Source Apportionment tools: advantages, limitations and complementarity

C.A. Belis

European Commission, Institute for Environment and Sustainability – JRC

and the source apportionment community

RTC ON METHODS AND TOOL TO IDENTIFY SOURCES OF AIR POLLUTION AND APPORTIONMENT OF APM

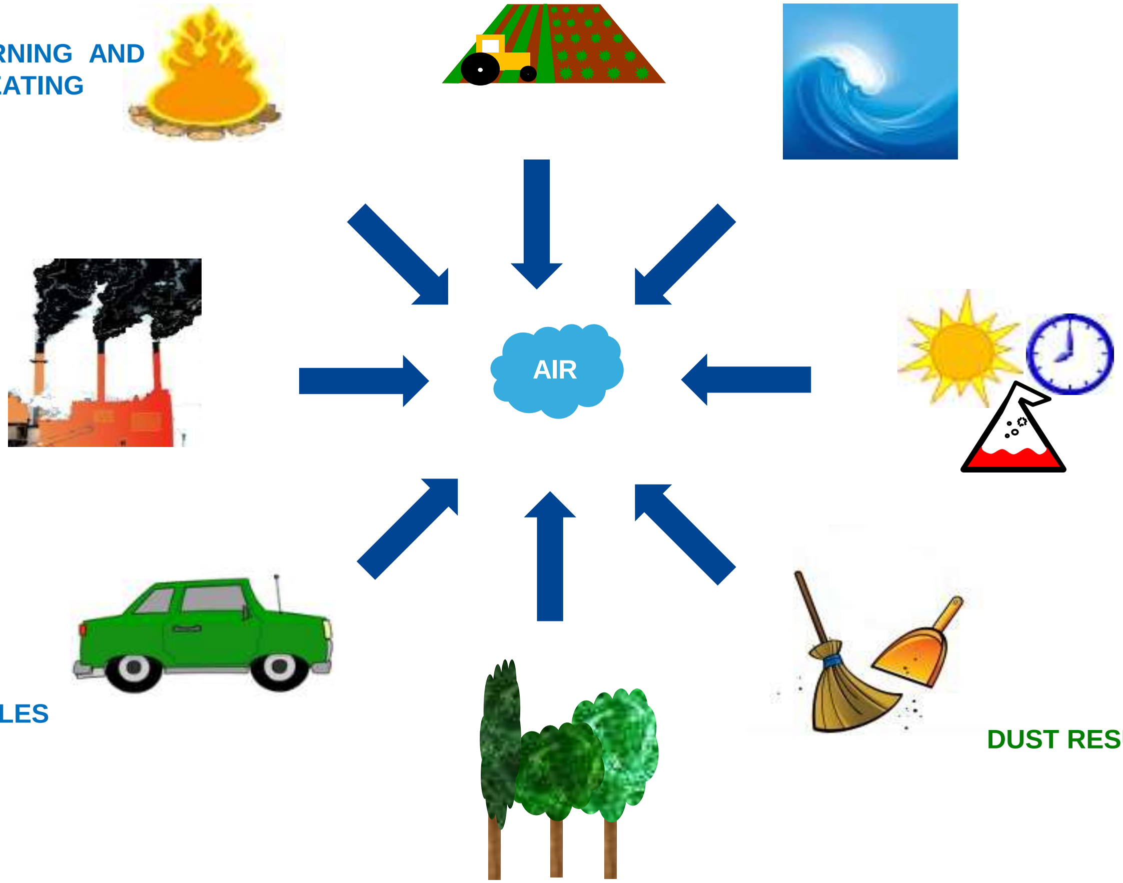
INTRODUCTION

What is source apportionment?

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

Pollution Sources

ANTHROPOGENIC



AGRICULTURE

SEA SALT

BIOMASS BURNING AND
DOMESTIC HEATING

INDUSTRY AND
POWER GENERATION

ATMOSPHERIC CHEMISTRY

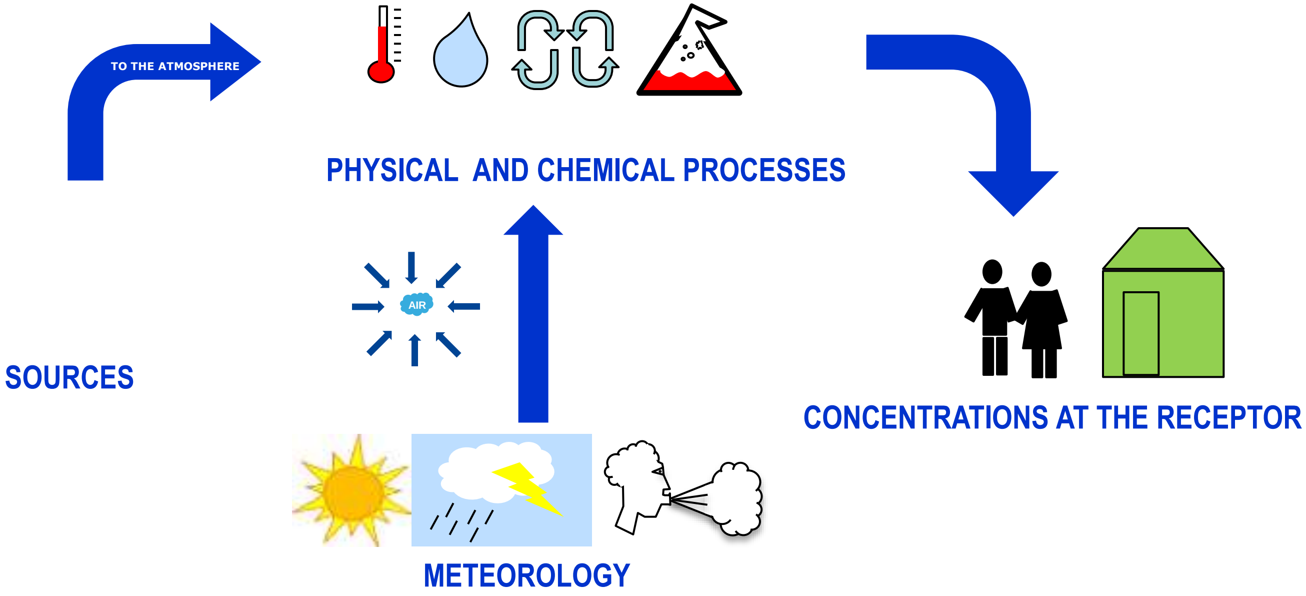
VEHICLES

DUST RESUSPENSION

BIOGENIC EMISSIONS

NATURAL

Atmospheric processes



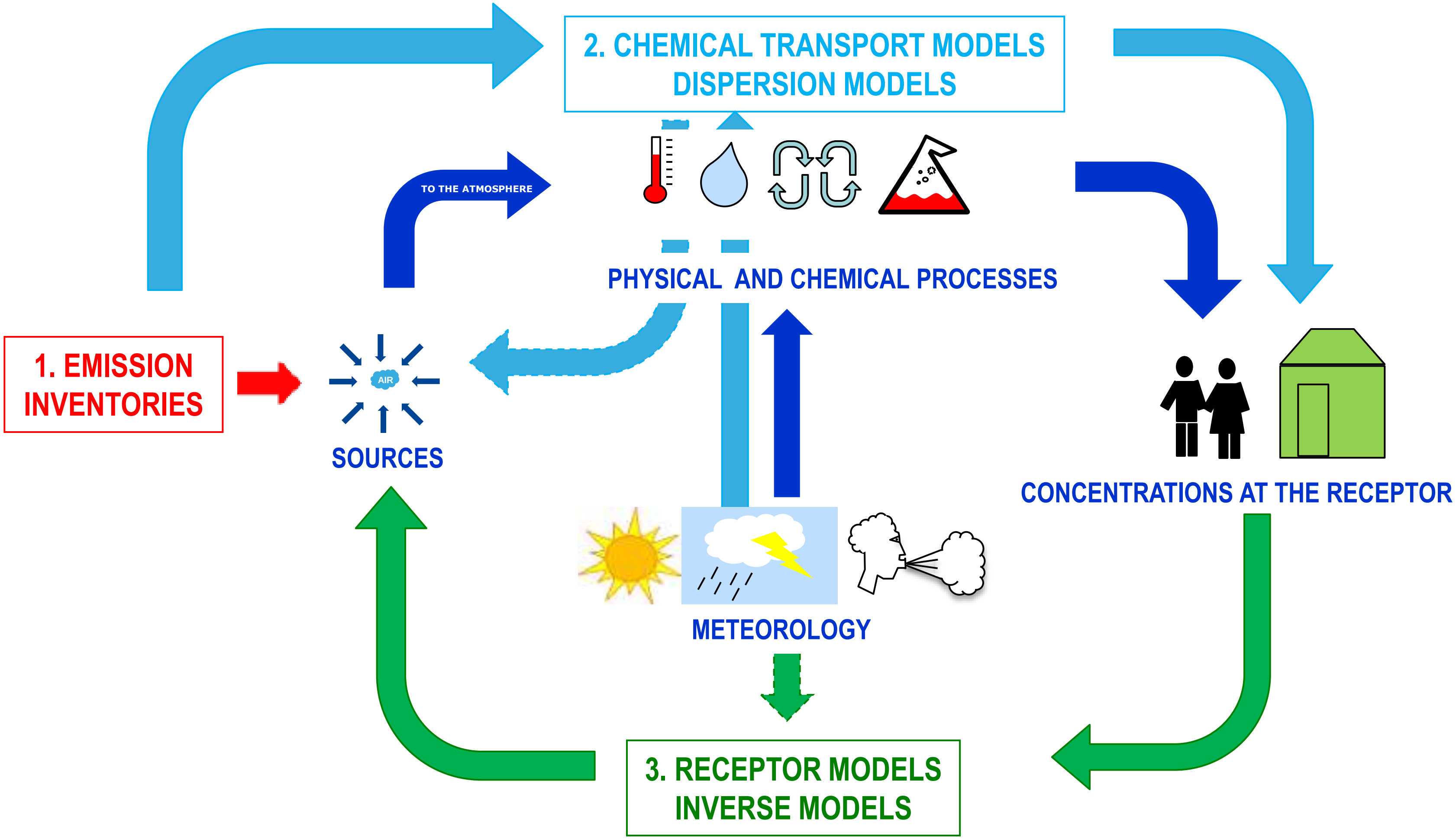
PHYSICAL AND CHEMICAL PROCESSES

SOURCES

CONCENTRATIONS AT THE RECEPTOR

METEOROLOGY

Source estimation methods



1. EMISSION INVENTORIES

2. CHEMICAL TRANSPORT MODELS DISPERSION MODELS

3. RECEPTOR MODELS INVERSE MODELS

TO THE ATMOSPHERE
AIR
SOURCES
PHYSICAL AND CHEMICAL PROCESSES
METEOROLOGY
CONCENTRATIONS AT THE RECEPTOR

Source estimation methods

1. EMISSION INVENTORIES

Required for reporting obligations
Do not consider atmospheric processes
Official data could be sketchty/inconsistent

2. CHEMICAL TRANSPORT MODELS DISPERSION MODELS

Consider atmospheric processess
Provide high resolution spatial and temporal estimations
Intensive computing resources and good parametrization neede
Simulation for short time windows
Output depends on input data quality

3. RECEPTOR MODELS INVERSE MODELS

Derive directly from data collected at the point of interest
Have good uncertainty estimation
Require field work and chemical analyses
Not applicable to all pollutants

Receptor Models

Use chemical composition of the pollutant measured at the receptor

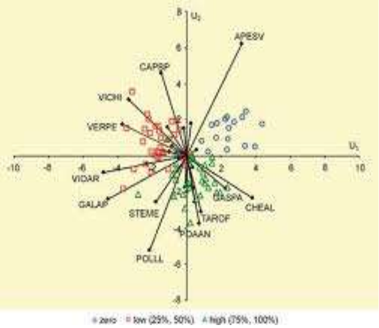
Equations refers to the chemical mass balance principle

Adjustments are needed for extremely non conservative species

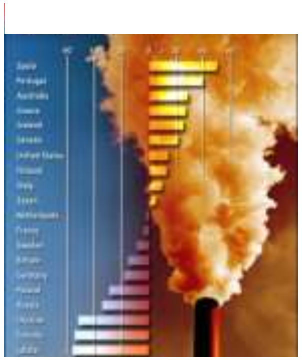
$A_e + B_e \rightarrow A_r + B_r$

Are based on statistical analysis (multivariate analysis)

At the first step do not consider physical and chemical processes but evolved hybrid models can process additional information to constrain results



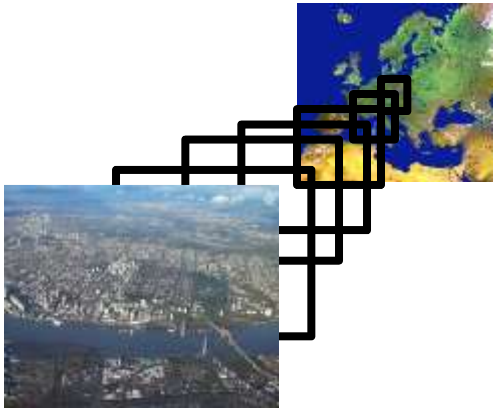
Receptor Models



Do not depend on the quality of Emission Inventories

Do not require complex meteorological and chemical processors
Low computational intensity

Mainly used in the apportionment of PM, gaseous hydrocarbons and PAHs



Suitable for urban and regional scales

Receptor Models

Assumptions

1) source profiles do not change significantly over time or do so in a reproducible manner so that the system is quasi-stationary.

2) that receptor species do not react chemically or undergo phase partitioning (solid/gas or solid/ liquid) during transport from source to receptor (i.e., they add linearly).

3) that data are representative of the studied geographical area and consistent with the conceptual model and

4) that comparable/equivalent analytical methods are used for the receptor site(s) throughout the study as well as for the characterization of the source profiles

Source and receptor models derive from the same physical construct

$$C_{ikl} = \Sigma_j \Sigma_m \Sigma_n F_{ij} T_{ijklmn} D_{kln} Q_{jkmn}$$

i = pollutant

j = source type

k = time period

l = receptor location

m = source sub-type, a specific source or groups of emitters with similar source compositions and/or locations

n = location of emitter m of source type j

C_{ikl} = ambient concentration

F_{ij} = fractional quantity of pollutant i in source j

T_{ijkmn} = transformation of pollutant i during transport

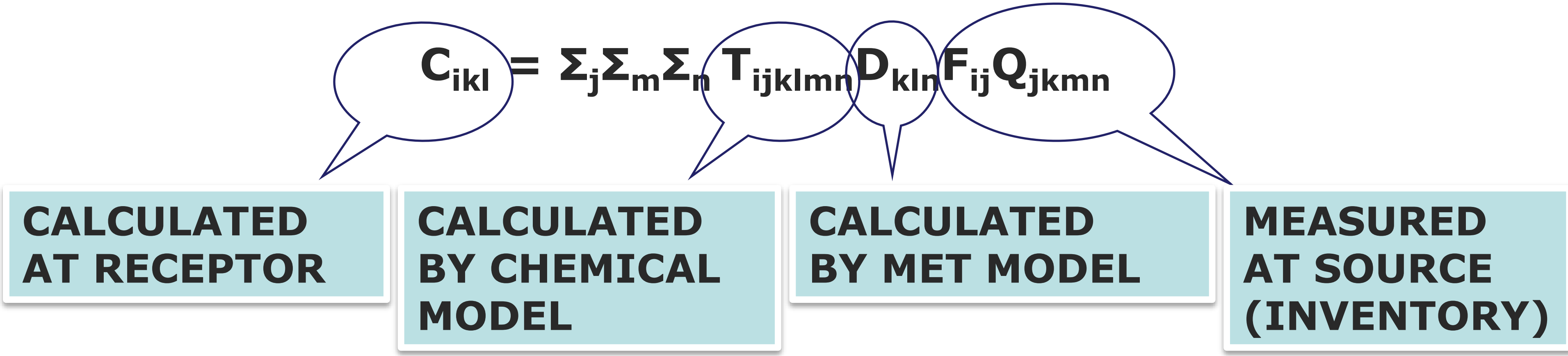
D_{kln} = dispersion and mixing between source and receptor

Q_{jkmn} = emissions rate

Watson, 1984

Source and receptor models are complementary

Source oriented Model



$$C_{ikl} = \Sigma_j\Sigma_m\Sigma_n T_{ijklmn} D_{kln} F_{ij} Q_{jkmn}$$

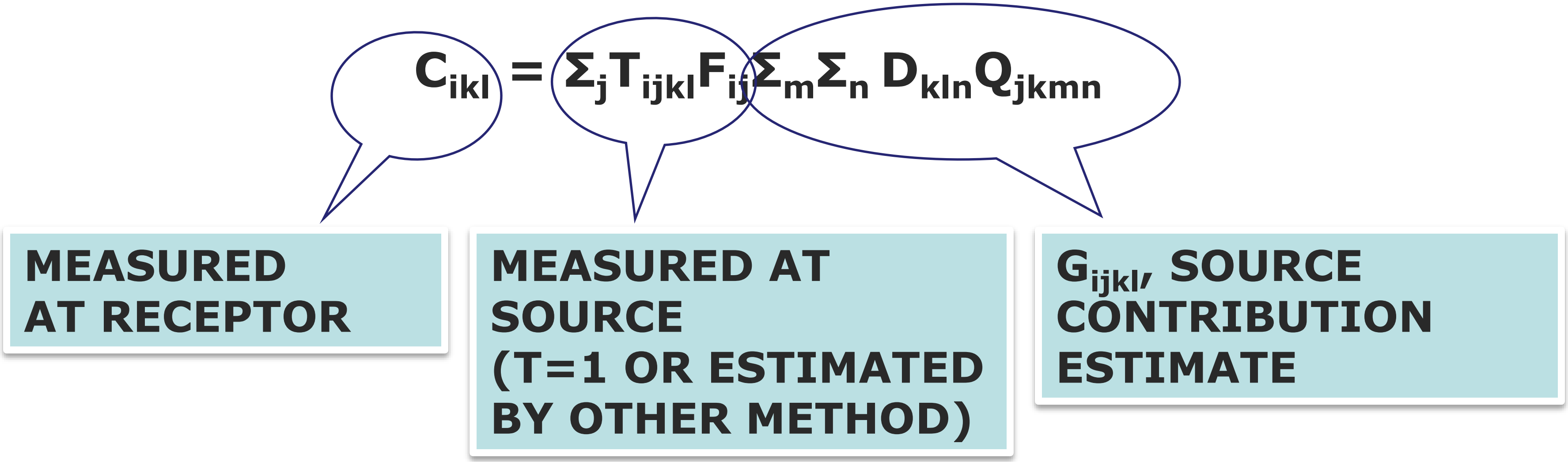
CALCULATED AT RECEPTOR

CALCULATED BY CHEMICAL MODEL

CALCULATED BY MET MODEL

MEASURED AT SOURCE (INVENTORY)

Receptor oriented Model



$$C_{ikl} = \Sigma_j T_{ijkl} F_{ij} \Sigma_m\Sigma_n D_{kln} Q_{jkmn}$$

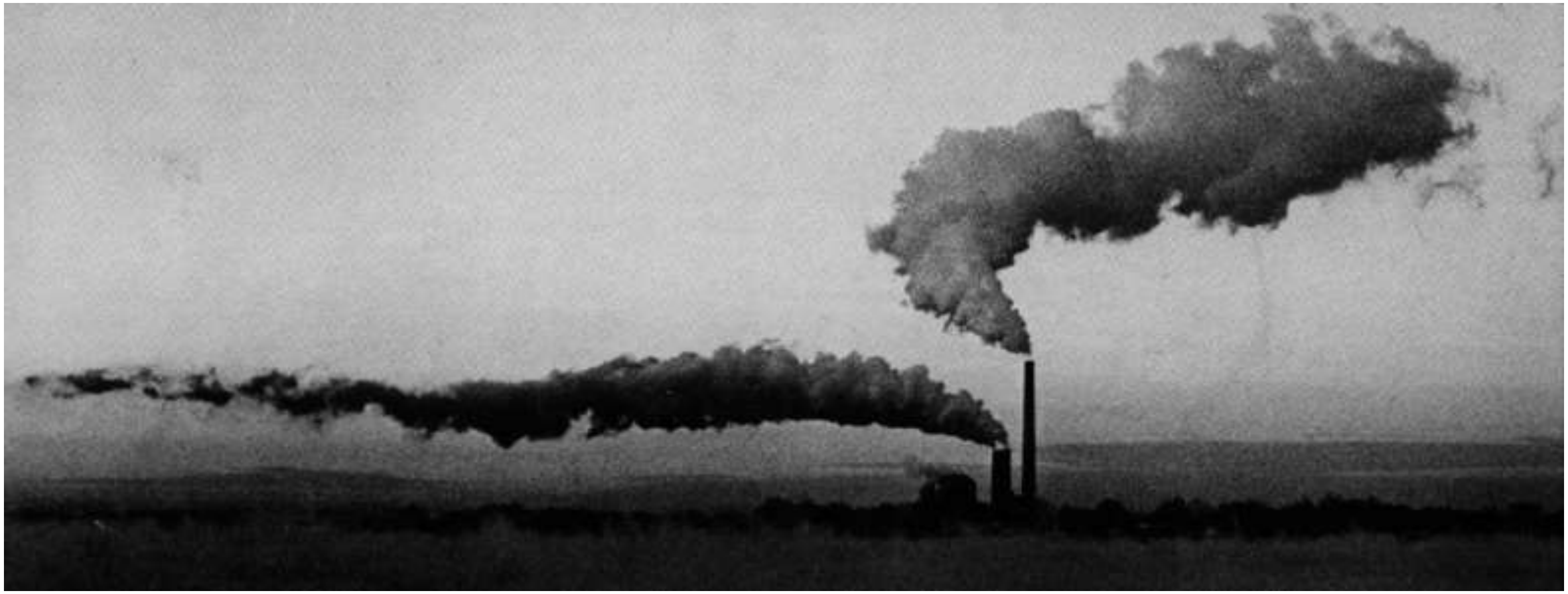
MEASURED AT RECEPTOR

MEASURED AT SOURCE (T=1 OR ESTIMATED BY OTHER METHOD)

G_{ijkl}, SOURCE CONTRIBUTION ESTIMATE

Watson, 1984

RECEPTOR MODELING



original slide from P. K Hopke

Receptor Models

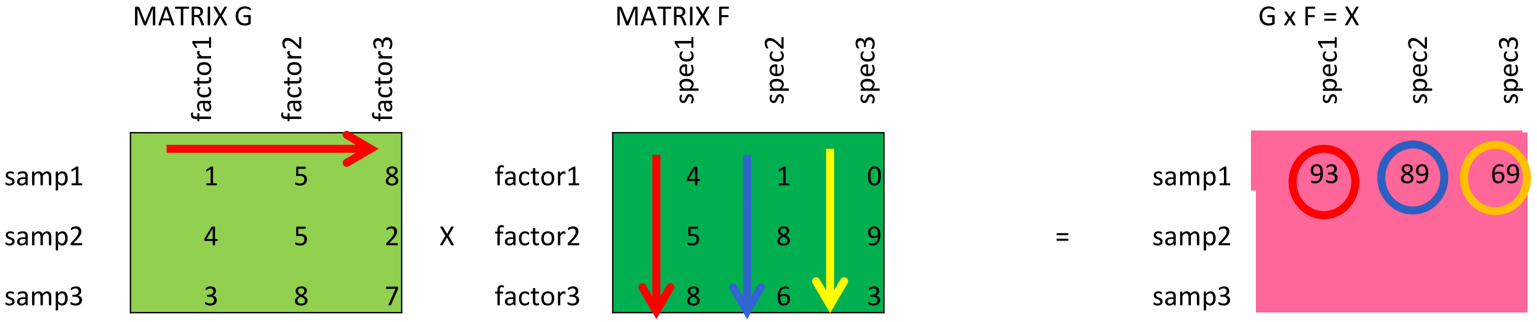
Concentration of the jth species in the ith sample

Concentration of the jth species in the kth source

$$x_{ij} = \sum_{p=1}^{P} g_{ik} f_{kj} + e_{ij}$$

Basic mass balance equation

Contribution of kth source to ith sample



$$\text{MATRIX G} \quad \begin{array}{c|ccc} & \text{factor1} & \text{factor2} & \text{factor3} \\ \hline \text{samp1} & 1 & 5 & 8 \\ \text{samp2} & 4 & 5 & 2 \\ \text{samp3} & 3 & 8 & 7 \end{array} \quad X \quad \text{MATRIX F} \quad \begin{array}{c|ccc} & \text{spec1} & \text{spec2} & \text{spec3} \\ \hline \text{factor1} & 4 & 1 & 0 \\ \text{factor2} & 5 & 8 & 9 \\ \text{factor3} & 8 & 6 & 3 \end{array} \quad = \quad G \times F = X \quad \begin{array}{c|ccc} & \text{spec1} & \text{spec2} & \text{spec3} \\ \hline \text{samp1} & 93 & 89 & 69 \\ \text{samp2} & & & \\ \text{samp3} & & & \end{array}$$

Receptor Models

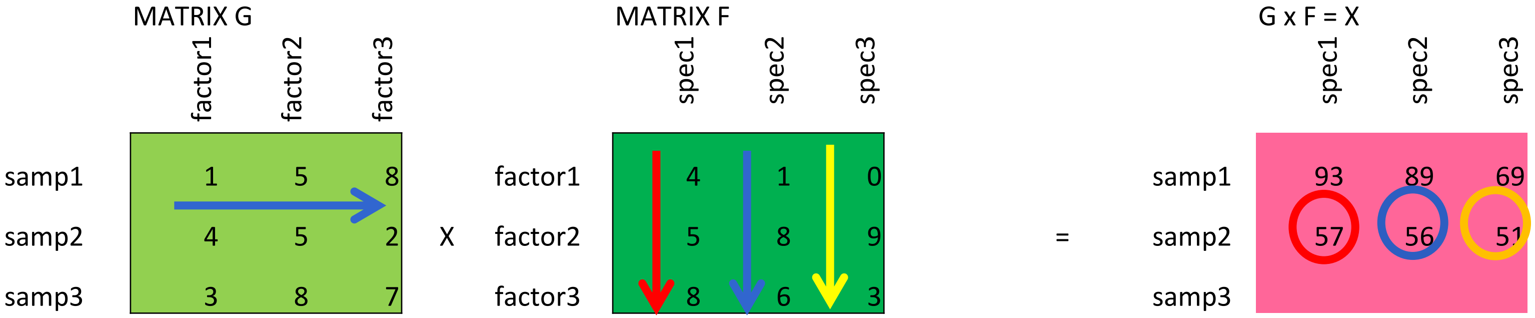
Concentration of the jth species in the ith sample

$$x_{ij} = \sum_{p=1}^{P} g_{ik} f_{kj} + e_{ij}$$

Basic mass balance equation

Concentration of the jth species in the kth source

Contribution of kth source to ith sample



MATRIX G

	factor1	factor2	factor3
samp1	1	5	8
samp2	4	5	2
samp3	3	8	7

X

MATRIX F

	spec1	spec2	spec3
factor1	4	1	0
factor2	5	8	9
factor3	8	6	3

=

G x F = X

	spec1	spec2	spec3
samp1	93	89	69
samp2	57	56	51
samp3			

Receptor Models

Concentration of the jth species in the ith sample

$$x_{ij} = \sum_{p=1}^{P} g_{ik} f_{kj} + e_{ij}$$

Concentration of the jth species in the kth source

Basic mass balance equation

Contribution of kth source to ith sample

MATRIX G

	factor1	factor2	factor3
samp1	1	5	8
samp2	4	5	2
samp3	3	8	7

X

MATRIX F

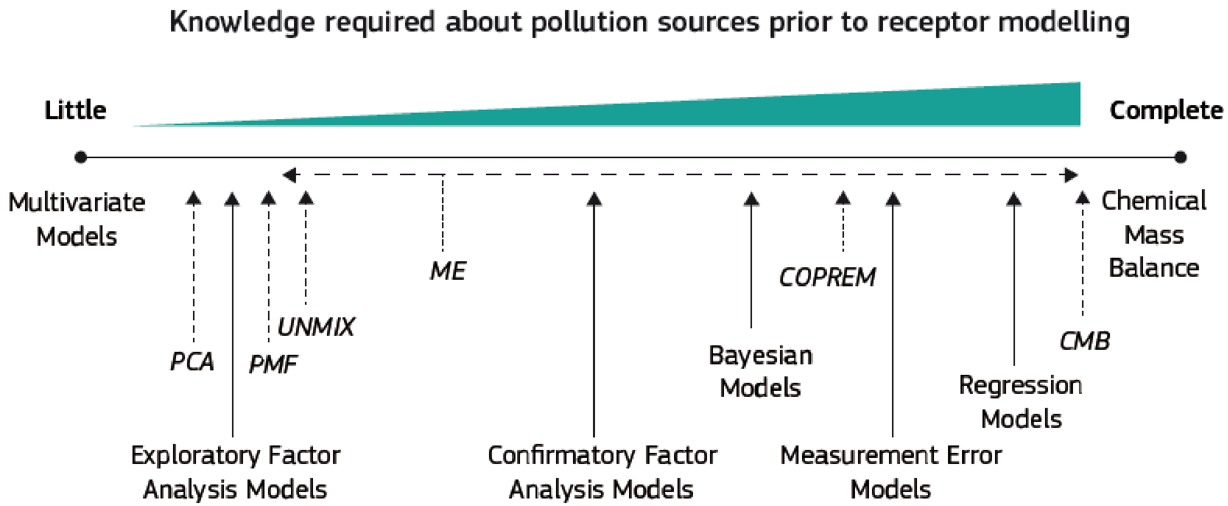
	spec1	spec2	spec3
factor1	4	1	0
factor2	5	8	9
factor3	8	6	3

=

G x F = X

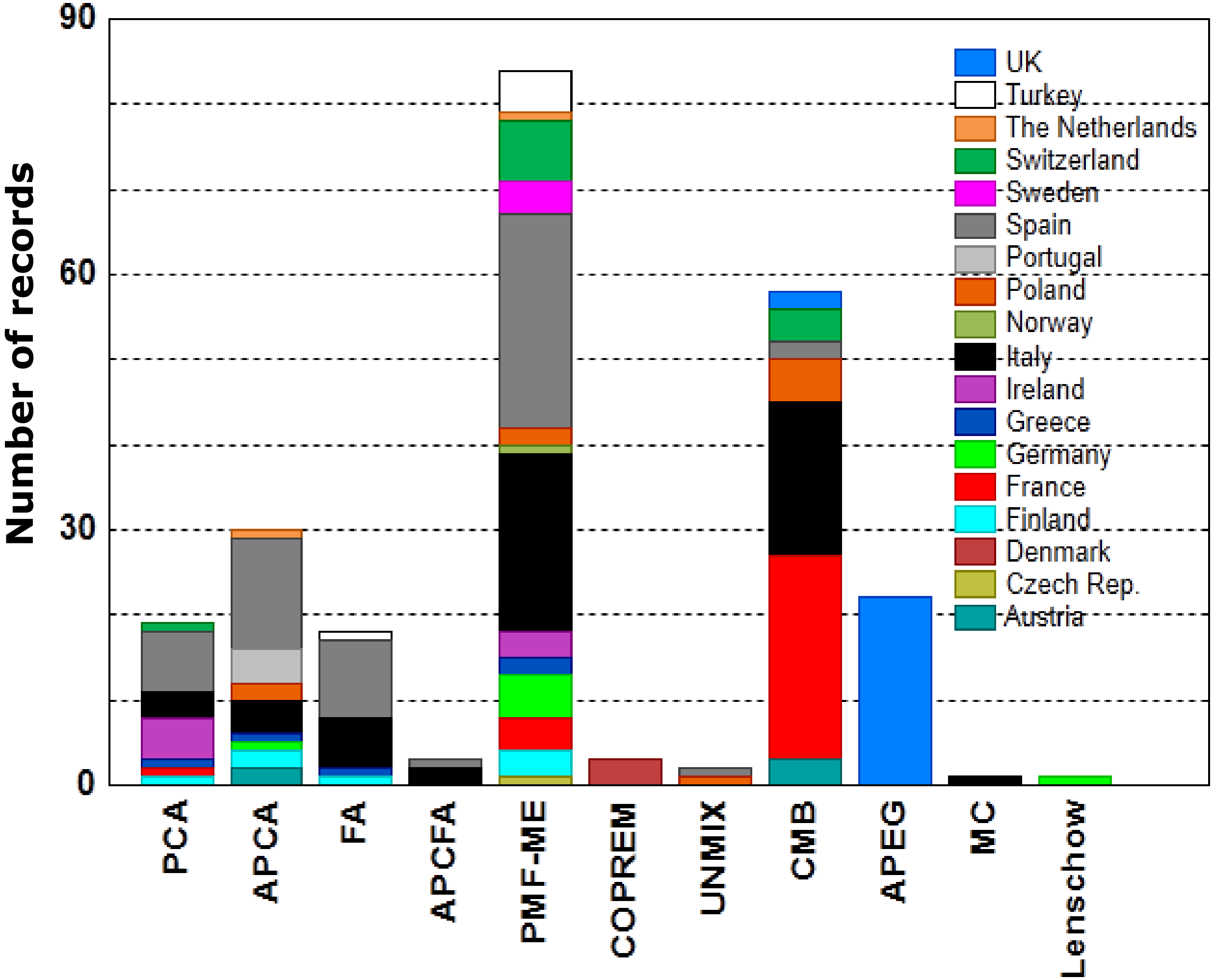
	spec1	spec2	spec3
samp1	93	89	69
samp2	57	56	51
samp3	108	109	93

Recpetor Models


Knowledge required about pollution sources prior to receptor modelling
Little
Complete
Multivariate Models
Chemical Mass Balance
PCA
PMF
UNMIX
ME
COPREM
CMB
Exploratory Factor Analysis Models
Confirmatory Factor Analysis Models
Bayesian Models
Measurement Error Models
Regression Models

Type of receptor model	Examples
Exploratory methods	Enrichment factor, tracer method, Lenschow approach, APEG
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, v4, v5
Hybrid trajectory based models	CPF, PSCF
Hybrid expanded models	PMF solved with ME-2, COPREM

European RM studies published between 2001 and 2011



80 studies – 224 (243) records

35% (PMF, ME), 24% (CMB), 21% (PCA, APCA), 8% (FA, APCFA), 12% APEG model, COPREM, Lenschow, UNMIX and Mass Closure (MC).

(AMS) data, mostly oriented to the apportionment of the PM_1 organic particulate (9 studies).

Karagulian&Belis, 2012 IJEP

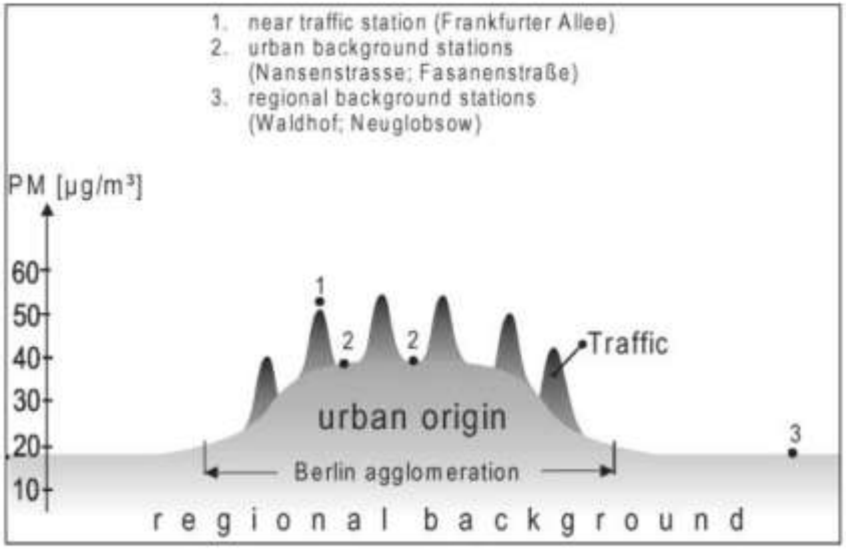
RECEPTOR MODELS USED IN EUROPE

Incremental or Lenschow Approach

Guidance to Decision 2011/850/EU

- **Regional background** is the split of total regional background in µg/m³ .
- **Urban background increment** represents the concentrations arising from emissions within towns or agglomerations, which are not direct local emissions (in µg/m³).

- **Local increment** identifies contributions from sources in the immediate vicinity of the exceedance situation.


1. near traffic station (Frankfurter Allee)
2. urban background stations
(Nansenstrasse; Fasanenstraße)
3. regional background stations
(Waldhof; Neuglobsow)
PM [µg/m³]
60
50
40
30
20
10
1
2
2
Traffic
urban origin
Berlin agglomeration
3
regional background

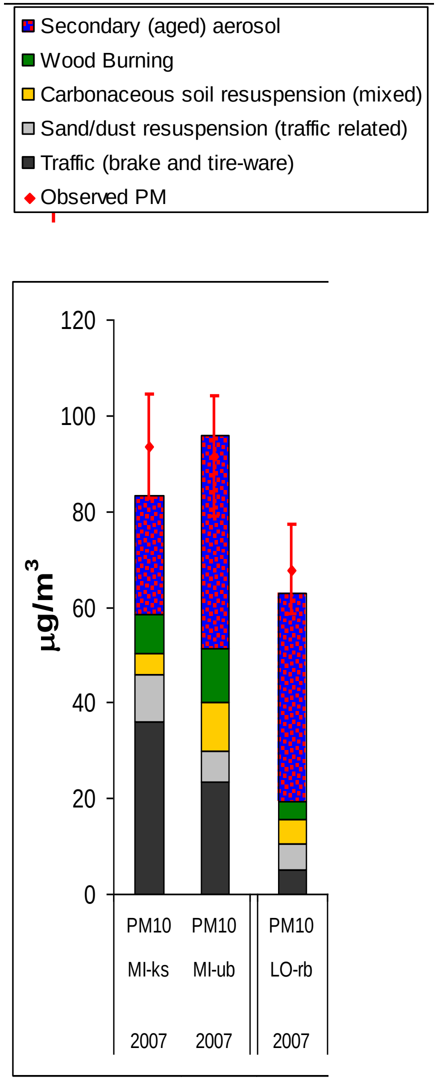
Lenschow et al., 2001 AE

Incremental Approach

- In studies carried out in a single urban background site no increment can be calculated
- Satisfactory results for estimation of traffic contributions
- The contribution of sources to primary and secondary pollution is assumed to be proportional to their emission estimations derived from emission inventories
- There are situations where this approach would lead to negative increments

Po Valley 2007 (Larsen et al, 2012):

<u>SIA – aged</u> higher in rural bkg. than urban bkg.
<u>Soil resuspension</u> higher in urban bkg than kerbside



Receptor Models

Enrichment factor

(Dams and DeJonge, 1976; Lawson and Winchester, 1979)

The EF is the ratio between elemental ratios in the measured sample to that of a reference material (e.g. particle composition vs crustal abundance)

$$EF = \frac{x_{aerosol}/r_{aerosol}}{x_{crust}/r_{crust}} \quad \text{or} \quad EF = \frac{x_{aerosol}/x_{crust}}{r_{aerosol}/r_{crust}}$$

A simple application of EF analysis for PM source indication may be the study of heavy metals (e.g. brake-metals) at a road site. For those metals not emitted by traffic, the ratio between EF of ambient PM and the EF for mineral dust (crust) remains close to unity, while this ratio will be significantly higher than one for species like Cu

Receptor Models

Mass Balance methods

The simplest mass balance is the:

Tracer approach one species for each source (Miller et al., 1972; Winchester & Nifong, 1971)

-APEG model

$$x = a.\ [NO_X] + b.[SO_4^{-2}] + c$$

where *x* is the measured PM_{10} , $[NO_X]$, $[SO_4^{-2}]$ are measured nitrogen oxides and sulphate, and a, b c are the fitted MLR coefficients (µg m^{-3})

the fulfilment of the intrinsic assumptions cannot easily be verified

-macrotracer or organic-tracer approach for SA of OC

Receptor Models

Mass Balance methods
(known source number and composition)
solution of mass balance equation by multiple regression

CMB approach: Effective variance least square (Watson, 1979; Dunker, 1979)

the weighting is inversely proportional to the square of the uncertainty in the source profiles and ambient data for each species

$$(W_e)_{jj} = \frac{1}{\sigma_j{}^2 + \sum_{k=1}^{p} \sigma_{jk}^2 g_k^2}$$

CMB characteristics

- CMB relies strongly on the availability of source profiles, which ideally must be from the region where the receptor is located and that should be contemporary to the underpinning ambient air measurements.
- CMB requires a good knowledge of the emissions in the study area in order to assure that all relevant sources are included and to evaluate their uncertainty.
- CMBs is sensitive to collinearity of the source profiles, which impedes the mathematical solution of the mass balance, often it is necessary to merge sources into groups of source types in order to produce composite profiles. This exercise automatically builds in intrinsic assumptions into the CMB model.
- From the mathematical point of view, CMB can be carried out with one single sample. In general more samples are needed but less than those required by factor analytical methods. for a limited number of samples.
- However, small data sets may not fully characterize the source-receptor relationships at a given site.

CMB application and validation steps

(1) assessing the general applicability of the CMB model to the situation under study;
(2) configuring the model with appropriate sources, source profiles, and chemical species concentrations at receptor sites;
(3) examining model statistics and diagnostics;
(4) determining agreement with model assumptions;
(5) identifying problems, changing the model configuration and rerunning;
(6) testing the consistency and stability of model results; and
(7) evaluating the validity of model results by comparing them with other receptor or dispersion model results.

There are four main categories of situations that can be addressed to improve model performance. These are:
(1) incorrect ambient data,
(2) incorrect source profiles,
(3) incorrect source list, and
(4) profile uncertainty collinearity.

(Thomson and Watson, 1987)

Identification of relevant sources and selection of source profiles (CMB)

Source Identification

(1) include ubiquitous area sources, such as motor vehicle exhaust, residual oil combustion, and resuspended dust. These sources are almost universally present in all urban areas;

(2) include natural sources, such as sea salt, if the receptor is in an area likely to be affected by such sources;

(3) include point sources which have been identified from an emissions inventory; and

(4) include "single constituent source types" in cases where substantial amounts of secondary nitrate, sulfate, and organic carbon are expected. These are profiles which represent only a single compound such as sulfate.

Identification of relevant sources and selection of source profiles (CMB)

Source profile representativeness and uncertainty

- review wind direction data and eliminate sources downwind

- eliminate those source types which are not likely to be emitting during the period of time being studied (e.g. woodsmoke emissions during hot summer months)

- eliminate those sources or source categories that are minor contributing sources

-select only one source profile per source type, it must also represent the range of variability expected from a number of individual emitters in the same source type category.

- this variability must be reflected in the uncertainties

CMB related methods

Methods related to CMB are:

- Non Negative Least Squares (Wang and Hopke, 1989) and
- Partial Least Squares Regression, which is a generalization of Multiple Linear Regression (MLR) suitable for analysing data with collinear, noisy, and numerous x-variables (Vong et al., 1988).

Receptor Models

Factor Analysis methods

(source number and composition unknown)

-eigenvector methods PCA (Blifford and Meeker, 1967)
UNMIX (Henry and Kim, 1989)

Principal component Analysis (PCA) is based on singular value decomposition.
It estimates X that gives the lowest possible value for:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} e^{2}{}_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[x_{ij} - \sum_{p=1}^{P} g_{ip} f_{jp} \right]^2$$

Eigenvector analysis is an implicit least squares analysis.

Scaling to normalize data led to distortions in the analysis

PCA characteristics

- PCA is to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components (PCs)
- In analyses with PCA the first principal component accounts for as much of the variability in the data as possible, and each succeeding component in turn has the highest variance possible under the constraint that it be uncorrelated with the preceding components .
- PCA is sensitive to the relative scaling of the original variables and is based upon the intrinsic assumption that the data set jointly is normally distributed.
- The artificial positioning of variance into the first few components can be partially solved by orthogonal rotations (e.g., varimax).

PCA characteristics 2

- To uncenter PCs a zero-valued pseudosample is subtracted and then regressed against the total PM mass (Absolute Principal Component Scores, APCS (Thurston and Spengler, 1985), Absolute Principal Component Analysis, APCA (Swietlicki and Krejci, 1996) and PCA-MLR (Tauler et al., 2008).

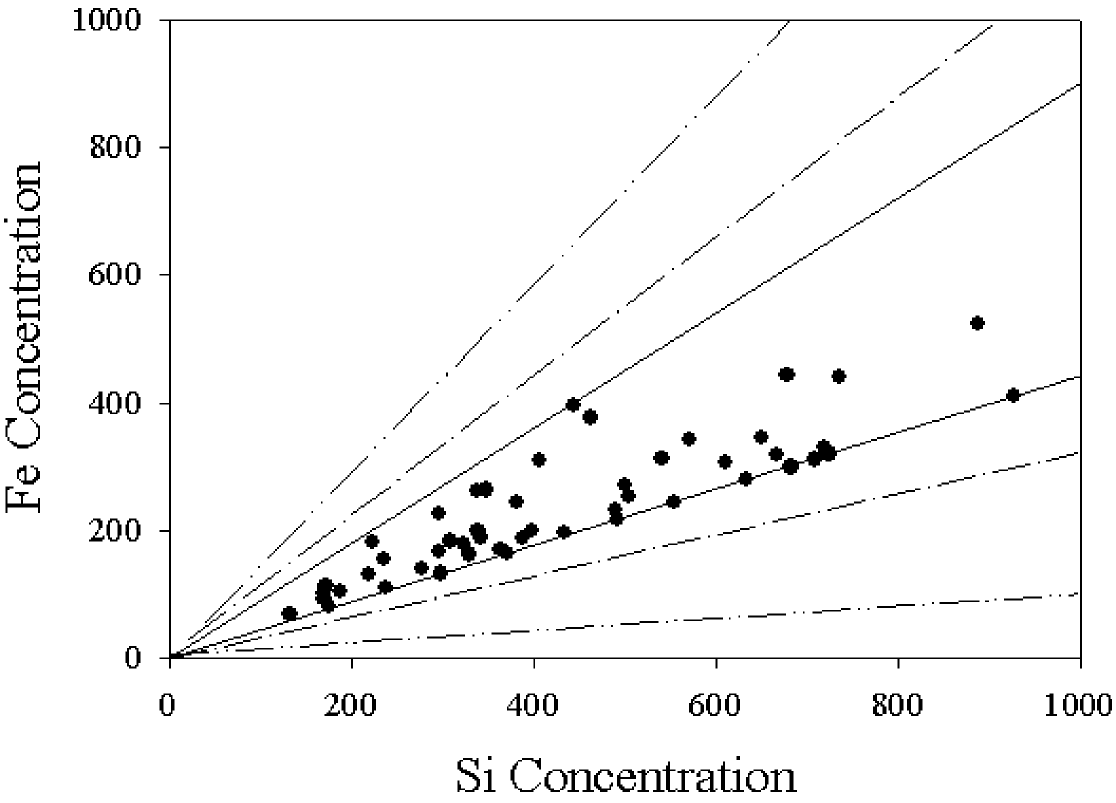
- PCA does not perform explicit data uncertainty treatment. Therefore, noise deriving from the uncertainty structure of the datasets is incorporated by PCA into the PCs

- the basic assumption for PCA of orthogonal component does not reflect the structure of real world data (many source profiles have a degree of collinearity)

UNMIX characteristics

- Uses eigenvalue analysis to reduce the dimensionality of the dataset without centring the original data
- The number of PCs is estimated by the NUMFACT algorithm (Henry, 1997) relying on the signal to noise ratio of PCs in advance.
- An edge-finding algorithm based on Self-Modelling Curve Resolution (SMCR) techniques is applied.
- Edges are hyperplanes determined by points in which a source profile is absent or has a very low relative contribution.
- Edges are used as explicit physical constraints to define a region of the real solution where source contributions are greater than or equal to zero.
- UNMIX does not incorporate errors into the analysis and suffers from some of the same concerns as PCA.

Ratios



original slide from P. K Hopke

Factor analysis and PCA

- PCA and FA are similar in the way they operate linear transformation of the original variables to create a new set of variables, which better explain cause-effect patterns
- PCA aims to maximize the variance by minimizing the sum of squares, FA relies on a definite model including common factors, specific factors and measurement errors.
- PCA has a unique solution while factors in FA are not exact linear functions of x.
- In PCA, variables are almost independent from each other while common factors (communalities) contribute to at least two variables
- 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.
- FA is considered more efficient than PCA in finding the underlying structure of data (Joliffe, 2002). However, PCA and FA produce similar results when there are many variables and their specific variances are small.

Receptor Models

Factor Analysis methods 2

(source number and composition unknown)

- unconstrained FA

- Positive constraints
  explicit least squares fit PMF (Paatero, 1997)

Positive Matrix Factorization (PMF) scales each data individually, more precise data have more influence

$$Q=\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{x_{ij}-\sum_{k=1}^{p}g_{ik}f_{kj}}{u_{ij}}\right)^{2}\rightarrow Q=\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{e_{ij}}{u_{ij}}\right)^{2}$$

PMF points to minimize Q with respect to g and f
with the constraints that each of the elements of g and f is to be non negative
However there is still rotational ambiguity

-multivariate curve resolution alternating least squares
MCR-ALS (Tauler et al.,1995; Tauler et al.,2009)

Positive Matrix Factorization characteristics

Strengths

- No need for *a priori* assumptions of sources with source profiles
- Good at isolating minor contributors to mass
- Gives a range of solutions for interpretation (analyst can inspect the range to determine robustness of solutions)
- No need to select "relevant" species
- Accounts for uncertainty in every sample

Weaknesses

- Requires large data matrix
- Depends on source profiles which may vary with time
- Often "secondary" factors that are not true sources are isolated
- Gives a wide range of solutions for interpretation (need to determine which solutions make sense within the conceptual model)
- Highly collinear sources are difficult to isolate

Outputs profiles and factor strengths which must be processed and interpreted by the user

Brown and Hafner, 2005

Positive Matrix Factorization characteristics

- In PMF non negativity constraints are set to avoid solutions without physical meaning (negative contributions).
- PMF incorporates the measurement uncertainties in the model. Every single value is scaled
- The original version used the alternating least squares iterative method (Paatero and Tapper, 1993), but convergence was very slow and a faster algorithm was developed by computing G and F matrices simultaneously (PMF2). In version 3 the PMF problem is solved using the Conjugate Algorithm.
- FPEAK is used to test the rotational ambiguity
- Fkey is used to introduce small additional constraints
- Uncertainty and stability of the solution is estimated with bootstrapping
- The new method of displacement has been recently developed to furhter explore the rotational ambiguity

Receptor Models

Two categories of hybrid methods have been used for SA of PM:

i) Constrained or expanded receptor models;

ii) Trajectory based receptor models.

Hybrid methods (sources partially known)
(three way least squares)
-Conjugate gradient algorithm PMF with ME (Paatero, 1999)

Multilinear Engine (ME) is a flexible programme used to further develop the PMF approach by applying the conjugate gradient algorithm (three way least squares) .

-Weighted least square methods: COPREM (Wåhlin, 2003)

Extended Factor Analysis Models

- Classical factor analysis is performed on a two dimensional matrix (two way model)
- This approach was extended to solve “n way” models applying the conjugate algorithm.
- The Multilinear Engine platform is suitable to deal with this kind of approach.
- Extended models further reduce the rotational ambiguity by adding additional constraints (e.g. known source contributions, known source profiles, etc.)

Least squares minimization solutions are often referred to as “the CMB” model, but PMF and UNMIX are also solutions, not separate models

TRACER-CMB: $S_j = C_i / F_{ij}$

Tracer solution, Hidy and Friedlander (1971), Winchester and Nifong (1971), single sample

OWLS-CMB: $\varkappa^2 = \min \Sigma_i \left[(C_i - \Sigma_j F_{ij} S_j)^2 / \sigma_{Ci}^2 \right]$

Ordinary Weighted Least Squares, Friedlander (1973), single sample

EV-CMB: $\varkappa^2 = \min \Sigma_i \left[(C_i - \Sigma_j F_{ij} S_j)^2 / (\sigma_{Ci}^2 + \Sigma_j \sigma_{Fij}^2 S_j^2) \right]$

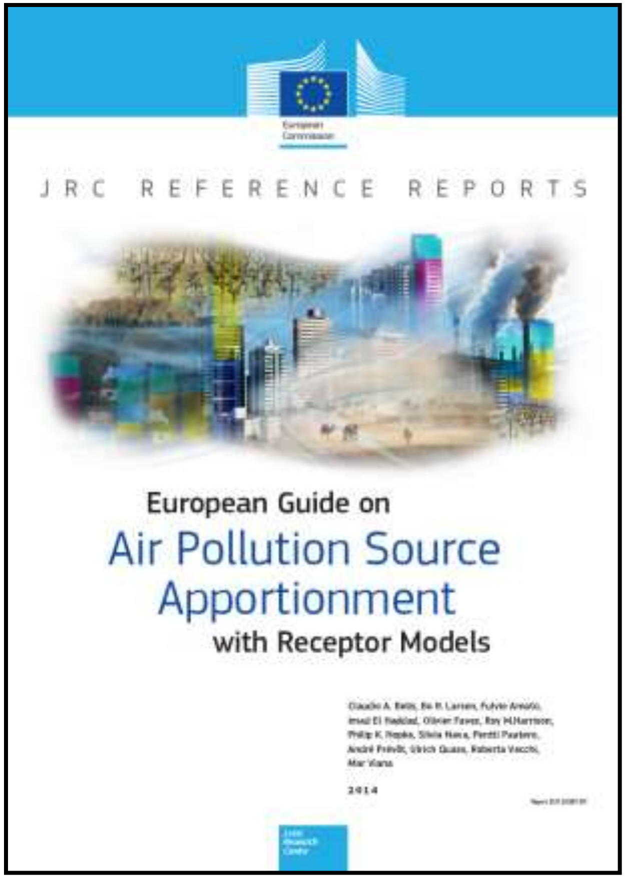
Effective Variance, Watson et al., (1984), single sample

PMF-CMB: $\varkappa^2 = \min \Sigma_i \, \Sigma_k \left[(C_{ik} - \Sigma_j F_{ij} S_{jk})^2 / \sigma_{Cik}^2 \right]$

Positive Matrix Factorization, Paatero (1997), multiple samples

original slide from J. Watson

Common Protocol: Driving elements



European Commission

JRC REFERENCE REPORTS

European Guide on
Air Pollution Source Apportionment
with Receptor Models

Claudio A. Belis, Bo R. Larsen, Fulvio Amato,
Imad El Haddad, Olivier Favez, Roy M.Harrison,
Philip K. Hopke, Silvia Nava, Pentti Paatero,
André Prévôt, Ulrich Quass, Roberta Vecchi,
Mar Viana

2014

- The main objective is to promote the best available operating procedures and to harmonize their application across Europe.
- Promote implementation of the protocol in new studies
- Establish a feed-back mechanism from users in Ms
- Schedule dissemination and capacity building activities

Common Protocol: Driving elements

- The main objective is to promote the best available operating procedures and to harmonize their application across Europe.
- There are sections targeted to customers interested in source contribution estimations for abatement measures design
- The text is structured in different levels of complexity according to the reader skills
- Contains tutorials, technical recommendations and check lists
- It is not meant to report all the information but to orient the reader to the relevant information sources

COMMON RM PROTOCOL

Common Protocol Outline

PART A: INTRODUCTION TO SOURCE APPORTIONMENT WITH RECEPTOR MODELS

Presents the work and provides the unskilled reader with basic elements on Source Apportionment and Receptor Modelling

PART B: STANDARD RECEPTOR MODEL TECHNICAL PROTOCOL

Is the core of the document. Contains description of the steps required in the most traditional and widespread Receptor Modelling techniques with particular reference to CMB and Factor Analysis

PART C: ADVANCED TOOLS

This section contains innovative and advanced methods most of which under continuous development. Also methods on trajectories that although have been available for long time their potentials have not been completely exploited

COMMON RM PROTOCOL

Common Protocol: structure

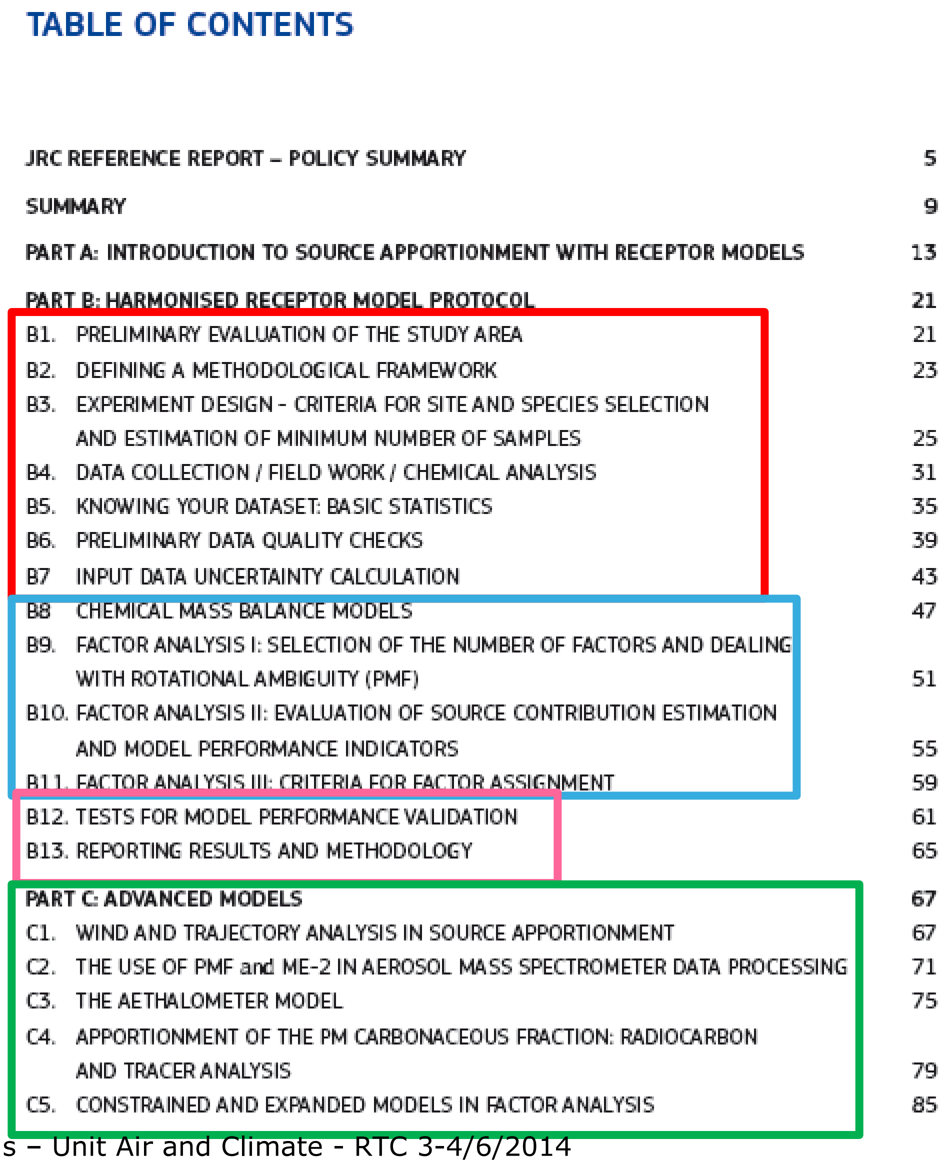
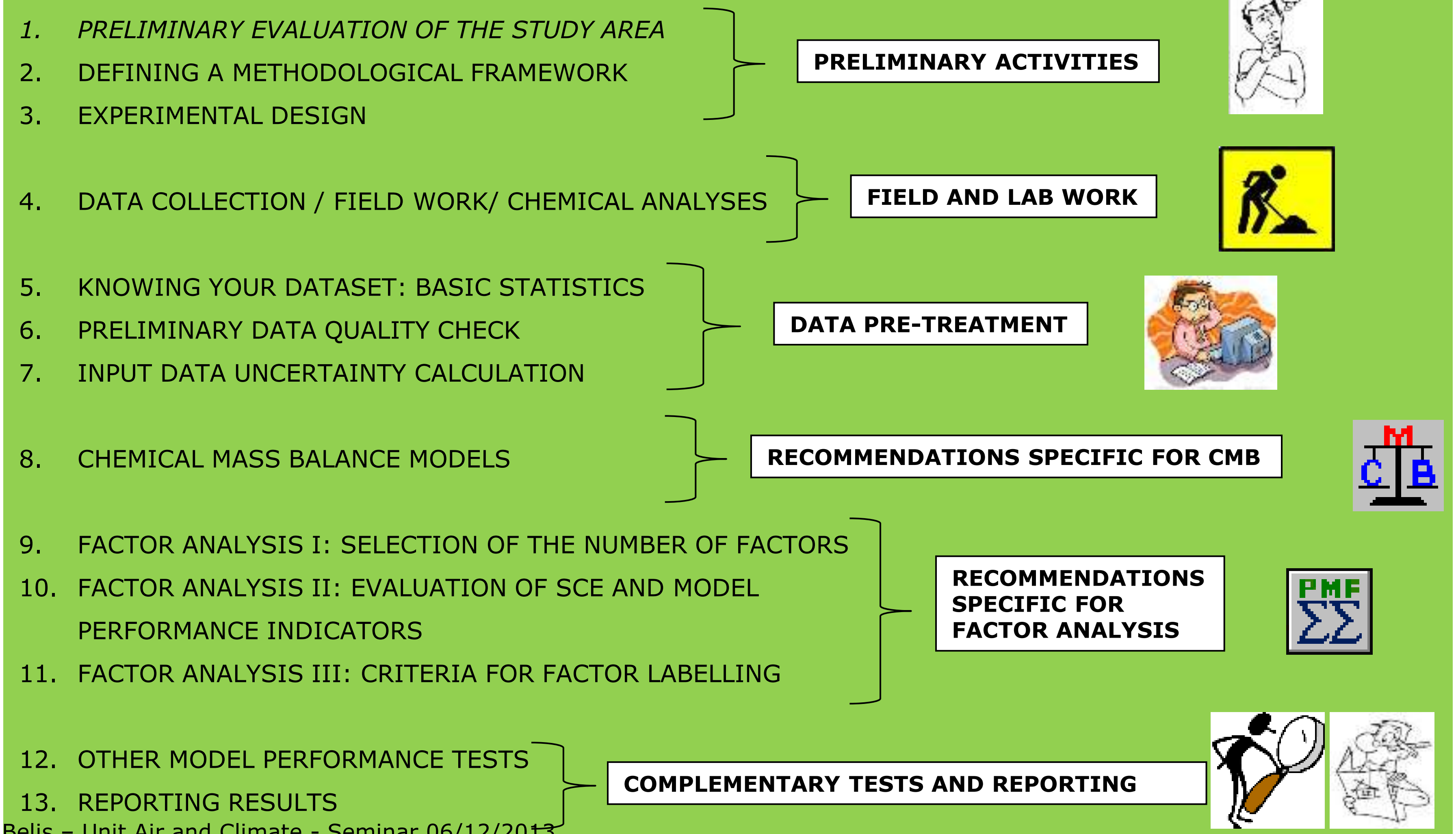


TABLE OF CONTENTS

COMMON RM PROTOCOL

PART B: STANDARD RECEPTOR MODEL TECHNICAL PROTOCOL



1. *PRELIMINARY EVALUATION OF THE STUDY AREA*
2. DEFINING A METHODOLOGICAL FRAMEWORK
3. EXPERIMENTAL DESIGN

PRELIMINARY ACTIVITIES

4. DATA COLLECTION / FIELD WORK/ CHEMICAL ANALYSES

FIELD AND LAB WORK

5. KNOWING YOUR DATASET: BASIC STATISTICS
6. PRELIMINARY DATA QUALITY CHECK
7. INPUT DATA UNCERTAINTY CALCULATION

DATA PRE-TREATMENT

8. CHEMICAL MASS BALANCE MODELS

RECOMMENDATIONS SPECIFIC FOR CMB

9. FACTOR ANALYSIS I: SELECTION OF THE NUMBER OF FACTORS
10. FACTOR ANALYSIS II: EVALUATION OF SCE AND MODEL PERFORMANCE INDICATORS
11. FACTOR ANALYSIS III: CRITERIA FOR FACTOR LABELLING

RECOMMENDATIONS SPECIFIC FOR FACTOR ANALYSIS

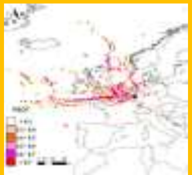
12. OTHER MODEL PERFORMANCE TESTS
13. REPORTING RESULTS

COMPLEMENTARY TESTS AND REPORTING

COMMON RM PROTOCOL

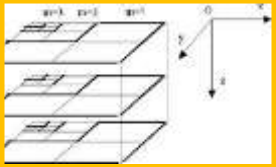
PART C: SPECIFIC TOPICS

1. TRAJECTORY ANALYSIS IN SOURCE APPORTIONMENT



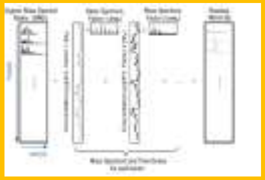
Combination of trajectories and wind direction analysis with receptor models makes it possible to evaluate the geographic provenience of sources. These techniques are also useful for RM output validation.

2. CONSTRAINED AND EXPANDED MODELS IN FACTOR ANALYSIS



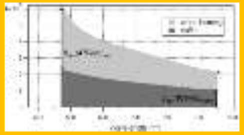
These models represent the new frontier of RM since make it possible to combine different type of data and make advanced data treatment

3. THE USE OF PMF IN AMS DATA PROCESSING



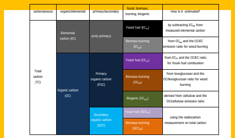
Thanks to the application of (PMF) to find factors this methodology has many analogies with the traditional RMs. The experience gained by the community working with these tools may be useful for RM experts and viceversa.

4. THE AETHALOMETER MODEL



This is a promising technique that opens the opportunity for mutual validation with traditional RMs.

5. CARBONACEOUS FRACTION: RADIOCARBON AND TRACER ANALYSIS



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 elemental carbon
			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 on total carbon
			Biomass burning (SOC_{bb})	

Radiocarbon analyses allow the distinction betweeen fossil and recent sources of carbon, macrotracers can be used to distinguish primary, secondary, biogenic fractions

COMMON RM PROTOCOL

SA studies can be considered as being consistent with the present protocol if :

1. The results are described according to the steps proposed in sections B1- B12.
2. Expert decisions are described and evidence of the objective information that support them is provided. (essential for critical steps).
3. The documentation includes the references of the source profiles used as input or to validate factor assignment.
4. The model and version used are clearly reported.
5. The quantitative uncertainty of the output is estimated and reported.
6. Estimation of overall uncertainty and validation is achieved by comparing outputs from independent models on the same dataset and/or using Monte Carlo permutation and/or displacement analysis techniques.
7. Sensitivity analysis is performed to demonstrate that there are no substantial deviations from the mass conservation assumption.
8. Only solutions that implement the quality assurance steps described in this guide can claim state-of-the-art performance supported by community-wide intercomparison exercises.

THANKS TO ALL THE COLLEAGUES WHO CONTRIBUTED TO THIS INTIATIVE

European common protocol for receptor models:

B. R. Larsen, F. Amato, O. Favez, I. El Haddad, R.M. Harrison, A.S.H. Prévôt, S. Nava, U. Quass, R. Vecchi, M. Viana, P. Paatero

FACTOR ANALYSIS SPECIFIC PROCEDURES IN THE COMMON PROTOCOL FOR RECEPTOR MODELS

SELECTION OF THE NUMBER OF FACTORS

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
- Each good (not weak) data point, contributes a value of approximately 1 to the Q-value
- 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).

SELECTION OF THE NUMBER OF FACTORS

Examining the Q-value (cont)

- **Q(robust)** is calculated excluding outliers
- **Q(true)** includes all points

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.

Solutions where Q(true) is 1.5 times greater than Q(robust) may indicate that the model is inconsistently modelling the data

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.

SELECTION OF THE NUMBER OF FACTORS

Examining the scaled residuals

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

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

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).

Too large scaled residuals for certain variables →, uncertainties specified for these variables are too small.

Too small scaled residuals (close to zero) for one variable:

a) large uncertainties have been specified or

b) this variable is explained by a unique factor.

A strong skewness in the scaled residual plots suggests that the fit is not correct and that other solutions should be sought

SELECTION OF THE NUMBER OF FACTORS

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.

- 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.

SELECTION OF THE NUMBER OF FACTORS

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 .
- To overcome this issue, regressions with weighted values should be calculated manually

SELECTION OF THE NUMBER OF FACTORS

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.

SELECTION OF THE NUMBER OF FACTORS

Rotational Ambiguity

Solutions with reduced RA are likely those with the proper number of factors

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.
- In the absence of any other consideration such as G-space, and unless there is a substantial improvement in the interpretability of the profiles, the best fit is given by FPEAK = 0.0.

SELECTION OF THE NUMBER OF FACTORS

Rotational Ambiguity

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.
- Additional flexibility in applying external constraints are available in the new version of the EPA PMF.

SELECTION OF THE NUMBER OF FACTORS

Rotational Ambiguity

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.

- 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.

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.

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.

Factor 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

CRITERIA FOR FACTOR ASSIGNMENT

Proposed steps to support factor assignment

Compare the obtained factor profiles with those reported in previously published PMF studies;

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);

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;

Perform local and/or regional source sampling along with the ambient PM sampling to develop source profiles needed to identify PMF profiles;

CRITERIA FOR FACTOR ASSIGNMENT

Proposed steps to support factor assignment (cont)

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 (see also section B12).

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 level of 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.

TESTS FOR MODEL PERFORMANCE VALIDATION

The fundamental, natural physical constraints that must be fulfilled in any source apportionment study are as follows:

- 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.

TESTS FOR MODEL PERFORMANCE VALIDATION

Ratios

- Unique source tracers are rare
- 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.
- 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.

TESTS FOR MODEL PERFORMANCE VALIDATION

Ratios (cont)

The species examined should have similar reactivity with respect to atmospheric oxidants and solar radiation

Also particle size distributions is important to exclude differences in particle scavenging by precipitation or particle dry deposition.

TESTS FOR MODEL PERFORMANCE VALIDATION

Ratios (examples)

- 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. (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).

.

TESTS FOR MODEL PERFORMANCE VALIDATION

Residuals

- A well-modelled species instead shows normally distributed residuals within the range +3 and -3.
- Many large-scaled residuals or displays a non-normal curve, may be an indication of a poor fit.
- In weighted ‘least squares’ analysis, weighted residuals must be used in graphical residual analysis, so that the plots can be interpreted as usual. (not available in EPA PMF v3 default residual graphs).
- 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.

TESTS FOR MODEL PERFORMANCE VALIDATION

Residuals (cont)

- 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).
- 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).

TESTS FOR MODEL PERFORMANCE VALIDATION

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.
-

TESTS FOR MODEL PERFORMANCE VALIDATION

Time trends

- The user can further explore their time variability in relation to concentrations of
- gaseous pollutants such as SO2, CO and NOx 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
- NH3 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 PM2.5 to carry out a PCA analysis.

TESTS FOR MODEL PERFORMANCE VALIDATION

A posteriori wind direction and trajectory 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.
- 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.
- The **nonparametric regression analysis** technique 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).

TESTS FOR MODEL PERFORMANCE VALIDATION

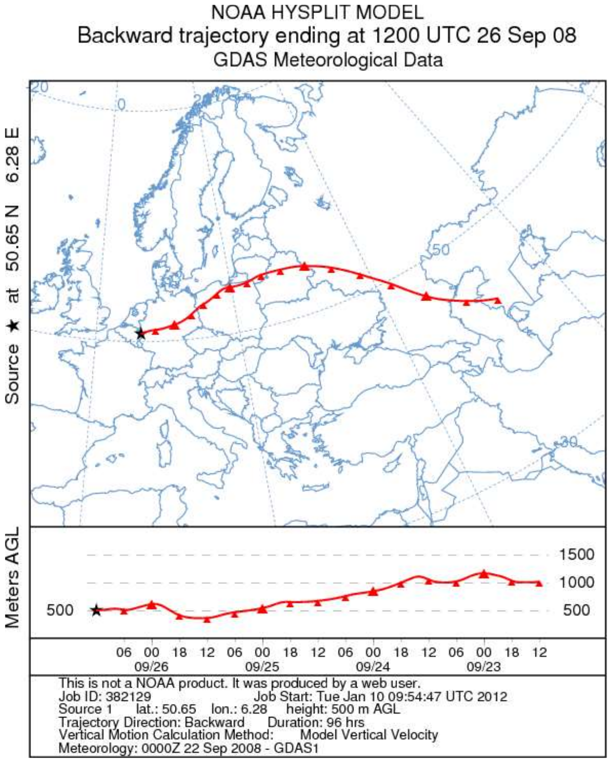
Backward Trajectory Analysis

- In a source apportionment study, back trajectories can be used either to:
  - ✓ pre-select datasets for analysis (e.g. in cases where specific sources and source regions are of major interest) or,
  - ✓ to check the plausibility of identified sources/processes and to get information about their geographical distribution and locations.
- The HYSPLIT model can process different meteorological file types that may also be downloaded via the program
- In addition to the models operated commercially by national weather forecast organisations, there is a variety of research-oriented models available in Europe which allow back-trajectory plots to be produced, e.g. FLEXPART (NILU); REM-CALGRID (TRUMF), EURAD (RIU).
- Potential Source Contribution Function (PSCF)

TESTS FOR MODEL PERFORMANCE VALIDATION

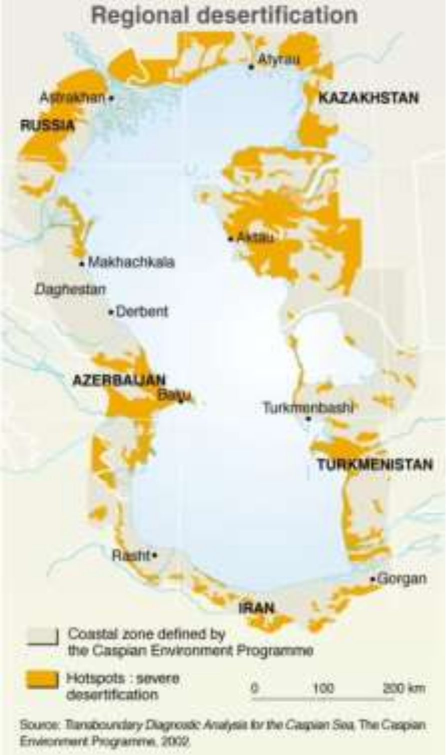
Backward Trajectory Analysis (example)

Figure C1.1. 96-hour back trajectory calculated with HYSPLIT 4.9 for a day with high mineral dust concentration (left, Draxler, 2012), pointing to arid source regions close to the Caspian Sea (right; Abasova, 2010)



NOAA HYSPLIT MODEL
Backward trajectory ending at 1200 UTC 26 Sep 08
GDAS Meteorological Data

This is not a NOAA product. It was produced by a web user.
Job ID: 382129 Job Start: Tue Jan 10 09:54:47 UTC 2012
Source 1 lat.: 50.65 lon.: 6.28 height: 500 m AGL
Trajectory Direction: Backward Duration: 96 hrs
Vertical Motion Calculation Method: Model Vertical Velocity
Meteorology: 0000Z 22 Sep 2008 - GDAS1



▭ Coastal zone defined by the Caspian Environment Programme

🟧 Hotspots : severe desertification

Source: *Transboundary Diagnostic Analysis for the Caspian Sea*, The Caspian Environment Programme, 2002.

REFERENCES

RECEPTOR MODELS IN EUROPE:

Belis, C. A., Karagulian, F., Larsen, B. R., and 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

RECEPTOR MODEL COMMON PROTOCOL (find specific references at the end of each chapter):

Belis C. A., Larsen B. R., Amato F., El Haddad I., Favez O., Harrison R. M., Hopke P. K., Nava S., Paatero P., Prévôt A., Quass U., Vecchi R., Viana M., 2013. European Guide on Air Pollution Source Apportionment with Receptor Models. JRC Reference Report EUR 26080. Luxemburg Publication Office of the European Union. ISBN 978-92-79-32514-4. doi: 10.2788/9332.

MORE INFO AND DOWNLOADS AT

JRC Source apportionment site :

http://source-apportionment.jrc.ec.europa.eu/

New FAIRMODE website

http://fairmode.jrc.ec.europa.eu/index.html