooking-type hydrocarbon-like organic aerosol particle emissions from renewable diesel compared to ultra low sulfur diesel in at-sea operations of a research vessel. *Aerosol Science and Technology*, 51, 135–146. https://doi.org/10.1080/02786826.2016.1238033

Russell, L.M., 2014. Carbonaceous Particles: Source-Based Characterization of Their Formation, Composition, and Structures, in: Holland, H.D., Turekian, K.K. (Eds.), *Treatise on Geochemistry (Second Edition)*, Elsevier, Oxford, pp. 291–316. https://doi.org/10.1016/B978-0-08-095975-7.00415-0

Russell, L.M., 2003. Aerosol organic-mass-to-organic-carbon ratio measurements. *Environmental Science & Technology*, 37, 2982–2987. https://doi.org/10.1021/es026123w

Russell, L.M., Bahadur, R., Ziemann, P.J., 2011. Identifying organic aerosol sources by comparing functional group composition in chamber and atmospheric particles. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 3516–3521. https://doi.org/10.1073/pnas.1006461108

Russell, L.M., Hawkins, L.N., Frossard, A.A., Quinn, P.K., Bates, T.S., 2010. Carbohydrate-like composition of submicron atmospheric particles and their production from ocean bubble bursting. *Proceedings of the National Academy of Sciences of the United States of America*, 107, 6652–6657. https://doi.org/10.1073/pnas.0908905107

Russell, L.M., Takahama, S., Liu, S., Hawkins, L.N., Covert, D.S., Quinn, P.K., Bates, T.S., 2009. Oxygenated fraction and mass of organic aerosol from direct emission and atmospheric processing measured on the R/V Ronald Brown during TEXAQS/GoMACCS 2006. *Journal of Geophysical Research-atmospheres*, 114, D00F05. https://doi.org/10.1029/2008JD011275

Schwartz, R.E., Russell, L.M., Sjostedt, S.J., Vlasenko, A., Slowik, J.G., Abbatt, J.P.D., Macdonald, A.M., Li, S.M., Liggio, J., Toom-Sauntry, D., Leaitch, W.R., 2010. Biogenic oxidized organic functional groups in aerosol particles from a mountain forest site and their similarities to laboratory chamber products. *Atmospheric Chemistry and Physics*, 10, 5075–5088. https://doi.org/10.5194/acp-10-5075-2010

Shaw, P.M., Russell, L.M., Jefferson, A., Quinn, P.K., 2010. Arctic organic aerosol measurements show particles from mixed combustion in spring haze and from frost flowers in winter. *Geophysical Research Letters*, 37. https://doi.org/10.1029/2010GL042831

Takahama, S., Johnson, A., Morales, J.G., Russell, L.M., Duran, R., Rodriguez, G., Zheng, J., Zhang, R., Toom-Sauntry, D., Leaitch, W.R., 2013a. Submicron organic aerosol in Tijuana, Mexico, from local and Southern California sources during the CalMex campaign. *Atmospheric Environment*, 70, 500–512. https://doi.org/10.1016/j.atmosenv.2012.07.057

Takahama, S., Johnson, A., Russell, L.M., 2013b. Quantification of Carboxylic and Carbonyl Functional Groups in Organic Aerosol Infrared Absorbance Spectra. *Aerosol Science and Technology*, 47, 310–325. https://doi.org/10.1080/02786826.2012.752065

Takahama, S., Schwartz, R.E., Russell, L.M., Macdonald, A.M., Sharma, S., Leaitch, W.R., 2011. Organic functional groups in aerosol particles from burning and non-burning forest emissions at a high-elevation mountain site. *Atmospheric Chemistry and Physics*, 11, 6367–6386. https://doi.org/10.5194/acp-11-6367-2011

Ward Jr., J.H., 1963. Hierarchical Grouping to Optimize an Objective Function. *J. American Statistical Association*, 58, 236–244. https://doi.org/10.1080/01621459.1963.10500845

Zhou, L.M., Hopke, P.K., Stanier, C.O., Pandis, S.N., Ondov, J.M., Pancras, J.P., 2005. Investigation of the relationship between chemical composition and size distribution of airborne particles by partial least squares and positive matrix factorization. *Journal of Geophysical Research-atmospheres*, 110, D07S18. https://doi.org/10.1029/2004JD005050.

18. WIND AND TRAJECTORY ANALYSIS IN SOURCE APPORTIONMENT

Introduction

Source apportionment results are frequently complemented by procedures to identify the direction of air masses with high pollution levels or where certain compounds of interest come from (table 18.1).

Table 18.1. Hybrid wind-, trajectory- and circulation patterns-based models (adapted from Belis et al., 2013)

ANALYSIS OF WIND DIRECTION
Conditional probability function (CPF)
Non-parametric wind regression (NWR)
Pseudo deterministic receptor model (PDRM)
ANALYSIS OF BACKWARD TRAJECTORIES
Trajectory sector analysis (TSA)
Potential source contribution function (PSCF)
Concentration fields analysis (CFA)
Concentration-weighted trajectory (CWT)
Simplified quantitative transport bias analysis (SQTBA)
Trajectory mass balance (TRMB) or TRMB regression (TMBR)
ANALYSIS OF ATMOSPHERIC CIRCULATION PATTERNS
Cluster analysis-CA
Principal component analysis-PCA

For low- to medium-spatial scales this can be done by, for example, wind rose analysis (see section 12). However, medium and long-range transport may be better assessed using backward trajectories calculated with a suitable dispersion model (Stohl, 1998). In fact, it is presently assumed that the best way for interpreting atmospheric pollutant measurements at a receptor site using back-trajectories is by performing the analysis of large sets of them using specific trajectory statistical methods (TSM). Besides, some methodologies devoted to objectively characterize synoptic meteorological patterns favouring high concentrations of some air pollutants across a specific region can be used to complement source apportionment studies. These methodologies are known as classifications of atmospheric circulation patterns (Philipp et al., 2014).

Some of these models are extensively described hereafter. A few toolkits (e.g., Openair and ZeFir) have recently been developed to easily and efficiently run some of them from a single software package (Carslaw and Ropkins, 2012; Petit et al., 2017). Using such toolkits, model outputs can be plotted as gridded or smoothed maps.

Wind Direction Analysis (CPF and NWR)

The potential location of emission sources or the origin of polluted air masses can be investigated a posteriori, once the source contributions are already obtained. As a starting point, simple concentration roses (polar plots of sector-averaged wind contributions) can be used. The conditional probability function (CPF; Ashbaugh et al., 1985) is a common tool used to analyse point source impacts from varying wind directions using the source contribution estimates from receptor models coupled with the wind direction values measured on site (Kim et al., 2003). When particulate matter (PM) measurements are performed over 24 hours, the same daily source contribution is assigned to each hour of a given day in order to match to the hourly wind data. The conditional probability function (CPF) estimates the probability that a given source contribution from a given wind direction will exceed a predetermined threshold criterion. It is defined as:

$$CPF_{\Delta\theta} = \frac{m_{\Delta\theta}}{n_{\Delta\theta}} \qquad (18.1)$$

where $m_{\Delta\theta}$ is the number of occurrences from wind sector $\Delta\theta$ that exceeded the threshold criterion, and $n_{\Delta\theta}$ is the total amount of data from the same wind sector. Typically, 12 sectors are used ($\Delta\theta$ = 15 degrees) and calm wind periods are excluded due to the isotropic behaviour of wind vane under calm winds. The threshold criterion should be chosen based on sensitivity tests with several different percentiles of the fractional contribution from each source. A commonly used threshold is the 75th percentile (e. g. Amato and Hopke, 2012; Jeong et al., 2011; Kim et al., 2004).

The sources are likely to be located based on the wind directions that have high conditional probability values. A large number of papers have been published on the application of these approaches to the receptor modelling problem (Zhao and Hopke, 2006; Kim and Hopke, 2004, among others). However, Zhou et al. (2004) showed that the conditional probability function can provide misleading results when many directions are used with very few (or no) wind occurrences and when the distribution of concentrations is far from normal. The nonparametric wind regression analysis technique (NWR) is an alternative that can be used to locate sources. In this technique, the relationship of the contribution and wind direction is determined by kernel regression and confidence intervals are also given (Henry et al., 2002; Henry, 2002). The expected concentration *C* at *θ* is computed by:

$$\overline{C}(\theta, \Delta\theta) = \frac{\sum_{i=1}^{n} K((\theta - W)/\Delta\theta)C_i}{\sum_{i=1}^{n} K((\theta - W_i)/\Delta\theta)} \qquad (18.2)$$

where *K* is a Gaussian kernel function, *Wi* and *Ci* are the wind direction and concentration of the i^{th} sample respectively, and *Δθ* is the smoothing parameter, the only adjustable parameter in nonparametric regression (Zhou et al., 2004). Kim and Hopke (2004) showed that conditional probability function and nonparametric wind regression results are generally very consistent.

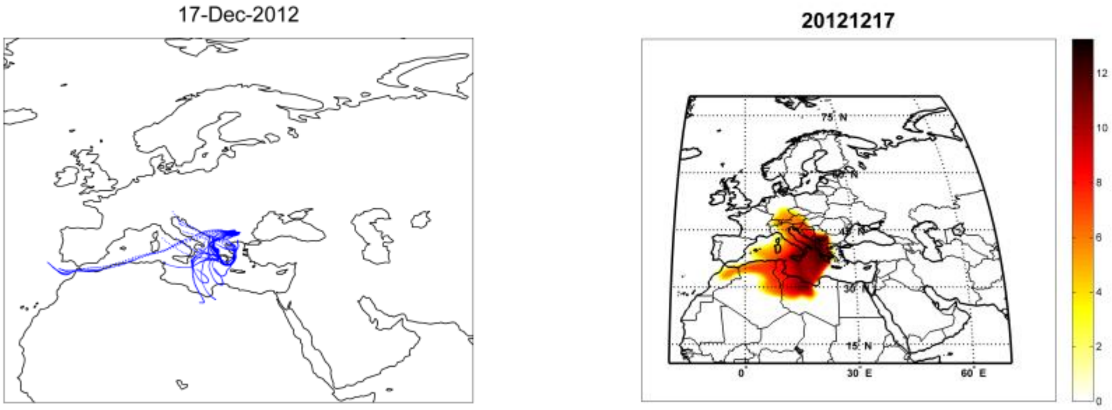
Backward Trajectory Computation

Back-trajectory models process gridded four-dimensional (*x*, *y*, *z*, *t*) meteorological fields to compute an estimation of the path followed by the air parcel during a user-defined run time before arriving at a specific time and height at a specified receptor. Depending on the receptor the height is defined as above ground or sea level (agl and asl, respectively). A model run generates a data file with the geographical coordinates (longitude, latitude and altitude or pressure) of the air parcel estimated position for every endpoint. The obtained back-trajectory is an estimate of the transport pathway of an infinitesimally air parcel. In itself the pathway is an estimate of the centerline of an advected air mass subject to vertical and horizontal dispersion (Kahl, 1993).

The most widely used tool to calculate back-trajectories is the freely available NOAA Hybrid Single Particle Lagrangian Integrated Trajectory (HYSPLIT) model (Draxler and Hess, 1998). Two different versions of the model are available, the stand-alone model and the web version (Rolph, 2012; Rolph et al., 2017). The latter allows for rapid feedback, but several limitations have been imposed to reduce computational activities on the NOAA Atmospheric Research Laboratory servers. Hence, local installation of the stand-alone model is recommended for routine use. The HYSPLIT model can process different meteorological file types, which needs to be downloaded prior to using the stand-alone program. The default meteorological files are provided by the Global data assimilation system (GDAS) and can be used for calculations at European geographical locations. The GDAS files are available in several spatial resolutions with 1-degree longitude and latitude being the default (http://www.arl.noaa.gov/HYSPLIT_info.php, accessed September 2018).

In the example given below (Figure 18.1), HYSPLIT model provides backward air mass trajectories every hour, going backward in time 5 days, beginning at a height of 500 m. above ground level.

Figure 18.1: 5 day backward trajectories and sensitivity plumes corresponding to daily filter measurements in Sofia. There are 24 hourly backward trajectories for HYSPLIT and 24 hourly releases of 20000 computational particles for FLEXPART. *Adapted from Perrone et al. (2018).*



In addition to HYSLPIT, there is a variety of European research-oriented models, such as FLEXTRA or FLEXPART (NILU) and REM-CALGRID (TRUMF), which allow back-trajectory plots to be produced. Back-trajectories at some specific locations are routinely calculated at NILU (Stohl et. al., 1995; Stohl and Seibert, 1998) and meteorological data provided from the European Centre for Medium Range Weather Forecast. Selected receptors include EMEP and EARLINET monitoring stations. Output of these runs, dating back to 1996, can be found on the NILU website (http://www.nilu.no/trajectories, accessed September 2018). In particular, FLEXPART model can be used to calculate backward sensitivity plumes for periods extending even to 30 days, in order to assess the impact of remote source areas. FLEXPART takes into account not only grid scale wind (as simple trajectory models do) but also turbulent wind fluctuations and mesoscale wind fluctuations. It also incorporates drift correction (to prevent accumulation of computational particles released) and density correction (to account for the decrease of air density with height).

Trajectory models are subject to uncertainty arising from interpolation of sparse meteorological data, assumptions regarding vertical transport, observational errors, sub-grid scale phenomenon, turbulence, convection, evaporation, and condensation. Moreover, it should be taken into account that the shorter the run-time, i.e. the length of the trajectory, the better the accuracy. Different studies agree that errors of 20% of the distance travelled seem to be typical for trajectories computed from analyzed wind fields (Rolph and Draxler, 1990; Stohl, 1998). Average horizontal trajectory errors were estimated to be 140-290 km in 24 h (Polissar et al., 2001). Stohl (Stohl et al., 1998a) investigated a 1-year set of terrain-following 96 h back trajectories at a level of 800 m above ground terminating at T=24, T=48 and T=72 h. He found relative errors at the origin of the trajectories of 16, 26 and 36% of the travel distance, respectively. Users can feel fairly confident in the trajectory position out to 48 hours. Longer term trajectories are best used in a statistical sense to determine transport regimes (Rolph, 2008). Otherwise back-trajectory models cannot represent spatial variations of the air movement at short- or medium-scales unless high resolution data are available. In general, the higher the resolution of the data, the better the terrain will be modeled. However, the model simulation may not be appropriate in very complex local terrain and will also depend upon the complexity of the meteorological conditions for the simulation time. It should be stressed that trajectory computations only compute advection from the *u*,*v*,*w* wind components. No independent thermodynamic effects are considered. For this reason, under strong diabatic conditions, i.e. in the boundary layer or through regions of latent heat release, the model simulation of the trajectory position may be erroneous (Kahl, 1993; Draxler, 1996). Thus, selection of the back-trajectories starting heights is a key factor to take into account. Due to the fact that diabatic effects are neglected by back-trajectory models, starting heights must be taken at relatively high or medium levels far from the ground influence, to avoid also the influence of complex topography. Moreover, differences between the model topography and the real topography can be high and should also be assessed before starting trajectories computation (Stohl, 1998).

Trajectory Statistical Methods (TSM)

Back-trajectories have been very frequently used for analyzing air pollution measurements. However, they have been commonly misinterpreted as a direct meteorological link between a source region and a downwind receptor (e.g., Abasova, 2010; Beuck et al., 2011).

As denoted by Kahl (1993) it is unrealistic to interpret a single estimated trajectory as a direct meteorological link between two distant regions. Otherwise, trajectories can provide a useful qualitative estimate of the general pathway of an advected air mass. TSM allow simultaneous computational treatment of both air mass back-trajectories and time series of atmospheric pollutants. These time series can be constituted by concentrations of gaseous pollutants, acidic compounds in rain water, particulate matter (PM) at specific grain sizes (e.g., PM_{10}, $PM_{2.5}$, PM_1), chemical components of PM_x or specific source contributions of PM_x derived from source apportionment analysis.

The use of TSM can help to:

- Characterize synoptic meteorological situations which give rise to low and high concentration levels of atmospheric pollutants at a specific sampling site.
- Identify potential source regions of air pollutants which can be transported towards the sampling site.

Moreover, the use of TSM deploying large trajectory ensembles can significantly reduce the trajectory uncertainty generated by interpolation and truncation processes, low temporal or spatial resolution of wind data, or an inappropriate selection of the starting heights (Stohl, 1998; Lupu and Maenhaut, 2002).

An important aspect to bear in mind is that sampling periods of atmospheric pollutants usually range between some hours to several days (typically being 24 hours). For this reason, many authors computed several back-trajectories for each PM sample, frequently with a time interval of 6 hours along the sampling period, with the aim to characterize the prevailing winds over the receptor site. Then, the endpoints of all trajectories calculated during the sampling period, were combined into a single composed trajectory for any sample. Finally, there will be a single composed trajectory associated to each PM sample that will be used for the subsequent calculations regarding the TSM. However, it must be argued that the origin of air mass can change dramatically during the sampling period. With the aim to exclude those samples which were influenced by many different air mass origins, some techniques can be applied. It is recommended the computation of a centroid that represents the averaged flow of the complete group of back-trajectories for each sampling period:

$$\overline{\mathrm{X}_\mathrm{k}} = \frac{1}{n_m} \sum_{\mathrm{j}=1}^{\mathrm{n_m}} \mathrm{X}_{\mathrm{jk}} \qquad (18.3)$$

where $\overline{X}_k$ is the centroid of the sample k, n_m is the number of back-trajectories calculated during the sampling period of the sample k, and X_{jk} is back-trajectory j that was calculated during the sampling period of the sample k. $\overline{X}_k$ and X_{jk} each consist of the same number of endpoints, each containing a coordinate of latitude, longitude and altitude, that will depend on the trajectory run time defined by the user.

In order to assess quantitatively the deviation of any single back-trajectory from the corresponding centroid, some statistical parameters that measure differences between trajectories can be used (Stohl, 1998 and references therein). Between the possible options, it is worth emphasizing the Absolute Horizontal Transport Deviation (AHTD). It is defined as:

$$\mathrm{AHTD}(t) = \frac{1}{N}\sum_{n=1}^{N}\left([X_n(t) - x_n(t)]^2 + [Y_n(t) - y_n(t)]^2\right)^{1/2} \qquad (18.4)$$

where *N* is the number of trajectories used, *X* and *Y* are the locations of the test trajectories and *x* and *y* are the locations of some reference trajectories at travel time *t*. A similar measure can be defined in the vertical. When the AHTD is divided by the average length of the test and reference trajectories (the centroid), the parameter known as Relative Horizontal Transport Deviation (AHTD) is thus obtained. Jorba et al. (2004) concluded that when the RHTD between any individual trajectory and the centroid exceeded a pre-defined threshold (50%) the trajectory had a substantial different origin than the averaged flow represented by the centroid. Therefore, those composed trajectories in which more than a relatively high number of individual trajectories had different origins, could be discarded from the trajectory analysis. Salvador et al. (2010; 2016) rejected those PM samples whose composed trajectories had more than 40% of the individual trajectories exceeding 50% of the RHTD from the centroid.

Three TSM that have been extensively used in previous source apportionment studies of atmospheric pollutants, namely: Cluster Analysis of trajectories, Residence Time Analysis based on Potential Source Contribution Function, and Concentration Fields Analysis, which are presented hereafter.

Cluster Analysis (CA) of trajectories

CA is a statistical method used to examine a sample of objects and group them into sets (clusters), each comprising objects of highest similarity. Since trajectories tend to be clustered along certain paths according to persistent synoptic situations, it can be very useful performing CA using air masses trajectory endpoints as the clustering variables. Many clustering methods exists as it has been described by Kalkstein et al. (1987) and Gong and Richman (1995):

- Hierarchical Methods (agglomerative "there are as many clusters as there are objects" and divisive "all the objects begin in one cluster"):

1. Ward's minimum variance method.
2. The centroid method.
3. The average linkage method.

- Non-hierarchical Methods (choose a specified number of clusters k and then loops though all the objects in the data set):

1. k-means method.
2. Nucleated agglomerative method.
3. Rotated principal component method.

Each of the methods removes the subjectivity inherent in classification procedures to a certain extent, although leaving some decisions on the classifying subject. In practice, the non-hierarchical k-means CA is the most frequently used procedure. It is an iterative algorithm that partitions the data by comparing each object to each of the k cluster centers by a dissimilarity measurement, and comprises 4 basic steps (Owen, 2003):

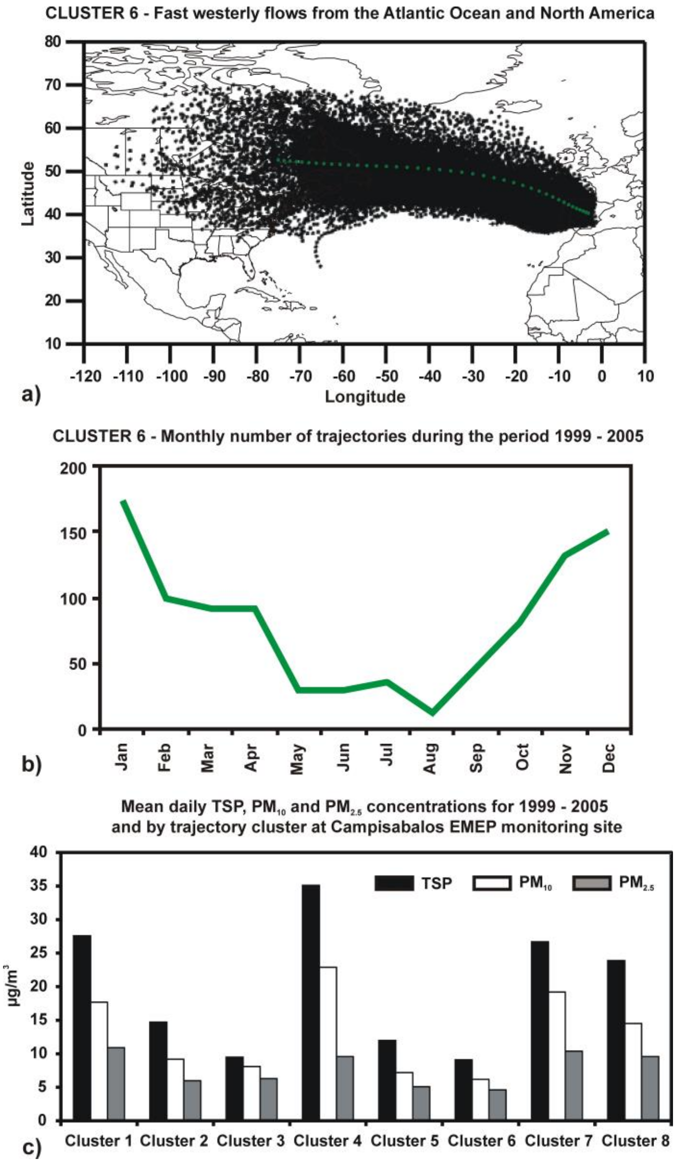
STEP 1: An initial partitioning of the trajectories is defined: *k* back-trajectories covering most regions of the study area are selected as initial seeds or "cluster centers". Subjectivity appears in the decision of how many clusters (k) should be retained for further analysis and in the selection procedure of the first *k* cluster centers (by a random choice from the data, by visual inspection, using the results of another clustering procedure…). It should be noted that k-means CA is sensitive to the selection of the initial "cluster centers".

STEP 2: calculate the change in clusterion criterion that result from changes in membership and reassign trajectories. The clustering criterion is a measure of the distance between two trajectories. The method of calculation of the clustering criterion must be selected (city-block distance, euclidean distance, …. Gong and Richman, 1995). Then, the distance from each trajectory *j* to each cluster-center *k* (d_{jk}) is calculated for every endpoint *i* along the trajectory (*n* endpoints) and summed. For the shake of an example the euclidean distance between endpoint *i* on trajectory *j* and endpoint *i* on cluster center *k* (ed_{ijk}) is used in equation (18.5) to compute d_{jk} as:

$$d_{jk} = \sum_{i=1}^{n} ed_{ijk} = \sum_{i=1}^{n} \sqrt{\left(x_{ij} - \overline{x_{ik}}\right)^2 + \left(y_{ij} - \overline{y_{ik}}\right)^2 + \left(z_{ij} - \overline{z_{ik}}\right)^2} \qquad (18.5)$$

where x_{ij} is the distance east/west or longitude from the origin to point i on trajectory j and $\overline{x}_{ik}$ is the corresponding distance to point i on cluster center k, y_{ij} and $\overline{y}_{ik}$ are similarly defined as distances north/south or latitude and z_{ij} and $\overline{z}_{ik}$ are similarly defined as the vertical distance or elevation (Owen, 2003 and references therein). The trajectory is assigned to the cluster with the smallest total distance from its cluster center.

Figure 18.2: Cluster of back-trajectories representing fast westerly flows from the Atlantic Ocean and North America towards the center of the Iberian Peninsula during the period 1999-2005 (a); Monthly number of back-trajectories grouped in this cluster (b); Mean daily PM concentrations registered at a regional-background monitoring site in central Spain (Campisábalos) by trajectory cluster (c). Adapted from Salvador et al. (2008).



STEP 3: recalculate the cluster centers after all the trajectories have been examined and assigned. The cluster centers are recalculated as the arithmetic mean of all members of any cluster using equation (18.3), where $\overline{X}_k$ is the cluster center of the cluster k, n_m is the number of members in cluster k and X_{jk} is trajectory j in cluster k.

STEP 4: repeat the steps 2 and 3 a specified number of times or until no trajectory or a specified number of trajectories changes its cluster assignment. The point at which no trajectories change cluster assignment is referred to as convergence.

Each cluster defines a characteristic meteorological synoptic situation for a given receptor site.

These meteorological scenarios have frequently specific seasonal patterns. Figure 18.2a shows the cluster 6 from the study of Salvador et al. (2008) that grouped fast-moving back-trajectories arriving to central Spain from the Atlantic Ocean and North America. This type of air flows was more frequently produced during the winter months (Figure 18.2b).

Mean *air pollutants* levels can thus be calculated for all trajectory clusters arriving at the sampling sites. *The non-parametric Kruskal-Wallis test can be used to test the significance of inter-cluster variation in pollutants* concentrations (Brankov et al., 1998). This technique tests the null hypothesis that several samples have been drawn from the same population. If the test leads to the rejection of the null hypothesis, it is interpreted as the levels of the atmospheric pollutants were influenced by the origin of air masses arriving at this site (i.e., by their transport paths or meteorological scenarios) which are represented by the clusters. To find out which clusters are significantly different from which others, several multiple sample comparison tests can be used. Figure 18.2c indicates that the regional background mean levels of Total Suspended Particles (TSP), PM_{10} and $PM_{2.5}$ concentrations in central Spain associated with fast moving maritime air masses (cluster 6) were significantly lower than in any of the other air mass transport patterns. This atmospheric transport situation generated high precipitation rates and strong winds producing a drop in surface PM concentration levels.

CA is a useful “objective” methodology for back-trajectories classifications that has worked very well for discriminating distinct flow patterns and large-scale circulation features in many air pollution studies (Abdalmogith and Harrison, 2005; Brankov et al., 1998; Cape et al., 2000; Dorling et al., 1992; Salvador et al. 2008; 2010) but also in combination with receptor models in a few previous source apportionment studies. For instance, Lee et al. (2002) used trajectories CA on days of high- and low- contributions of sources identified with PMF that contributed to $PM_{2.5}$ mass concentrations in northeastern USA, with the aim to link source types to potential source regions. Baudic et al. (2016) used trajectories CA to interpret a source apportionment analysis of VOCs performed with PMF in Paris. Salvador et al. (2016) carried out a cluster analysis to back trajectories starting at different heights (500 and 1500 m agl) from the Cape Verde archipelago and used the results to interpret source categories of PM_{10} identified by PMF.

Residence Time Analysis (RTA) based on Potential Source Contribution Function (PSCF)

This technique makes use of the Potential Source Contribution Function (PSCF) to determine the geographic origin and transport pathways of air masses that arrive at a receptor site under any given pollution scenario. It allows the establishment of a connection between atmospheric pollutants sources and the most probable origin of the air mass sampled using back-trajectory information. PSCF values are computed essentially in the same way than CPF values (Ashbaugh et al., 1985). PSCF and CPF have been indistinctly used for performing RTA using back-trajectories in many published papers (Zeng and Hopke, 1989; Lupu and Maenhaut, 2002; Salvador et al, 2004 among others). For the sake of simplicity, only the term PSCF will be used in this section. Once a grid has been created over the region of study, the PSCF for the ijth cell is defined as:

$$PSCF_{ij} = m_{ij} / n_{ij} \qquad (18.6)$$

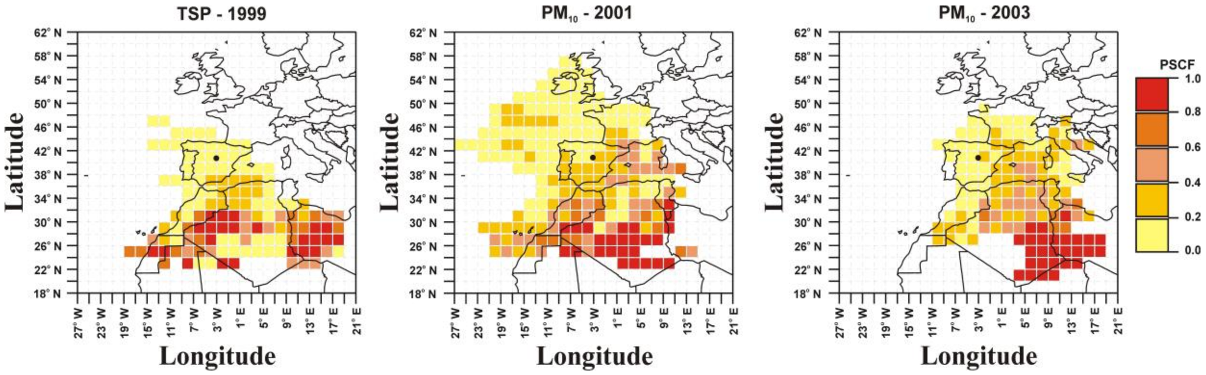
Where n_{ij} = number of endpoints from all the trajectories computed during the sampling period, residing or contained in the grid cell ij, and m_{ij} = number of endpoints from all the trajectories computed during “high incident days” residing or contained in the grid cell ij.

“High incident days” are defined as days with atmospheric pollutants concentrations values higher than a pre-specified given value. The criteria for determining them depends on the structure of the source contribution time series for each factor. For sources with relatively low concentrations for most of the study but displaying a few concentration peaks (much greater than the background concentrations) on some days, the 90th percentile of the total data set could be used. On the other hand, for sources with very frequent high concentration levels, the 75th percentile could be used (Pekney et al., 2006). Equation (18.6) represents the probability that an air mass arrives at the receptor site during “high incident days”, after having been observed to reside in a specific geographic area (the one contained in the grid cell). In order to check the statistical significance of every $PSCF_{ij}$ value, different tests have been proposed:

- A "point filter". Namely, grid cells counting less than a pre-specified number of endpoints or trajectories are not taken into account (Charron et al., 1998). It is very subjective. Another approach was to assign weights to the $PSCF_{ij}$ values depending on the number of endpoints falling into the in the ij^th grid cell (Zeng and Hopke, 1989; Zíková et al., 2016). The higher the number of endpoints, the higher the weight. It should be noted that the weighing values were arbitrary settings.
- The "binomial test" which assumes that the PSCF is the parameter p of the binomial distribution (Vasconcelos et al., 1996).
- Using bootstrapping techniques (Hopke et al., 1995).

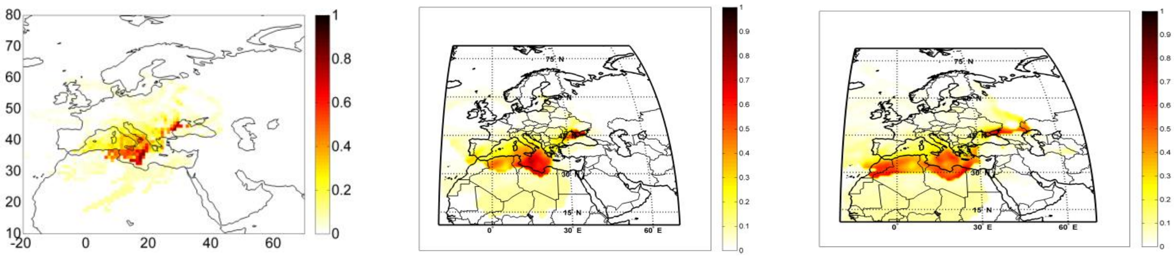
PSCF results can be reported on geographical maps. Hence, areas with high statistically significant PSCF values can be considered as source regions of the atmospheric pollutant under study or as preferred air mass pathways. To assess quantitatively the PSCF values, many authors have used the criteria proposed by Zeng and Hopke (1989). These criteria interpret in a qualitative way the possibility of each grid cell as being a source of the specific elements, according to the PSCF values ranges. According to them, PSCF values in the 0.0 - 0.2 range means very weak, 0.2 - 0.4 weak, 0.4 - 0.6 intermediate and higher than 0.6 means very strong source of the specific compounds. Results of a RTA/PSCF analysis, performed at central Spain, are showed in Figure 18.3 for examples.

Figure 18.3. PSCF for days with TSP (a) and PM_{10} (b and c) concentrations values higher than the 90th percentile of the annual data sets at the EMEP regional background monitoring site of Campisábalos. Only significant PSCF values at the 95% confidence level are displayed. Adapted from Salvador et al. (2008).



In practice, PSCF analysis could be applied to HYSPLIT backtrajectories and FLEXPART residence time in each grid cell (sensitivity plumes) in order to determine possible source areas for each pollutant (see Figure 18.4 for examples).

Figure 18.4: HYSPLIT and FLEXPART (5 day and 10 day backward run) PSCF for Dust aerosol source. *Adapted from Perrone et al. (2018).*



5 days Hysplit PSCF: Dust mainly originates from Libya and Algeria, but there is also a source in the North-East. The hot spot over the Meditteranean sea could be attributed to the prefereble pathway of transport from Africa.

5 day FLEXPART PSCF: Dust aerosol originates from the North Coast of Africa. We observe many similarities to the HYSPLIT PSCF (origin from the African coast and a transport path over the Black Sea).

10 day FLEXPART PSCF: We can now see the actual origin of the dust aerosol from Africa, as in 5 day PSCF the hotspots were mainly in the Mediterranean Sea.

Concentration Fields Analysis (CFA):

This analysis was originally proposed by Seibert et al. (1994) to identify potential source areas of air pollutants by creating concentration fields (CF) maps. It comprises several steps:

STEP 1: for each cell ij of the grid which was superimposed to the domain of the trajectory calculations, a logarithmic mean concentration is calculated:

$$CF_{ij} = \frac{\sum_{l=1}^{L} n_{ijl} \cdot log C_l}{\sum_{l=1}^{L} n_{ijl}} \qquad (18.7)$$

where (i, j) = index of the ijth grid cell, l = index of the trajectory, L = total number of trajectories, C_l = concentration of the air pollutant observed on arrival of the trajectory l, and n_{ijl} = time spent in the ijth grid cell ij by trajectory l (i.e., the number of endpoints of the trajectory l contained in the ijth grid cell).

Equation (18.7) is a weighted mean concentration of CF values in grid cells, in which the weight for each C_l value is the time spent in the ijth grid cell by trajectory l. Since atmospheric pollutants concentrations are frequently log-normally distributed, time series are transformed into a normally distributed form because it is required for standard statistics. It allows computing a confidence interval for the mean value CF_{ij}.

STEP 2: a confidence interval of the 100(1- α)% for the mean value CF_{ij} is calculated as follows:

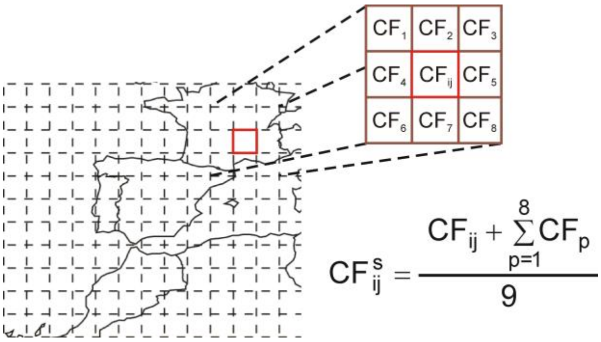
$$CF_{ij} \pm t_{(1-\frac{\alpha}{2},n-1)}(\frac{SD_{ij}}{\sqrt{n_{ij}}}) \qquad (18.8)$$

where SD_{ij} = standard deviation of CF_{ij}, n_{ij} = total number of trajectories passing through the grid cell ij, $t_{(a,b)}$ = value of the a^{th} percentile of the t distribution with b degrees of freedom, and α = significant level.

STEP 3: The concentration field is smoothed with a 9-point filter, imposing the restriction that the values must be kept within their confidence interval. A 9-point filter is a low pass filter used for digital image data treatment. It removes the local variability and the random noise and consequently the overall pattern becomes more clearly apparent (Mather, 2004). For each grid cell ij a smothed CF_{ij}^S value is calculated as the average between the concentration values in the cell itself and the neighboring cells according to equation (18.9) and Figure 18.5:

$$CF_{ij}^S = \frac{CF_{ij}+\sum_{p=1}^{8} CF_p}{9} \qquad (18.9)$$

Figure 18.5: Diagram illustrating the operation of a 9-point filter. At each stage equation (18.9) is applied on the central pixel CF_{ij} of a 3 x3 filter matrix to obtain the smoothed value CF_{ij}^S.



Stohl (1996) argued that the Seibert et al. (1994) procedure does not extract all the information from the data, because the measured concentration is equally attributed to all endpoints of the related trajectory. Since real sources of air pollutants are often concentrated in “hot spots”, this method underestimates the gradients of the “true” CF. Stohl (1996) proposed using the initial CF obtained with the Seibert et al. (1994) method to redistribute the concentrations along the trajectories in such a way that the concentration for the endpoint p of the trajectory l, C_{pl} could be estimated as:

$$C_{pl} = C_l \cdot {CF_{ijpl}} \Big/ {\overline{CF_{ijpl}}} \qquad (18.10)$$

where C_l = concentration observed on arrival of the trajectory l, CF_{ijpl} = concentration of the grid cell ij from the initial CF, where the endpoint p of the trajectory l resided, $\overline{\mathrm{CF}}_{\mathrm{ijpl}}$

= average of the concentrations of the grid cells in the initial CF, hit by the N_l endpoints of the trajectory l, and p = index of the endpoints of the trajectory l (p = 1, …, P).

This advanced approach is commonly known as Redistributed Concentration Fields (RCF).

CF and RCF results are also reported on geographical maps showing the weighted concentrations in the grid cells (Figure 18.6).

STEP 4: After the redistribution is applied to all trajectories, a new CF can be computed with:

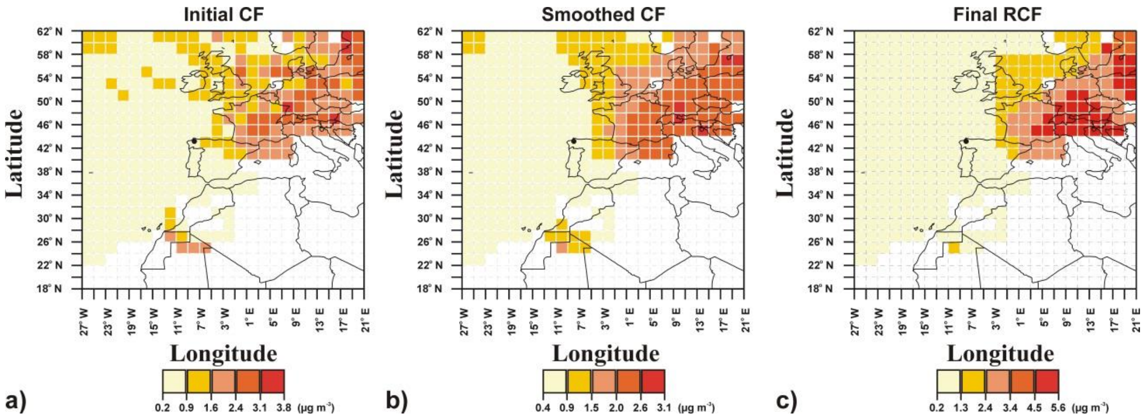
$$RCF_{ij} = \frac{\sum_{l=1}^{L}\sum_{p=1}^{P} n_{ijpl} \cdot logC_{pl}}{\sum_{l=1}^{L}\sum_{p=1}^{P} n_{ijpl}} \qquad (18.11)$$

n_{ijpl} = is the residence time in the grid cell ij of the endpoint p of the trajectory l.

STEP 5: The concentration field is smoothed with a 9-point filter again, imposing the restriction that the values must be kept within their confidence interval (Figure 18.6c).

Note that the smoothing procedure of the map by the 9-point filter assures that significant variations are preserved while most of the insignificant ones are removed (Figure 18.6a and 18.6b). Moreover, when using weighted concentrations, regions with weighted concentrations in the higher and lower value ranges indicated that, on average, air parcels residing over them resulted in high and low concentrations, respectively, of the atmospheric pollutant contributions at the receptor sites. These maps show those potential source areas whose emissions can be transported to the measurement site by prevailing synoptic winds.

Figure 18.6: CF for NO_3^- in PM_{10} registered in Bemantes (NW Spain) corresponding to back-trajectories starting at 1500 m AGL. Initial CF (a), Initial CF smoothed by a 9-point filter (b) and Final RCF smoothed by a 9-point filter (c). *Adapted from Salvador et al. (2007).*



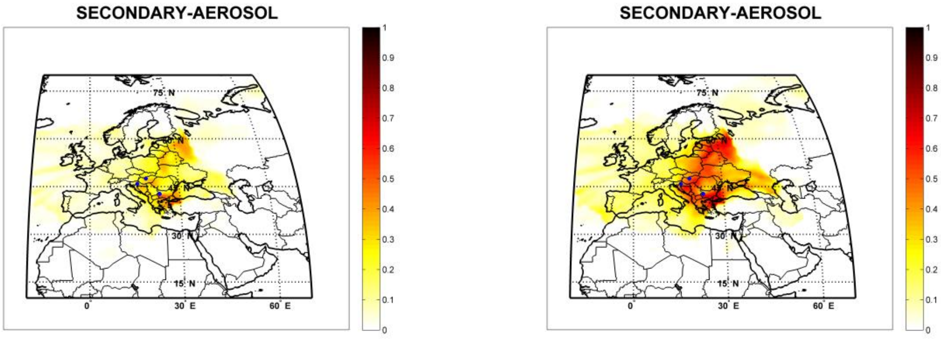
It must be highlighted that CF has been extensively applied in the last 15 years by many authors but most of them applied partially the aforementioned methods. For the shake of an example CF maps have been frequently computed using raw concentration data instead logarithmic data (Avila and Alarcón, 2003). Hsu et al. (2003) called Concentration Weighted Trajectory (CWT) to this method. It should be noted that this option prevents the confidence intervals of the CF_{ij} values from computing with equation (18.8). Thus, the statistical significance of the CFij values can not be checked with the application of the 9-point filter operator and the restriction that the values must be kept within the confidence intervals. The most frequently used test to check the statistical significance of the PSCF and CF maps has been the "point filter" (Hsu et al., 2003). As it was mentioned before it is based on highly arbitrary settings. For this reason, its use should be avoided as much as possible.

Further comments on the role of TSM in the identification of air pollution sources

As detailed above, TSM allow for identifying potential source regions of atmospheric pollutants and preferred transport pathways of air masses at receptor sites not strongly impacted by local emission sources. However, it is worth bearing in mind that TSM such as PSCF and CF can help obtaining the average contributions of source regions to the concentration at the receptor site rather than the true distribution of sources (Seibert and Jost, 1994). The term "potential source regions" should be preferably used due to the fact that influences of vertical exchange, chemical transformation and scavenging are not taken into account. The users of TSM should be aware that real source regions are at best a subset of these, which can be identified with additional knowledge. The interpretation of the PSCF, CF and RCF maps must be achieved on the basis of the existence of true emission sources over the potential source areas and the occurrence of specific atmospheric circulations that caused their transport. For this reason, it is advisable performing additionally a CA for back-trajectories arriving at different heights and characterizing the meteorological scenarios causing the synoptic flows associated to each cluster over the region of study. Otherwise, Vinogradova (2000) showed that the most polluting source contributing to a receptor site by medium- or long-range transport, may not be the one with the highest levels of emission. It will depend on the specific air mass transport patterns prevailing over the receptor site and the residence time of the air masses over the potential source areas before arriving the receptor site. Lupu and Maenhaut (2002) suggested calculating maps of probabilities or concentrations by making use of data from more than one receptor site, the more, the better, with the aim to improve the spatial resolution of the source areas. This option has been successfully applied by Perrone et al., (2018), using a methodology similar to the one reported by Han et al. (2007) to detect secondary aerosol source areas for the Danube region. A multi-site PSCF analysis was carried out for this region based on results from Sofia, Zagreb and Budapest (Figure 18.7).

Intercomparison studies (Charron et al., 1998; Hsu et al, 2003) have showed that all these techniques correctly identify the same potential source regions of atmospheric pollutants. Scheifinger and Kaiser (2007) validated various TSM, including PSCF, CF and RCF, by different approaches, achieving the best performance with the RCF method. Besides, Wotawa and Kröger (1999) successfully tested the ability of the RCF method to reproduce emission inventories of air pollutants.

Figure 18.7: Multi-site PSCF analysis for the Danube region based on results from Sofia, Zagreb and Budapest (Secondary aerosol). Secondary aerosol mainly originates from sources of European Russia and Turkey. At the 75th percentile, the identified source areas included not only Northern and North-Eastern areas from the Danube region, but also areas from central Europe and the Danube region itself. Adapted from Perrone et al., 218.

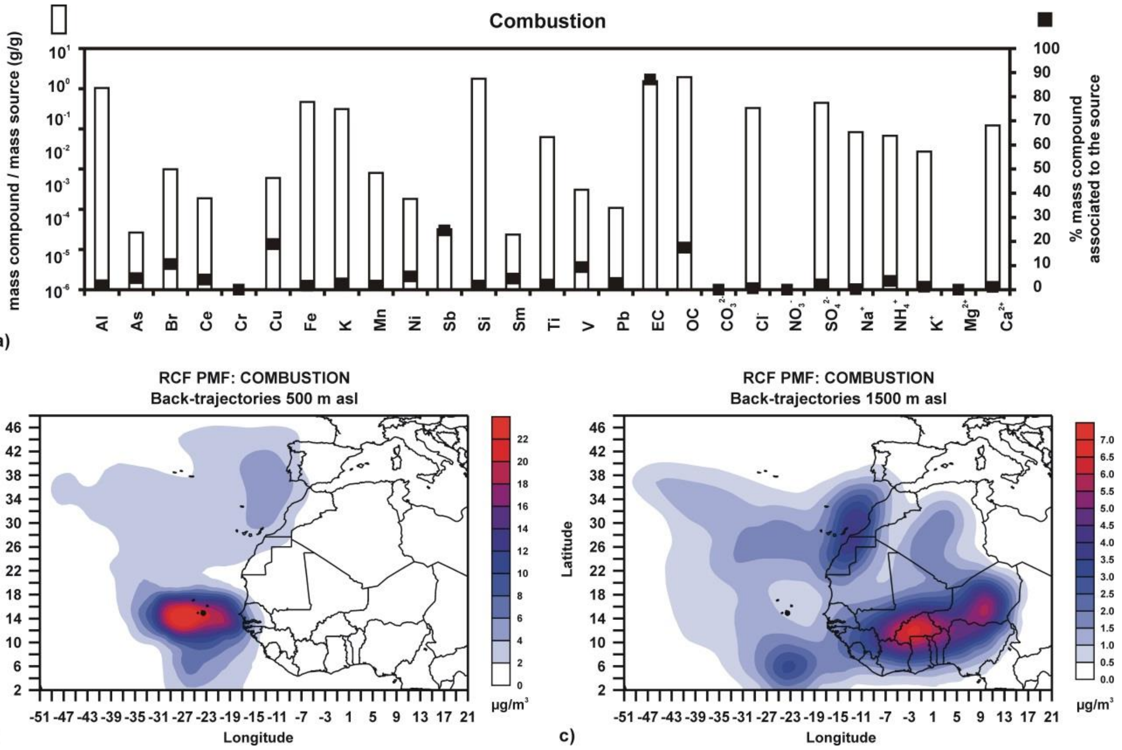


Combined Secondary aerosol PSCF analysis for Sofia, Zagreb, and Budapest at the 90th percentile, 5 days backward run.

Combined Secondary aerosol PSCF analysis for Sofia, Zagreb, and Budapest at the 75th percentile, 5 days backward run.

Some more examples of the advantage of applying TSM to receptor modeling results are given in the following. Polissar et al. (2001) used PMF to identify the main sources of $PM_{2.5}$ at Vermont (USA) and estimate their mass contributions. RTA results indicated a strong regional character to the upwind locations associated with daily source contributions from most of the $PM_{2.5}$ sources identified by PMF. Salvador et al. (2004) identified the main sources of PM_{10} at an urban-traffic site in Madrid (Spain) by means of a Varimax rotated factor analysis. Then, they computed CPF maps for the elements that showed the highest factor loadings for each source, i.e. the best source tracers.

Figure 18.8: RCF for the PM_{10} source “Combustion” obtained with PMF (a) corresponding to back-trajectories arriving at 500 (b) and 1500 (c) m asl over the Cape Verde archipelago. Adapted from Salvador et al. (2016).



External contributions from long-range transport processes of African dust and marine aerosol with a high content of Al - Ti and Na, respectively, could be detected. Salvador et al. (2007) used a similar procedure (Varimax rotated factor analysis plus RCF maps) to detect contributions of NO_3^- from anthropogenic sources of central and eastern Europe to the PM_{10} and $PM_{2.5}$ levels at a background coastal site in NW Spain. Another interesting application of the analysis of back-trajectories sets computed at different heights for interpreting receptor modelling results can be found in Salvador et al. (2016). They computed RCF maps for specific PM_{10} source contributions identified by PMF using back trajectories arriving at different heights (500 and 1500 m asl) at the Cape Verde archipelago. They helped to identify different source areas for the same source category. For a source profile denoted as “Combustion” (Figure 18.8a), the RCF map corresponding to back trajectories starting at 500 m asl showed the regions over and surrounding the archipelago as the main potential source areas (Figure 18.8b). It was attributed to combustion emissions from local-regional sources (mainly traffic) in summer when slow air flows due to weak baric gradients predominantly occurred. In contrast, the RCF map corresponding to back trajectories starting at 1500 m asl identified a big area over the sub-Sahelian West African corridor as the main potential source region (Figure 18.8c). It was interpreted as the long-range transport of the carbonaceous aerosols that are produced from intense wildfires occurring in this area during the October-April period every year.

Generally, it may be concluded that the application of TSM as a supplementary tool can be very useful in the studies aimed to characterize potential sources and formation processes of atmospheric pollutants. In fact, Poirot et al. (2001) recommended the simultaneous application of source contributions techniques and other TSM as a useful approach for improving the understanding of source-receptor relationships for PM, for improving the confidence in the individual model results and for developing a better understanding of the underlying aerosol data. However, TSM remain barely used in combination with other receptor modelling techniques, in source apportionment studies (Viana et al., 2008; Belis et al., 2013). This is probably because some basic computer programming skills are required to manage simultaneously back-trajectories and atmospheric pollutants concentrations data files and to create basic scripts for developing the algorithms required to perform the different TSM. The development of new user-friendly toolkits, such as Openair or ZeFir should help bridging this gap in near-future. Otherwise, a geographical information system (GIS) is needed to present TSM results in a clear and useful manner.

Classifications of atmospheric circulation patterns

A circulation pattern means a field of a meteorological variable describing atmospheric circulation, defined for each time instant of the analysis on a regular longitude-latitude grid. The majority of circulation patterns are represented by gridded sea level pressure and/or geopotential heights in lower to middle troposphere (up to 500 hPa). Classification of long data sets of circulation patterns are referred to “circulation classification” and individual groups as “circulation types” or “synoptic types”. Unlike weather-type classification, circulation classification is usually based on just one of the aforementioned meteorological parameters. The inclusion of additional levels yields only little extra information, due to the high degree of dependence among individual levels (Huth et al., 2008).

Different methods have been used for classification of circulation patterns (Huth, 1996 and references therein; Philipp et al., 2014) being cluster analysis-CA and principal component analysis-PCA the most widely used.

In this case of the CA, the grid-point values of the meteorological variable for each field are used as the clustering variables. The Nonhierarchical K-means CA procedure follows the same steps described before for back-trajectories. The main difference when CA is carried out to group circulation patterns is that it is commonly preceded by PCA for data reduction. PCA allows converting the input variables into linearly independent components. The number of variables necessary to account for the bulk variance of the data is thus reduced. Hence, PCA in S-mode is used, that is, the columns of the input data matrix refer to grid points and the rows correspond to individual time observation. As a result, a large number of grid-point values at each time are reduced to several PC scores, sufficiently representing circulation on each day. Then the PCs scores are used as the clustering variables.

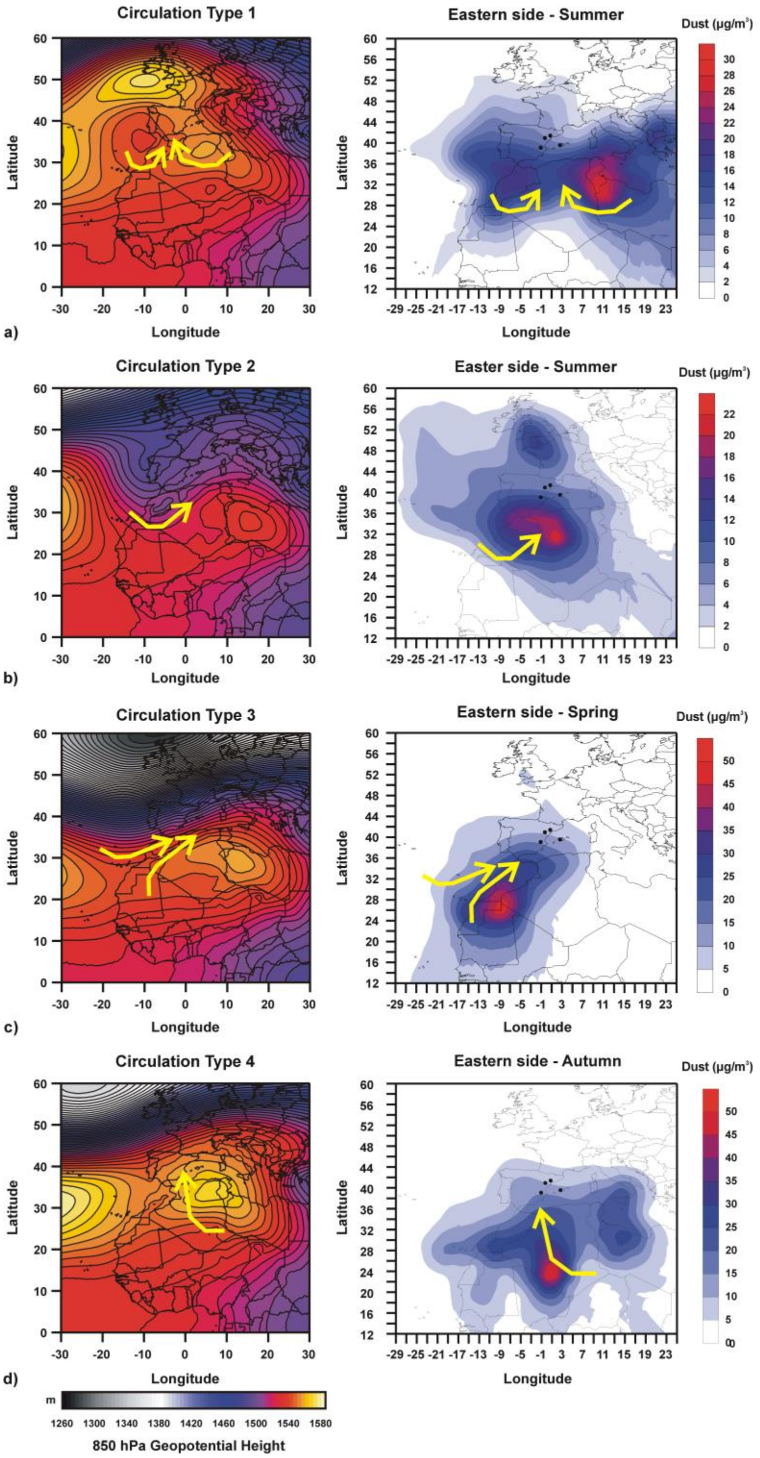
The principal component analysis classification technique requires that a PCA in T-mode must be used. In this case the columns of the input data matrix represent time observations and the rows correspond to grid points. Hence, the resulting PCs reproduce the main circulation types. The similarity of each day's circulation pattern to a PC is expressed by the corresponding loading, i.e., the higher the loading, the greater similarity.

Nonhierarchical K-means CA is regarded the best procedure if the preferred property is the separation among clusters as well as between clusters and the whole data set (Huth, 1996). Additionally, Huth (1996) and Huth et al. (2008) demonstrated that in comparison with other classification methodologies, K-means provides excellent separability among cluster, good temporal and spatial stability and a moderate ability to reproduce known underlying structure of data. Otherwise, PCA is the most accurate method for circulation pattern classification, if the goal of the study was centered on the ability to reproduce known patterns (Gong and Richman, 1995, Huth, 1996). However, these authors also demonstrated that there is not a classification method which is best in all aspects among other tested. In the end, each of the methods removes the subjectivity inherent in classification procedures to a certain extent, although leaving some decisions on the classification subject. Subjectivity appears in both methods in the decision of how many components should be retained, in the rotation procedure for the PCA and if "raw" data or "anomalies" (deviations from the mean period for each day) should be used. Other problems may appear associated to the fact that the fields of atmospheric variables used in these studies are so big, that many standard statistical software do not support them. In that case, programs for statistical computing and graphics created by users can be used as well as open source applications. In recent years, the cost733class open source software was developed during the COST733 Action for creating, comparing and evaluating circulation classifications (Philipp et al., 2014). It should be stressed that this FORTRAN software package runs on UNIX/LINUX.

A composite synoptic map representing each resultant circulation type can be calculated by averaging data fields of meteorological variables that describes atmospheric circulations (sea level pressure and geopotential height at 1000, 850 or 500 hPa levels). These type of information can be obtained free of charge for individual dates or user's defined data ranges from the NCEP/NCAR (NOAA, USA) Reanalysis datasets files (https://www.esrl.noaa.gov/psd/data/composites/hour/, accessed September 2018; Kalnay et al., 1996). Thus, composite synoptic maps can be extracted and used to highlight the fluctuations of time series of levels of different sources contributions at specific sites. Circulation classifications have been derived over specific regions to study atmospheric transport of air pollutants (Russo et al., 2014; Valverde et al., 2014). Circulation classification techniques and TSM were used in the study of Salvador et al. (2014) to determine the main circulation patterns that produced the transport of desert dust to the western Mediterranean area and the source areas associated to each of them. The study was based on the analysis of time series of African dust contributions to the PM_{10} levels at regional background monitoring sites, which were estimated by a specific receptor modelling methodology (Escudero et al., 2007). Figure 18.9 illustrates how depending on the circulation type that produces the advection of African dust towards the eastern side of the Iberian Peninsula and the Balearic Islands, the source area of dust was different. This fact can strongly influence the composition and amount of the transported dust.

It should be noted that classifications of atmospheric circulation patterns using sea level pressure data fields can be very useful for discriminating periods under stagnant meteorological conditions. They are produced by the stationary presence of high pressure systems which gives rise to highly stable atmospheric conditions and deep surface thermal inversions. It results in slow air flows and a thin surface atmospheric layer where pollutants accumulate day after day in urban areas. Under these synoptic types, the typical autumn-winter air pollution episodes are produced in European cities (Artíñano et al., 2003). These periods of high atmospheric stability in which the surface layer is decoupled from higher atmospheric layers, are hardly characterized by trajectories computed by numerical models.

Figure 18.9: Characterization of atmospheric circulation patterns that gave rise to African dust outbreaks over the eastern side of the Iberian Peninsula and the Balearic Islands in 2001-2011 (left column). Associated dust source areas were identified by computing RCF for African dust during episodes occurring in different seasons (right column). Adapted from Salvador et al. (2014).



References

Abbasova T., 2010. Detection and analysis of changes in desertification in the Caspian Sea Region. Master Thesis. Stockholm University, Faculty of Science, Department of Physical Geography and Quaternary Geology (INK), Stockholm University. http://su.diva-portal.org/smash/record.jsf?searchId=1&pid=diva2:354905; access 10.01.2012

Abdalmogith S.S., Harrison R.M., 2005. The use of trajectory cluster analysis to examine the long-range transport of secondary inorganic aerosol in the UK. Atmospheric Environment, 39, 6686–6695.

Amato, F., Hopke, P. K., 2012. Source apportionment of the ambient PM2.5 across St. Louis using constrained positive matrix factorization, *Atmospheric Environment*, 46(0), 329–337, doi:10.1016/j.atmosenv.2011.09.062.

ARL: Air Resouces Laboratory, http://www.arl.noaa.gov/ (accessed September 2018)

Artíñano B., Salvador P., Alonso D.G., Querol X. and Alastuey A., 2003. Anthropogenic and natural influence on the PM_{10} and $PM_{2.5}$ aerosol in Madrid (Spain). Analysis of high concentration episodes. *Environmental Pollution*, 125, 453-465.

Ashbaugh, L. L., Malm, W. C., Sadeh, W. Z., 1985. A Residence Time Probability Analysis of Sulfur Concentrations at Grand Canyon National Park. *Atmospheric Environment* - Part A General Topics 19 (8),1263-1270.

Avila, A., Alarcón, M., 2003. Precipitation chemistry at a rural Mediterranean site: Between anthropogenic pollution and natural emissions. *Journal of Geophysical Research*, Vol. 108 (D9), 4278, doi: 10.1029/2002JD002565.

Baudic A., Gros V., Sauvage S., Locoge N., Sanchez O., Sarda-Estève R., Kalogridis C., Petit J.E., Bonnaire N., Baisnée D., Favez O., Albinet A., Sciare J., Bonsang B., 2016. Seasonal variability and source apportionment of volatile organic compounds (VOCs) in the Paris megacity (France). *Atmospheric Chemistry and Physics*, 16, 11961–11989.

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. *Atmospheric Environment*, 69, 94-108.

Beuck, H., Quass, U., Klemm, O., Kuhlbusch, T.A.J., 2011. Assessment of sea salt and mineral dust contributions to PM10 in NW Germany using tracer models and positive matrix factorization. *Atmospheric Environment*, 45, 5813-5821

Brankov E., Rao S.T., Porter P.S., 1998. A trajectory-clustering-correlation methodology for examining the long-range transport of air pollutants. *Atmospheric Environment*, 32, 1525-1534.

Cape J.N., Methven J., Hudson L.E., 2000. The use of trajectory cluster analysis to interpret trace gas measurements at Mace Head, Ireland. *Atmospheric Environment*, 34, 3651–3663.

Carslaw D.C., Ropkins K., 2012. Openair - an R package for air quality data analysis. *Environmental Modelling & Software*, 27-28, 52–61.

Charron A., Plaisance H., Sauvage S., Coddeville P., Galloo J.C., Guillermo R., 1998. Intercomparison between three receptor-oriented models applied to acidic species in precipitation. *The Science of the Total Environment*, 223, 53-63.

Dorling S.R., Davies T.D., Pierce C.E., 1992. Cluster analysis: a technique for estimating the synoptic meteorological controls on air and precipitation chemistry – method and applications. *Atmospheric Environment*, 26A, 2575-2581.

Dimitriou, K., Kassomenos P., 2016. Combining AOT, Angstrom Exponent and PM concentration data, with PSCF model, to distinguish fine and coarse aerosol intrusions in Southern France. *Atmospheric Research*, 172, 74-82.

Draxler R.R., 1996. Boundary layer isentropic and kinematic trajectories during the August 1993 North Atlantic Regional Experiment Intensive. *Journal of Geophysical Research*, 101, D22, 29255-29268

Draxler R.R., Hess G.D., 1998. An overview of the HYSPLIT_4 modelling system for trajectories, dispersion and deposition. Australian Meteorological Magazine, 47, 295-308Draxler, R.R. and Rolph, G.D., 2012. HYSPLIT (HYbrid Single-Particle Lagrangian Integrated Trajectory) Model access via NOAA ARL READY Website (http://ready.arl.noaa.gov/HYSPLIT.php). NOAA Air Resources Laboratory, Silver Spring, MD.

Escudero M., Querol X., Pey J., Alastuey A., Pérez N., Ferreira F., Alonso S., Rodríguez S., Cuevas, E., 2007. A methodology for the quantification of the net African dust load in air quality monitoring networks. *Atmospheric Environment*, 41, 5516-5524.

Eleftheriadis K., Vratolis S., Nyeki S., 2009. Aerosol black carbon in the European Arctic: Measurements at Zeppelin station, Ny-Ålesund, Svalbard from 1998–2007. *Geophysical Research Letters*, 36: L02809, doi: 10.1029/2008GL035741.

Gong X., Richman M.B., 1995. On the application of cluster analysis to growing season precipitation data in North America east of the Rockies. *Journal of Climate*, 8, 897-931

Han Young-Ji, Thomas M. Holsen, Philip K. Hopke, 2007. Estimation of source locations of total gaseous mercury measured in New York State using trajectory-based models, Atmospheric Environment, 41, 28, 6033-6047.

Henry, R.C., 2002. Multivariate receptor models—current practice and future trends. *Chemometrics and Intelligent Laboratory Systems* 60, 43–48.

Henry, R.C., Changa, Y-S., Spiegelman, C.H., 2002. Locating nearby sources of air pollution by nonparametric regression of atmospheric concentrations on wind direction. *Atmospheric Environment* 36, 2237–2244.NILU: http://transport.nilu.no/flexpart

Hopke P.K., Li C.L., Ciszek W., Landsberger S., 1995. The use of bootstrapping to estimate conditional probability fields for source locations of airborne pollutants. *Chemometrics*, Intelligent Laboratory Systems, 30, 69-79.

Hsu, Y., Holsen, T.M., Hopke, P.K., 2003. Comparison of hybrid receptor models to locate PCB sources in Chicago. *Atmospheric Environment*, 37, 545–562.

Huth R., 1996. An intercomparison of computer-assisted circulation classification methods. *International Journal of Climatology*, 16, 893-922.

Huth R., Beck C., Philipp A., Demuzere M., Ustrnul Z., Cahynová M., Kyselý J., Tveito O.E., 2008. Classifications of atmospheric circulation patterns: recent advances and applications. *Annals of the New York Academy of Sciences*, 1146, 105-152

Jeong, C. H., McGuire, M. L., Herod, D., Dann, T., Dabek-Zlotorzynska, E., Wang, D., Ding, L., Celo, V., Mathieu, D., Evans, G., 2011. Receptor model-based identification of PM2.5 sources in Canadian cities, *Atmospheric Pollution Research*, 2(2), 158–171.

Jorba O., Pérez C., Rocadenbosch F., Baldasano J.M., 2004. Cluster analysis of 4-day back trajectories arriving in the Barcelona area, Spain, from 1997 to 2002. *Journal of Applied Meteorology*, 43, 887-901.

Kahl J., 1993. A cautionary note on the use of air trajectories in interpreting atmospheric chemistry measurements. *Atmospheric Environment*, 27A, 3037-3038.

Kalnay E., Kanamitsu M., Kistler R., Collins W., Deaven D., Gandin L., Iredell M., Saha S., White G., Woollen J., Zhu Y., Chelliah M., Ebisuzaki W., Higgins W., Janowiak J., Mo K.C., Ropelewski C., Wang J., Leetmaa A., Reynolds R., Jenne R., Joseph D., 1996. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77: 437-470.

Kalkstein L.S., Tan G., Skindlov J.A., 1987. An evaluation of three clustering procedures for use in synoptic climatological classification. Journal of Climate and Applied Meteorology, 26, 717-730.Kim, E., Hopke, P.K., Paatero, P., Edgeton, E.S., 2003. Incorporation of Parametric Factors into Multilinear Receptor Models Studies of Atlanta Aerosol. *Atmospheric Environment* 37, 5009-5021

Kim, E., Hopke, P. K., Larson, T. V., Covert, D. S., 2004. Analysis of ambient particle size distributions using unmix and positive matrix factorization, *Environmental science & technology*, 38(1), 202–209.

Kim, E. and Hopke, P.K., 2004. Comparison between conditional probability function and nonparametric regression for fine particle source directions. *Atmospheric Environment* 38 (28), 4667-4673

Lee J. H., Yoshida Y., Turpin B.J., Hopke P.K., Poirot R.L., Lioy P.J., Oxley J.C., 2002. Identification of sources contributing to mid-atlantic regional aerosol. *Journal of the Air & Waste Management Association*, 52:10, 1186-1205.

Lupu A., Maenhaut W., 2002. Application and comparison of two statistical trajectory techniques for identification of source regions of atmospheric aerosol species. *Atmospheric Environment*, 36, 5607-5618.

Mather P.M., 2004. Computer Processing of Remotely-Sensed Images: An Introduction, Third Edition. John Wiley &Sons, Ltd., 442 pp.

Owen R.C., 2003. Climatological study of transport to the PICO-NARE site using atmospheric backward trajectories. Master's Thesis, Michigan Technological University, 361 pp.

Pekney, N. J., Davidson, C. I., Zhou, L., and Hopke, P. K., 2006. Application of PSCF and CPF to PMF-modeled sources of PM2.5 in Pittsburgh. *Aerosol Science and Technology*, 40, 10, 952–961.

Petit J.-E., O. Favez, A. Albinet, F. Canonaco, 2017. A user-friendly tool for comprehensive evaluation of the geographical origins of atmospheric pollution: Wind and trajectory analyses. *Environmental Modelling & Software*, 88, 183-187.

Perrone, M. G., Vratolis, S., Georgieva, E., Torok, S., Sega, K., Veleva, B., Osan, J., Beslic, I., Kertesz, Z., Pernigotti, D., Eleftheriadis, K., Bellis, C. A., 2018. Sources and geographic origin of particulate matter in urban areas of the Danube macro-region: the cases of Zagreb (Croatia), Budapest (Hungary) and Sofia (Bulgaria). *Science of the Total Environment*, 619-620, 1515-1529.

Poirot R., Wishinski P., Hopke P.K., Polissar, A., 2001. Comparative application of multiple receptor methods to identify aerosol sources in Northern Vermont. *Environmental Science & Technology*, 35, 4622-4636

Polissar, A V, Hopke, P K, Harris, J M., 2001. Source regions for atmospheric aerosol measured at Barrow, Alaska. *Environmental Science & Technology*, 35, 4214-4226

Prospero, J.M., Ginoux, P., Torres, O., Nicholson, S.E., Gill, T.E., 2002. Environmental characterization of global sources of atmospheric soil dust identified with the Nimbus 7 Total Ozone Mapping Spectrometer (TOMS) absorbing aerosol product. *Reviews of Geophysics*, 40 (1), 21-31

Querol, X., Pey, J., Pandolfi, M., Alastuey, A., Cusack, M., Perez, N., Moreno, T., Viana, M., Mihalopoulos, N., Kallos, G., and Kleanthous, S., 2009. African dust contributions to mean ambient PM10 mass-levels across the Mediterranean Basin, *Atmospheric Environment*, 43, 4266–4427.

RIU: the EURopean Air Pollution Dispersion (EURAD) Project: http://www.eurad.uni-koeln.de/index_e.html?/modell/eurad_descr_e.html

RIU: the EURopean Air Pollution Dispersion (EURAD) Project: http://www.eurad.uni-koeln.de/index_e.html?/modell/eurad_descr_e.html

Rolph G.D., Draxler R.D., 1990. Sensitivity of three-dimensional trajectories to the spatial, temporal densities of the wind field. *Journal of Applied Meteorology*, 29, 1043-1054.

Rolph, G.D., 2008. NOAA ARL HYSPLIT FAQs website (https://www.arl.noaa.gov/faq_hg14.php). Accessed 29 August 2017.

Rolph, G.D., 2012. Real-time Environmental Applications and Display sYstem (READY) Website (http://ready.arl.noaa.gov). NOAA Air Resources Laboratory, Silver Spring, MD.

Rolph, G.D., Stein A., Stunder B., 2017. Real-time Environmental Applications, Display sYstem: READY. *Environmental Modelling & Software*, 95, 210-228.

Russo A., Trigo R.M., Martins H., Mendes M.T., 2014. NO_2, PM_{10} and O_3 urban concentrations, its association with circulation weather types in Portugal. *Atmospheric Environment*, 89, 768–785.

Salvador P., Artíñano B, Alonso D, Querol X, Alastuey A., 2004. Identification and characterisation of sources of PM_{10} in Madrid (Spain) by statistical methods. *Atmospheric Environment*, 38, 435-447.

Salvador P., Artíñano B., Querol X., Alastuey A., Costoya M., 2007. Characterisation of local, external contributions of atmospheric particulate matter at a background coastal site. *Atmospheric Environment*, 41, 1-17.

Salvador P., Artíñano B., Querol X., Alastuey A., 2008. A combined analysis of backward trajectories and aerosol chemistry to characterise long-range transport episodes of particulate matter: The Madrid air basin, a case study. *Science of the Total Environment*, 390, 495-506.

Salvador, P., Artíñano, B., Pio, C., Afonso, J., Legrand, M., Puxbaum H., Hammer, S., 2010. Evaluation of aerosol sources at European high-altitude background sites with trajectory statistical methods. *Atmospheric Environment*, 44, 2316-2329.

Salvador, P., Alonso-Pérez, S., Pey, J., Artíñano, B., de Bustos, J.J., Alastuey, A., Querol, X., 2014. African dust outbreaks over the western Mediterranean Basin: 11-year characterization of atmospheric circulation patterns and dust source areas. *Atmospheric Chemistry and Physics*, 14, 6759–6775.

Salvador P., Almeida S.M., Cardoso J., Almeida-Silva M., Nunes T., Cerqueira M., Alves C., Reis M.A., Chaves P.C., Artíñano B., Pio C., 2016. Composition and origin of PM_{10} in Cape Verde: Characterization of long-range transport episodes. *Atmospheric Environment*, 127, 326-339.

Scheifinger H., Kaiser A., 2007. Validation of trajectory statistical methods. *Atmospheric Environment*, 41, 8846-8856.

Seibert P., Kromp-Kolb H., Baltenspenger U., Jost D.T., Schwikowski M., Kasper A., Puxbaum H., 1994. Trajectory Analysis of Aerosol Measurements at High Alpine Sites. A contribution to subproject ALPTRAC. The Proceedings of EUROTRAC Symposium '94 edited by P.M. Borrell et al., pp. 689-693. Academic Publishing, The Hague.

Seibert P., Jost D. T., 1994. Investigation of potential source areas by statistical trajectory analysis of ALPTRAC aerosol measurements. EUROTRAC Newsletter, 14, 14-17.

Seibert, P. and Frank, A., 2004. Source-receptor matrix calculation with a Lagrangian particle dispersion model in backward mod., *Atmospheric Chemistry and Physics*, 4, 51-63.

Squizzato S., M. Masiol, F. Visin, A. Canal, G. Rampazzo and Bruno Pavoni, 2014. The PM2.5 chemical composition in an industrial zone included in a large urban settlement: main sources and local background, *Environ. Sci.: Processes & Impacts*, 16.

Stohl, A., G. Wotawa, P. Seibert, H. Kromp-Kolb. 1995. Interpolation errors in wind fields as a function of spatial and temporal resolution and their impact on different types of kinematic trajectories. *J. Appl. Meteor.* 34, 2149-2165.

Stohl A., 1996. Trajectory statistics-a new method to establish source-receptor relationships of air pollutants, its application to the transport of particulate sulfate in Europe. *Atmospheric Environment*, 30, 579-587.

Stohl, A., and P. Seibert. 1998. Accuracy of trajectories as determined from the conservation of meteorological tracers. *Q. J. Roy. Met. Soc.* 124, 1465-1484.

Stohl A., 1998. Computation accuracy and applications of trajectories - A review and bibliography. *Atmospheric Environment*, 32, 947-966.

Stohl A., M. Hittenberger, and G. Wotawa, 1998. Validation of the Lagrangian particle dispersion model FLEXPART against large scale tracer experiments. *Atmospheric Environment*, 32, 4245-4264.

Stohl Stohl, A., and D. J. Thomson, 1999. A density correction for Lagrangian particle dispersion models. *Bound-Layer Met.*, 90, 155-167.

Stohl A., Eckhardt S., Forster C., James P., Spichtinger N., Seibert P., 2002. A replacement for simple back trajectory calculations in the interpretation of atmospheric trace substance measurements. *Atmospheric Environment* 36, 4635-4648.

Stohl, A., C. Forster, A. Frank, P. Seibert, G. Wotawa, 2005. Technical Note: The Lagrangian particle dispersion model FLEXPART version 6.2. Atmos. Chem. Phys. 5, 2461-2474.

TRUMF (Troposphärische UmweltForschung): TRAMPER Trajectories: http://www.geo.fu-berlin.de/en/met/ag/trumf/Trajektorien/index.html

Vasconcelos L., Kahl J., Liu D., Macias E., White W., 1996. A tracer calibration of back trajectory analysis at the Grand Canyon. *Journal of Geophysical Research*, 101 (D14), 19329–19335.

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., Wahlin P., Vecchi R., Miaranda A.I., Kasper-Giebl A., Maenhaut W., Hitzenberger R., 2008. Source apportionment of particulate matter in Europe: a review of methods and results. *Journal of Aerosol Science*, 39, 827-849.

Vinogradova A.A., 2000. Anthropogenic pollutants in the Russian Arctic atmosphere: sources and sinks in spring and summer. *Atmospheric Environment*, 34, 5151-5160.

Wang Y.Q., Zhang X. Y., Draxler R.R., 2009. TrajStat: GIS-based software that uses various trajectory statistical analysis methods to identify potential sources from long-term air pollution measurement data. *Environmental Modelling & Software* 24, 938.
download of product from http://www.meteothinker.com/TrajStatProduct.html

Wotawa G., Kröger H., 1999. Testing the ability of trajectory statistics to reproduce emission inventories of air pollutants in cases of negligible measurement and transport errors. *Atmospheric Environment*, 33, 3037-3043.

Zhao, W., Hopke, P.K., 2006. Source investigation for ambient PM 2.5 in Indianapolis, IN. *Aerosol Science and Technology*, 40(10), 898-909.

Zeng Y., Hopke P. K., 1989. A Study of the Sources of Acid Precipitation in Ontario, Canada. *Atmospheric Environment*, 23, 1499-1509.

Zíková N., Wang Y., Yang F., Li X., Tian M., Hopke P.K., 2016. On the source contribution to Beijing PM2.5 concentrations. *Atmospheric Environment*, 134, 84-95.

Zhou, L., Hopke, P.K., Paatero, P., Ondov, J.M., Pancras, J.P., Pekney, N.J., Davidson C.I., 2004. Advanced factor analysis for multiple time resolution aerosol composition data. *Atmospheric Environment*, 38, 4909-4920.

19. TESTS FOR MODEL PERFORMANCE VALIDATION

The fundamental, natural physical constraints that must be fulfilled in any source apportionment study are as follows (Hopke, 2010):

- The original data must be reproduced by the model; the model must explain the observations;
- The predicted source compositions must be non-negative; a source cannot have a negative elemental concentration (slightly negative values are acceptable provided zero is in the confidence interval);
- The predicted source contributions to the aerosol must all be non-negative; a source cannot emit negative mass;
- The sum of the predicted elemental mass contributions for each source must be less than or equal to the total measured mass for each element; the whole is greater than or equal to the sum of its parts.

The assignment of a source factor to a source type (or source category) is a critical step in factor analysis. Therefore, it is important to carry out sensitivity tests that assess the variability of the results because of different combinations of sources and/or species in the model (Watson et al., 2008). Several diagnostics are available to evaluate the receptor model results.

Advanced User Box

Actually, ME-2 allows a certain degree of negativity in the source/factor contributions for the sake of better rotational uniqueness (Norris et al., 2009).

Ratios

Unique source tracers are rare, therefore elemental and/or molecular ratios have often been used to trace similar sources, such as combustion processes or mineral sources, for example. In factor analysis techniques, the resolved factor profiles are often evaluated by comparing relative amounts of elements/compounds with those expected in relevant sources (Galarneau, 2008). Robinson et al. (2006a, b and c) demonstrated that the ratio of marker species in a source profile, when compared with those from the same and/or different source types and from ambient samples, helps to interpret the source variability and identify the most important sources in a region. However, one should bear in mind that the two assumptions of unique ratios among sources and conservative ratios in the atmosphere are not always met in reality. Also, the species examined should have similar reactivities with respect to atmospheric oxidants and solar radiation and similar particle size distributions in order to exclude differences in particle scavenging by precipitation or particle dry deposition (Galarneau, 2008).

One of the first uses of the elemental ratio was proposed by Juntto and Paatero (1994) who compared the Na/Cl ratio in PMF factors with sea-water composition. Liu et al. (2003) showed that their long-range transported dust profiles correlated well with standard reference Chinese desert dust, with the exception of enrichment in sulphate. Hien et al. (2004) used several ratios to distinguish between Local Burning and Long-Range transport aerosols. Hien et al. (2005) used different Ca/Si ratios to separate coal fly ash from soil dust. Lanz et al. (2007) calculated ratios of the modelled primary organic aerosols (POA) and measured primary pollutants such as elemental carbon (EC), nitrogen oxides (NO_X), and carbon monoxide (CO), finding good agreement with literature values. Organic and inorganic ratio evolutions have been also examined as a function of photochemical age of aerosols (DeCarlo et al., 2010).

Residuals

The distribution of residuals (the percentage of all scaled residuals in a given bin, 0.5 for example) should be investigated in order to verify how well the model fits each species. If a species has many large-scaled residuals or displays a non-normal curve, it may be an indication of a poor fit. A well-modelled species instead shows normally distributed residuals within the range +3 and -3.

In weighted ‘least squares’ analysis, the distribution of residuals can vary substantially with the different values of the variables (species). Therefore, weighted residuals (Graybill and Iyer, 1994) must be used in graphical residual analysis, so that the plots can be interpreted as usual. This must be taken into account when evaluating EPA PMF v3 default unweighted residual graphs.

The scale of the histogram chart (y-axis) is important. Setting the maximum values as the maximum value of each species is helpful when examining individual species and the shape of their distributions. If the Y-axis maximum is fixed at 100%, a comparison between species can easily be made. Species with residuals beyond +3 and -3 need to be further evaluated by comparing the observed vs modelled concentrations by means of scatter plots and/or time series. Large positive scaled residuals may indicate that the model is not fitting the species or that the species is present in an infrequent source. Species that do not have a strong correlation between observed and modelled values or have poorly modelled peak values should be evaluated by the user to determine if they should be downweighted or excluded from the model.

Other useful statistics when comparing observed vs modelled values are the coefficient of determination (r^2), intercept, slope, and normal residual (EPA PMF v3 User Guide; Norris et al., 2008).

The Kolmogorov-Smirnoff test can be used to determine whether the residuals are normally distributed. If the test indicates that the residuals are not normally distributed, the user should visually inspect the histogram for outlying residuals. A very narrow (leptokurtic) distribution of residuals suggests that species are fitted too well and may be an indicator of “ghost factors”, which can explain most of the variation of one species (Amato and Hopke, 2012).

Advanced User Box

Residuals can also be compared between different runs of one model (different starting points). The sum of squared difference between residuals of a pair of runs can be used (as in EPA PMF v3) as a diagnostic of different solutions (rather than mere rotations of the same solution).

In EPA PMF v5, it is possible to examine the ratio of observed to expected Q-values using Q/Qexp charts. This is an efficient way to understand the residuals of the PMF solution, and in particular, what samples and/or species were not well modelled (i.e. have values greater than 2).

Time trends

Source strengths are often time-dependent due to the influence of atmospheric processes (nucleation, volatilisation, transport, etc.), meteorological parameters (solar radiation, humidity, precipitation, etc.), and variation in human activity (intra-day, day-to-day). As a result, the source contributions will also change over time, and this variation is a suitable diagnostic for evaluating interpretations of factor profiles.

Some programs such as EPA PMF v3 already implement tools for a quick check of the seasonal and weekday/weekend variation of factor contributions. However, the user can further explore their time variability in relation to concentrations of gaseous pollutants such as SO_2, CO and NO_x for combustion sources (Zhou et al., 2005; Yue et al., 2008; Brown et al., 2012), O_x (O_3+NO_2) for secondary sources (Huang et al., 2010), and NH_3 for agricultural activities (Eatough et al., 2010). In some cases, factor analysis can couple different pollutant categories in a unique dataset; for example, Pey et al. (2009) combined the size distribution of aerosols, meteorological parameters, gaseous pollutants and chemical speciation of $PM_{2.5}$ to carry out a PCA analysis.

A posteriori wind direction analysis

A simple but reliable method is to plot source contributions in a polar scatter plot in such a way that wind direction determines the angle and source contribution determines the radius of each plotted point. Such a plot shows at a glance the general behaviour of wind-directional dependence. Also, an overview of the individual points is helpful, as one or two high-concentration points cannot distort the picture, as discussed above. Additional information, such as winter/summer classification, may be indicated by using different colours when plotting winter and summer source contribution points. See section 18 for a more detailed discussion of wind direction analysis techniques.

Bootstrapping

Bootstrap (BS) tools are commonly used in factor analysis to identify if a small number of measurements (or values on the input matrix) has the potential to greatly influence the solution. As the law of large numbers dictates, the possibility of this fact to happen decreases as the size of the matrix increases, and thus small datasets are more sensitive to BS analysis. BS uncertainty intervals include the effects of random errors and partially rotational ambiguity. In each perturbed or resampled version, some randomly chosen rows of the original matrix occur multiple times, while other rows do not occur at all. Each resampled data set is decomposed into profile and contribution matrices using PMF (Norris et al., 2008). During bootstrapping the correlation between resampled factors against each base factor is tested. The resampled factor is assigned to the base factor with the highest correlation, provided that a correlation coefficient above a user-specified threshold is found. If no base factors have a correlation above the specified threshold that factor is considered “unmapped” (Norris & Brown, 2014). Even though BS is not designed to capture rotational uncertainty, it offers some insight to a certain extent BS results should be interpreted carefully especially in small datasets. A crucial factor to reduce the rotational uncertainty in a solution is the number of zero values in G, F matrices (Paatero & Hopke, 2008). Since in each resample BS omits a number of G values, it is possible that some zero values will be omitted as well. Thus, the rotational stability of the resampled solution is decreased leading to high uncertainty estimations. If the quality of the data is high and the zero values are certain, then the estimations are not realistic. A solution might be greatly influenced by a small number of observations (whether they are high or low/zero). This is not always alarming, when no erroneous observations are present in the dataset. The higher uncertainty for BS and BS-DISP, indicates that a number of peak events affect these factors. Those events might not be resampled in the BS runs leading to higher uncertainty. The user can check the 24h source contributions to investigate if such events do exist in the solution (Manousakas et al., 2017). A matter of discussion is whether such events should be considered as outliers and be subsequently removed from the analysis. Not every observation outside the pattern of the time series is an outlier and should therefore be omitted. Events of episodic nature such as forest fires or intense long-range transport events may appear as outliers, but if they are down weighted, then a serious modeling error is made, leading to loss of critical information (Paatero et al., 2014). There are certain cases that an event might take place that leads to a seemingly not normal observation in the dataset (high or low). For example, if a certain area is affected by the emissions of an industry, it is sometimes possible the contribution of this source to be zero, when for example the industry was shut down for a time period for maintenance. On the other hand, if a source such as mineral dust or biomass burning displays a very high contribution on a certain day, the reason may well be an event like African dust transportation or a forest fire. Those events provide very important information that will help the modeling process. If a small number of such events do exist, the dataset will be very sensitive in BS analysis. This is not always alarming. If the high/low values are reliable, then the large variation in the solution, expressed as high uncertainty intervals, cannot be considered as an indication of bad fitting. Another source of errors, investigated by the uncertainty estimation tools, is the modeling uncertainty. Regarding modeling errors, it is not known how well BS captures the uncertainty from this cause (P. Paatero et al., 2014). If the user specifies data uncertainties wrongly, this modeling error usually has minimal impact in BS results (Brown et al., 2015).

The first thing that should be examined at the BS results is whether the base run (the initial run before the resampling) is within the interquartile ranges of the profiles. If the base run for a species is outside the interquartile range, then the concentration of that species on the source profile is greatly affected by a small number of observations. As mentioned earlier, this indication by itself is not enough to support that modeling for this species is not well defined. However, the interpretation of the results should be made with caution. It is suggested that the dataset should be examined for the identification of those events affecting the solution. If those events have a physical meaning or no indications of some random error exist, the user can proceed with the interpretation of the solution. This problem is most often encountered when dealing with small datasets. In large datasets it is more probable that bad fitting is the cause of poor BS results. The second factor to be checked is the number of resamples that are attributed to the right factor, then to which factor the remaining resamples are attributed and finally the number of unmapped resamples. If there is a tendency to have a spread of BS resamples to different factors it should be examined if the factors have similar profiles, or common main tracers. If they do, then the physical meaning of the factors should be examined because this could be interpreted as an indication of too many factors employed. In case though that the factors have indeed physical meaning and the experience in the area leads to the conclusion that the situation is better described using all sources then the number of factors can remain as is. In the case though that the mixing between factors is random and the number of unmapped cases is high, then it is probably due to a poorly resolved solution.

From the above it is clear that the uncertainty intervals of BS analysis do not always offer a clear answer to whether or not the solution is stable and reliable enough to be used. Additional checks of the dataset may be required to interpret or complement BS analysis, but this type of checks involve a level of subjectivity. The user should report the number of resamples analyzed and the size of percentiles of the obtained distribution of results chosen for error limits (5th and 95th percentiles in EPA PMF v5). The percentage of BS factors assigned to each base case factor and the number of BS factors not assigned to any base case factor (unmapped) as well as the interval ratios of each factor's identifying species should be also reported (Brown et al., 2015).

Overall uncertainty

The output from source apportionment (SA) consists of source contribution estimates (SCEs) with a definite uncertainty. Special efforts must be taken by the SA scientist to analyse and communicate this uncertainty. Most receptor models compute the uncertainty of the output. However, in cases where results derive from more than one SA technique, the computation of the uncertainty of the combined SCEs is not straightforward. Larsen et al. (2012) have recently demonstrated how probabilistic uncertainty characterisation by Monte Carlo simulations yielding probability distributions can be used to combine results deriving from CMB, PMF, and emission factor analysis. The advantage of this approach is that it generates the uncertainty of the combined SCEs as well as essential data for sensitivity analysis.

The newest version of EPA PMF (v5) incorporates into the model substantially improved error estimation methods. Those methods attempt to estimate the errors arising from the inherent uncertainties in the measured data as well as the rotational ambiguity in the solution (Hopke, 2016). Previous versions of PMF offered only “bootstrapping” as a tool to estimate the effect of random errors, and to some extent, the rotational ambiguity in the dataset. Fpeak was a function for estimating the lower limit of rotational uncertainty (Reff et al., 2007). The newest version includes two additional tools apart from the classical bootstrap (BS), which are namely the displacement of factor elements (DISP), and the bootstrap enhanced by displacement of factor elements (BS- DISP). These two methods, designed to complement each other (P. Paatero, et al., 2014), are presented hereafter.

Displacement (DISP): this method estimates the uncertainty of individual variables in F (source profile) by repeatedly fitting the model until each variable is displaced from its original value. Each displacement is extended until the Q increases by a predefined level, given as a maximum allowable change in Q (dQmax). dQmax is defined as the difference between the initial Q value, Qbase and the modified Q value, Qdisp: (dQmax= Qbase- Qdisp)**.** The model generates results for the following dQmax values: 4, 8, 15, and 25. Each displacement is interpreted as the upper or lower interval estimate of the displaced variable. If the intervals for a certain variable are high, then the level of the presence of that variable in the factor is uncertain. Consequently, if that variable is an important tracer of the source the factor is assigned to, the identification of the source is uncertain as well. DISP is executed for every predefined level of dQmax and the minimum and maximum source profile values are summarized for each variable in each factor profile. Only the species that are set as strong variables in the PMF v5 model are actively displaced. Since DS intervals are directly related to the uncertainty of the variable, setting variables as weak so that they are actively displaced would produce unrealistically high error estimates. DISP analysis provides an estimate for all species, even those that are not actively displaced. In this case the intervals might be smaller than those obtained in the case where these species were actively displaced (Brown et al., 2015). DISP is sensitive to misspecified data uncertainties (errors). It underestimates the uncertainty of the solution if data errors are correlated or underestimated and overestimates the uncertainty if data errors are overestimated. To evaluate the DISP results the user needs to first check the summary that it is provided by the program. If the decrease in Q is higher than 1%, then the analysis should be repeated, as it is very likely that the used solution (base solution) is a local and not a true global minimum. During DISP the factors are gradually transformed without the Q changing too much. When the rotational ambiguity of the solution is high, the identity of the resolved factors may be exchanged or swapped during DISP runs. This is expressed in the diagnostic result as a number of factor swaps. The factor swaps that occur during the analysis are reported for the 4 levels of the predefined maximum allowable change in Q (dQmax). If factor swaps occur for the smallest dQmax, it indicates that there is significant rotational ambiguity and that the solution is not sufficiently robust to be used (Norris & Brown, 2014). In extreme cases, presence of large rotations may prevent well-defined modeling altogether (Emami & Hopke, 2017). Some measures that can be taken to reduce the rotational ambiguity are to decrease the number of factors and/or check the dataset for the identification of marginal species. The upper and lower intervals for each variable in every factor should be checked, especially for the variables that are used for source identification. If the intervals for those variables are too high, then the identification of the sources is uncertain. The PMF user should report the DISP analysis results in any published work. The species that are not actively displaced (downweighted or weak species), the number of factor swaps, the decrease in Q (%) and the interval ratios of each factors' identifying species should be reported.

Bootstrap-Displacement (BS-DISP): in this method, DISP is combined to aforementioned BS method. It can thus estimate uncertainty associated both with rotational ambiguity and random errors. In BS-DISP the fitted elements in factor profiles are displaced in every resample of the BS. As described by Brown et al 2015, this process may be viewed as follows: "each BS resample results in one solution that is randomly located within the rotationally accessible space. Then, the DISP analysis determines an approximation for the rotationally accessible space around that solution. Taken together, all the approximations of rotationally accessible spaces for randomly located solutions represent both the random uncertainty and the rotational uncertainty for the modeled solution to the complete dataset".

BS-DISP analysis takes many hours to be completed so it is suggested to be used as a diagnostic tool after the results from BS and DISP are evaluated and found to be satisfactory. BS-DISP displays some of the weaknesses of the classical BS method. If some of the rotation blocking values (zeros) or the factor defining values (high values) are omitted, the intervals will be large, a fact that cannot be regarded as definite sign of a bad fitting. In order to reduce computational time in BS-DISP the species that will be actively displaced are defined by the user. It is suggested that only species that are important for source identification are selected to be actively displaced. Passive estimation can still be used in the evaluation, unlike DISP where the passive estimations may produce very short intervals. BS-DISP output can be evaluated and reported in a similar way to that of DISP analysis. The decrease in Q, the number of factor swaps and the intervals especially of key species for source identification should be examined. After the analysis, the user should report the number of resamples and the size of percentiles of the error limits. As it is the case for BS the percentage of BS factors assigned to each base case factor and the number of BS factors not assigned to any base case factor should be reported as well.

Finally, rotational ambiguities could also be reduced by constraining individual factor elements, either scores and/or loadings, towards zero values, prescribing values for ratios of certain key factor elements (Paatero et al., 2002). It must be emphasized that application of these techniques must be based on some external information about acceptable or evidence-based factor profiles.

The base run can be modified (constrained) by the following methods:

- by setting some factor elements to a fixed value
- by specifying lower and/or upper limits
- by pulling a factor element towards a certain value and
- by setting an equation such as a ratio, a mass balance equation or a custom equation.

Some constraints are considered strong such as setting a variable on a fixed value because they can perturb the model results significantly. Pulling towards a value has the advantage that if the equation is incompatible with the result (Q changes too much), then the pulling will fall sort of the target value (Paatero & Hopke, 2008). In other words, if a free rotation is not available then the pulled factor will never reach the user defined outcome. Setting constraints in the form of equations such as a ratio is regarded as a rather strong constraint but a lesser one than setting a particular value. For all other constraints apart from setting a certain value the model offers the option to set the maximum allowed dQ% change. Giving low maximum dQ% change ensures that significant perturbation of the model results is less probable. It must be emphasized here that the application of constraints does not exactly reduce the rotational ambiguity, but the effect can be better described as pushing the solution to a user defined/selected optimum solution among the available rotations by using some external information and imposing this on the model.

Advanced User Box

RMs output uncertainty derives from both inaccuracy in the input data and model assumptions and ambiguities (Karagulian and Belis, 2012). Monte-Carlo probabilistic methods (such as bootstrapping) are suitable for estimating the random component of the output uncertainty in factor analytical methods. On the other hand, according to a study on the error estimation methods implemented in EPA PMF v5 (Paatero et al., 2013; Hopke, 2016), analysis of controlled perturbations of the F and G matrices elements (displacements) is most appropriate for estimating the rotational uncertainty (which is a non-random component).

If the number of factors is large, as is typical when analysing speciated data, rotational uncertainty is often the leading cause of uncertainty in results and relying on Monte-Carlo methods may produce error intervals that are much too narrow. On the other hand, if there is only a small number of rotation-limiting zero values in true time series (G) factors, customary bootstrapping may also lead to uncertainties that are much too large. Whenever the resampling process happens to eliminate such zero values, rotational uncertainty may increase dramatically, and bootstrapped results may deviate dramatically from the original full-data results.

Nevertheless, more experience is still needed on the application of these new methodologies to real-world datasets.

References

Amato, F. and Hopke, P.K., 2012. Source Apportionment of the Ambient PM2.5 in East St. Louis using Constrained Positive Matrix Factorization. *Atmospheric Environment* 46, 329-337

Brown, S.G., Lee, T., Norris, G.A., Roberts, P.T., Collett, J.L., Paatero, P., Worsnop, D.R., 2012. Receptor modeling of near-roadway aerosol mass spectrometer data in Las Vegas, Nevada, with EPA PMF. *Atmospheric Chemistry and Physics* 12, 309-325

Brown, S. G., Eberly, S., Paatero, P., & Norris, G.A., 2015. Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results. Science of The Total Environment, 518–519, 1–10

Decarlo, P.F., Ulbrich, I.M., Crounse, J., De Foy, B., Dunlea, E.J., Aiken, A.C., Knapp, D., (...), Jimenez, J.L., 2010. Investigation of the sources and processing of organic aerosol over the Central Mexican Plateau from aircraft measurements during MILAGRO. *Atmospheric Chemistry and Physics* 10 (12), 5257-5280.

Eatough, D.J., Kuprov, R., Hansen, J.C., Olsen, N., 2010. PMF apportionment of primary and secondary fine particulate material in the Salt Lake Valley: Winter 2009. Proceedings of the Air and Waste Management Association's Annual Conference and Exhibition, AWMA 8, 6258-6263.

Emami, F., & Hopke, P. K., 2017. Effect of Adding Variables on Rotational Ambiguity in Positive Matrix Factorization Solutions. Chemometrics and Intelligent Laboratory Systems, 162(January), 198–202. https://doi.org/10.1016/j.chemolab.2017.01.012

Galarneau, E., 2008. Source specificity and atmospheric processing of airborne PAHs: Implications for source apportionment. *Atmospheric Environment* 42, 8139–8149

Graybill, F.A., and H.K. Iyer, 1994. Regression Analysis: Concepts and Applications. Duxbury Press, Belmont, California.

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. *Atmospheric Environment* 38 (2), 189-201.

Hien, P.D., Bac, V.T., Thinh, N.T.H., 2005. Investigation of sulfate and nitrate formation on mineral dust particles by receptor modelling. *Atmospheric Environment* 39 (38), 7231-7239.

Hopke, P. K., 2010. The application of receptor modeling to air quality data. *Pollution Atmospherique special issue* 91-109

Hopke, P. K. (2016). Review of receptor modeling methods for source apportionment. Journal of the Air and Waste Management Association, 66(3), 237–259. https://doi.org/10.1080/10962247.2016.1140693

Huang, X., Zhao, Q., He, L., Hu, M., Bian, Q., Xue, L., Zhang, Y., 2010. Identification of secondary organic aerosols based on aerosol mass spectrometry. *Science China Chemistry* 53 (12), 593-2599

Juntto, S. and Paatero, P., 1994. Analysis of daily precipitation data by positive matrix factorization. *Environmetrics* 5 (2) ,127-144, Karagulian F. Belis C.A., 2012. Enhancing Source Apportionment with receptor models to Foster the Air Quality Directive Implementation. International Journal of Environmental Pollution 50, 190-199.

Lanz, V.A., Alfarra, M.R., Baltensperger, U., Buchmann, B., Hueglin, C., Prevot, A.S.H., 2007. Source apportionment of submicron organic aerosols at an urban site by factor analytical modelling of aerosol mass spectra. *Atmospheric Chemistry and Physics* 7 (6), 1503-1522.

Larsen, B.R., Gilardoni S., Stenström K., Niedzialek J., Jimenez J., Belis C.A., 2012. Sources for PM air pollution in the Po Plain, Italy: II. Probabilistic uncertainty characterization and sensitivity analysis of secondary and primary sources. *Atmospheric Environment* 50, 203-213.

Liu, W., Hopke, P.K., Van Curen, R.A., 2003. Origins of Fine Aerosol Mass in the Western United States Using Positive Matrix Factorization. *Journal of Geophysical Research- Atmospheres* 108 (D23), D04716.

Manousakas, M., Papaefthymiou, H., Diapouli, E., Migliori, A., Karydas, A. G., Bogdanovic-Radovic, I., & Eleftheriadis, K., 2017. Assessment of PM2.5 sources and their corresponding level of uncertainty in a coastal urban area using EPA PMF 5.0 enhanced diagnostics. Science of the Total Environment, 574, 155–164. https://doi.org/10.1016/j.scitotenv.2016.09.047

Norris G., Vedantham R., K. Wade, S. Brown, J. Prouty, C. Foley, 2008. EPA Positive Matrix Factorization (PMF) 3.0 Fundamentals & User Guide. EPA 600/R-08/108.

Norris G., Vedantham,R., Wade D., Zahn P., Brown S., Paatero P., Eberly S. Foley C., 2009. Guidance Document for PMF Applications with Multilinear Engine.EPA 600/r-09-032.

Norris, G., & Brown, S., 2014. EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide. EPA PMF 5.0 Manual.

Paatero, P., Hopke, P. K., Song, X.-H., & Ramadan, Z., 2002. Understanding and controlling rotations in factor analytic models. Chemometrics and Intelligent Laboratory Systems, 60(1–2), 253–264. https://doi.org/10.1016/S0169-7439(01)00200-3

Paatero P., & Hopke, P. K., 2008. Rotational tools for factor analytic models. Journal of Chemometrics, 23(2), 91–100. https://doi.org/10.1002/cem.1197

Paatero, P., Eberly, S., Brown, S. G., & Norris, G.A., 2014. Methods for estimating uncertainty in factor analytic solutions. Atmospheric Measurement Techniques, 7(3), 781–797. https://doi.org/10.5194/amt-7-781-2014

Pey, J., Querol, X., Alastuey, A., Rodriguez, S., Putaud, J.P., Van Dingenen, R., 2009. Source apportionment of urban fine and ultra-fine particle number concentration in a Western Mediterranean city. *Atmospheric Environment* 43 (29), 4407-4415.

Reff, A., Eberly, S. I., & Bhave, P. V., 2007. Receptor Modeling of Ambient Particulate Matter Data Using Positive Matrix Factorization: Review of Existing Methods. Journal of the Air & Waste Management Association, 57(2), 146–154. https://doi.org/10.1080/10473289.2007.10465319

Robinson, A.L., Subramanian, R., Donahue, N.M., Bernardo-Bricker, A., Rogge, W.F., 2006. Source Apportionment of Molecular Markers and Organic Aerosols. 1. Polycyclic Aromatic Hydrocarbons and Methodology for Data Visualization. *Environmental Science and Technolology* 40, 7803-7810.

Robinson, A.L., Subramanian, R., Donahue, N.M., Bernardo-Bricker, A., Rogge, W.F., 2006. Source Apportionment of Molecular Markers and Organic Aerosol, 2. Biomass Smoke. *Environmental Science and Technology* 40, 7811-7819.

Robinson, A.L., Subramanian, R., Donahue, N.M., Bernardo-Bricker, A., Rogge, W.F., 2006. Source Apportionment of Molecular Markers and Organic Aerosol. 3. Food Cooking Emissions. *Environmental Science and Technology* 40, 7820-7827.

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*ournal of Air & Waste Management* 58, 265–288.

Yue, W., Stolzel, M., Cyrys, J., Pitz, M., Heinrich, J., Kreyling, W.G., Wichmann, H.-E., (...), Hopke, P.K., 2008. Source apportionment of ambient fine particle size distribution using positive matrix factorization in Erfurt, Germany. *Science of the Total Environment* 398 (1-3), 133-144.

Zhou, L., Kim, E., Hopke, P.K., Stanier, C., Pandis, S.N., 2005. Mining airborne particulate size distribution data by positive matrix factorization. *Journal of Geophysical Research D: Atmospheres* 110 (7), art. no. D07S19, 1-15.

20. REPORTING RESULTS AND METHODOLOGY

Due to the large number of variables to be considered, source apportionment (SA) studies are complex. They often require adaptation of existing methods to the specific problem or the development of tailor-made solutions. In addition, there are many steps in which decisions have to be taken by the modeller. Therefore, it is essential to support the final results with an appropriate description of the methodological choices and documentation of the objective qualitative or quantitative information that support expert decisions. In this way, reviewers and final users are provided with the elements to assess the relevance of the study and other modellers get a chance to reproduce the methodology. If the results are reported in a peer-reviewed scientific journal, much detail can be provided as supplementary material that most journals now support.

The present protocol has been conceived as a reference document that cannot substitute for experience and competence. For that reason, documented participation of experts in training and intercomparisons should be promoted in order to develop and demonstrate individual and institutional capacities.

Although this protocol aims at promoting the highest quality standards, it has to deal with the intrinsic limitation of any SA study: the “true” contribution of sources to atmospheric pollution at a given point cannot be directly measured.

SA studies can be considered as being consistent with the present protocol if they comply with the following requisites:

- The results are described according to the steps proposed in sections 1-12. When limited space is available, these technical notes should be allocated in an annex or, in the case of scientific journals, as supporting information or supplementary material.
- Expert decisions are described and evidence of the objective information (e.g. quantitative tests, sensitivity analysis, external information) that supports them is provided. This point is essential for critical steps such as the selection of source profiles in chemical mass balance modelling, and the identification of the number of sources and factor assignment in factor analysis.
- The documentation includes the references of the source profiles used as input or to validate factor assignment.
- The model and version used are clearly reported and the modifications adopted for the specific case well described.
- The quantitative uncertainty of the output is estimated and reported using the techniques described in the present document or other robust methodologies available in the literature. Sources of uncertainty that cannot be quantified should be acknowledged, bearing in mind

that both inaccuracy in the input data and model assumptions and ambiguities contribute to the total uncertainty budget in receptor models.

- Estimation of overall uncertainty and validation is achieved by comparing outputs from independent models on the same dataset and/or using Monte Carlo permutation analysis techniques.
- Sensitivity analysis is performed to demonstrate that there are no substantial deviations from the mass conservation assumption.
- Only solutions that implement the quality assurance steps described in this guide can claim state-of-the-art performance in documentation prepared for community-wide intercomparison exercises.

KJ-NA-29816-EN-N

The European Commission's science and knowledge service
Joint Research Centre

JRC Mission
As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

EU Science Hub
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office
of the European Union

doi:10.2760/439106

ISBN 978-92-76-09001-4