

Positive Matrix Factorization mediante EPA-PMF5: strumenti per applicazioni avanzate e metodi di analisi dell'incertezza (S. Nava)

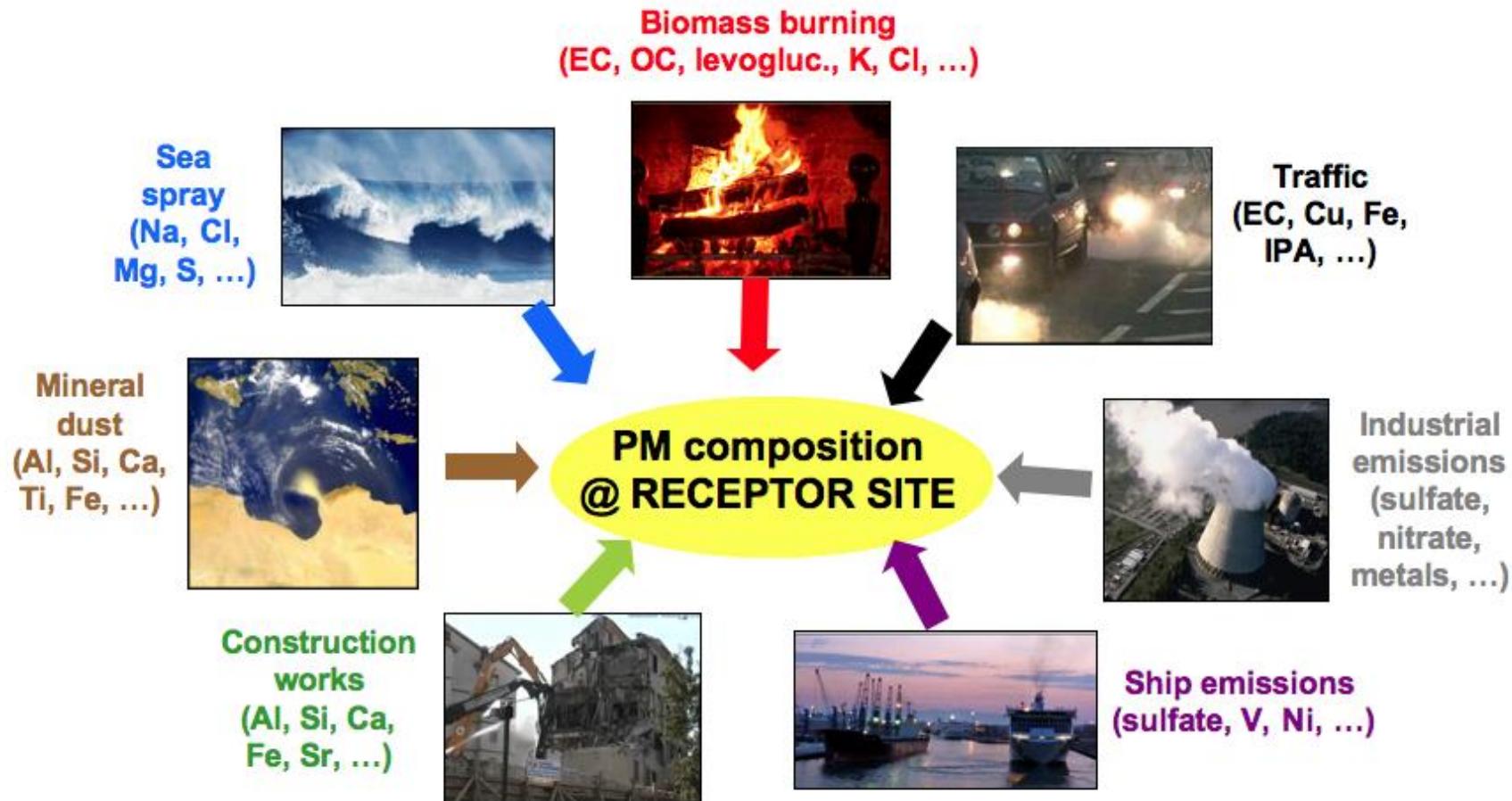
Silvia Nava

**Dip. Di Fisica – Università di Firenze
INFN – Sezione di Firenze
nava@fi.infn.it**



Idea di base dei modelli a recettore

Le particelle di particolato mantengono la composizione elementale/chimica caratteristica della loro origine: la composizione del PM in un sito “recettore” è una combinazione delle composizioni del PM emesso dalle diverse sorgenti.



Equazione del bilancio di massa

Tutti i modelli a recettore si basano sul BILANCIO DI MASSA:

$$x_{ij} \approx \sum_k g_{ik} \cdot f_{kj}$$

concentrazioni misurate

peso della sorgente

profilo della sorgente

somma sui contributi delle diverse sorgenti

x_{ij} = concentrazione della specie j nel campione i

g_{ik} = contributo della sorgente k nel campione i

f_{kj} = frazione della specie j nel particolato prodotto dalla sorgente k

Lo scopo dei modelli a recettore è quello di determinare il peso delle sorgenti (matrice G) a partire dalle concentrazioni misurate (matrice X). In alcuni modelli i profili delle sorgenti (matrice F) sono assunti noti, in altri sono determinati dal modello stesso.

Ipotesi dell'equazione

$$x_{ij} @ \sum_k g_{ik} \times f_{kj}$$

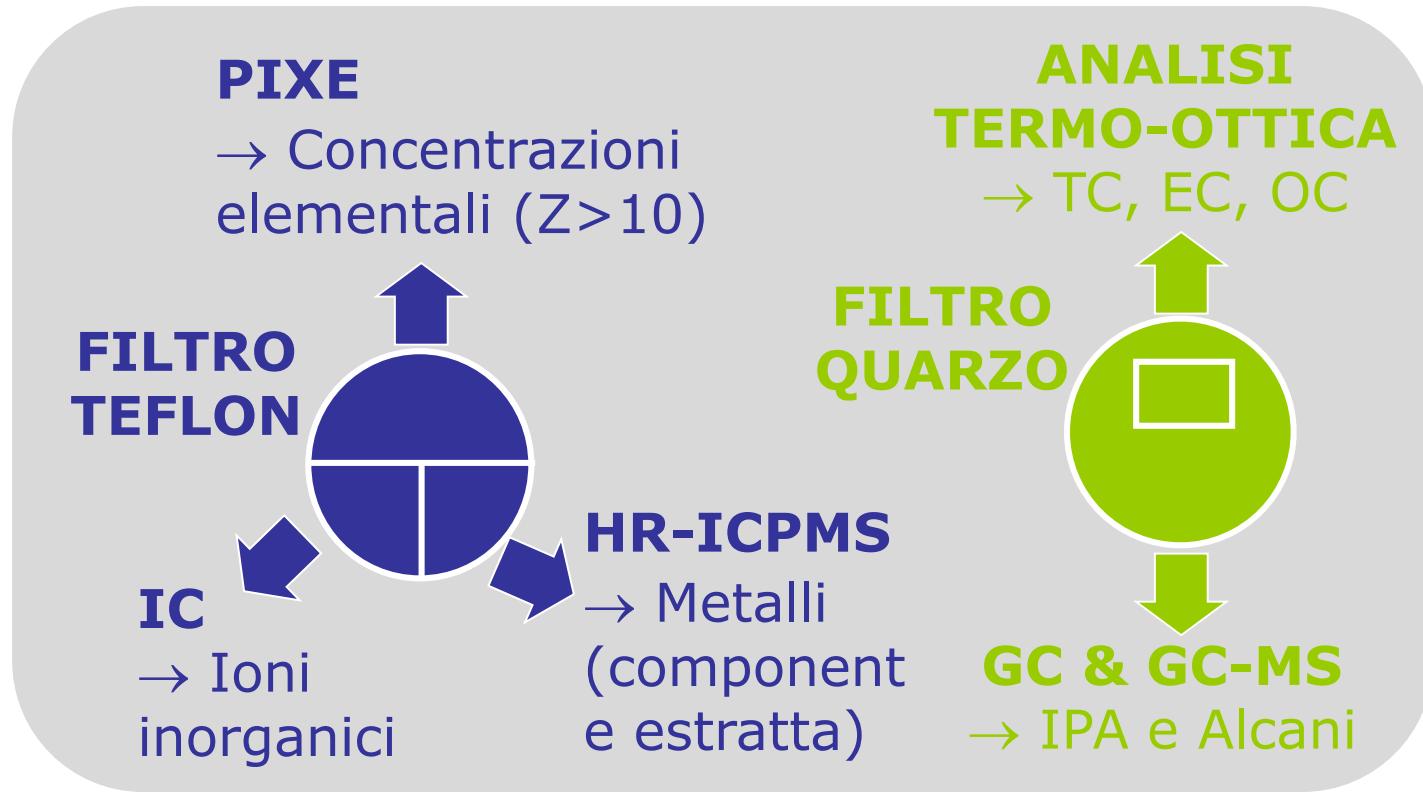
- Il profilo chimico delle sorgenti è costante nel tempo
 - difficoltà nell'identificare ad es. processi industriali con cicli produttivi variabili
- Il profilo chimico delle emissioni è costante nello spazio, ovvero non si trasforma nel tragitto dalla sorgente al recettore
 - è importante che i marker siano specie chimicamente STABILI (ad es. elementi); difficoltà con traccianti organici/volatili
 - problema dei secondari
- I profili e gli andamenti delle sorgenti sono linearmente indipendenti fra loro (non collineari), affinché i loro contributi siano correttamente distinti dal modello
 - sorgenti con profili simili (ad es. local, road and sharan dust) o andamento simile (anche per meteorologia) possono essere accorpate dal modello in un unico fattore: speciazione chimica più dettagliata e maggiore risoluzione temporale
- Sono state identificate tutte le sorgenti che danno un contributo significativo
 - speciazione più completa possibile, con traccianti delle diverse sorgenti

Raccomandazioni generali valide per tutti i modelli a recettore

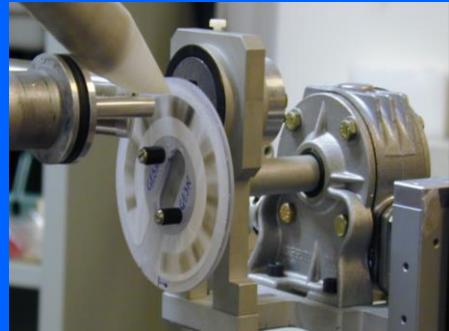
1. Il database da analizzare e il modello non sono due entità a sé stanti
 - Il database deve essere realizzato in modo opportuno progettando accuratamente il campionamento e l'analisi dei campioni in termini di: scelta di punti recettore rappresentativi, numero di campioni, risoluzione temporale del campionamento, specie chimiche misurate.
2. Prima di far girare il modello i dati devono essere ben vagliati ed osservati
 - controllo della qualità dei dati e validazione
 - Osservazione delle serie temporali, statistica di base, correlazioni e scatterplot
 - confronto fra i siti, diagrammi polari (rose dei venti)
3. I risultati del modello devono essere vagliati molto criticamente
 - Confronto con i profili di letteratura e conoscenza del luogo
 - Ragionevolezza degli andamenti temporali delle contributi delle sorgenti (anche in relazione alla loro posizione e alla meteorologia)

Es. di strategia di campionamento e analisi

Campioni
GIORNALIERI
su lungo periodo
(stagioni, anno)
e con diversi tipi
di filtri

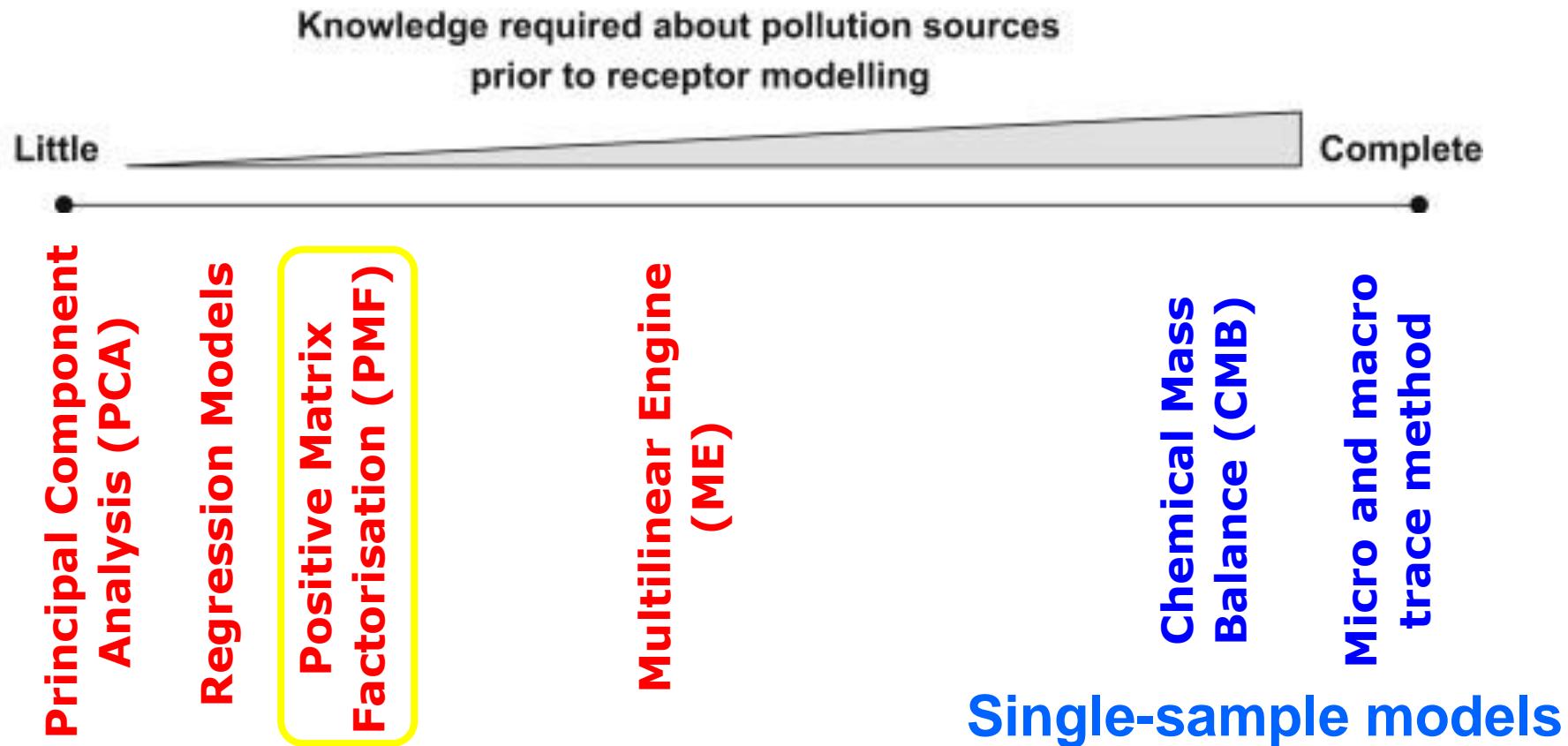


Campioni
ORARI su
periodi più
brevi



STREAKER SAMPLER + PIXE
→ Concentrazioni elementali ($Z>10$)

Classificazione dei modelli a recettore



Multivariate models

Analisi condotta su campioni multipli (l'apporzioneamento è estratto simultaneamente dall'intera matrice dei dati); la maggiore complessità è ripagata dalla quantità di informazioni fornite: non solo i pesi ma anche i profili

Modelli a singolo campione

$$x_{ij} @ \sum_k g_{ik} \times f_{kj}$$

- Analisi indipendente di ogni singolo campione

$$x_1 @ g_1 \times f_{11} + g_2 \times f_{21} + \dots + g_P \times f_{P1}$$

...

$$x_M @ g_1 \times f_{1M} + g_2 \times f_{2M} + \dots + g_P \times f_{PM}$$

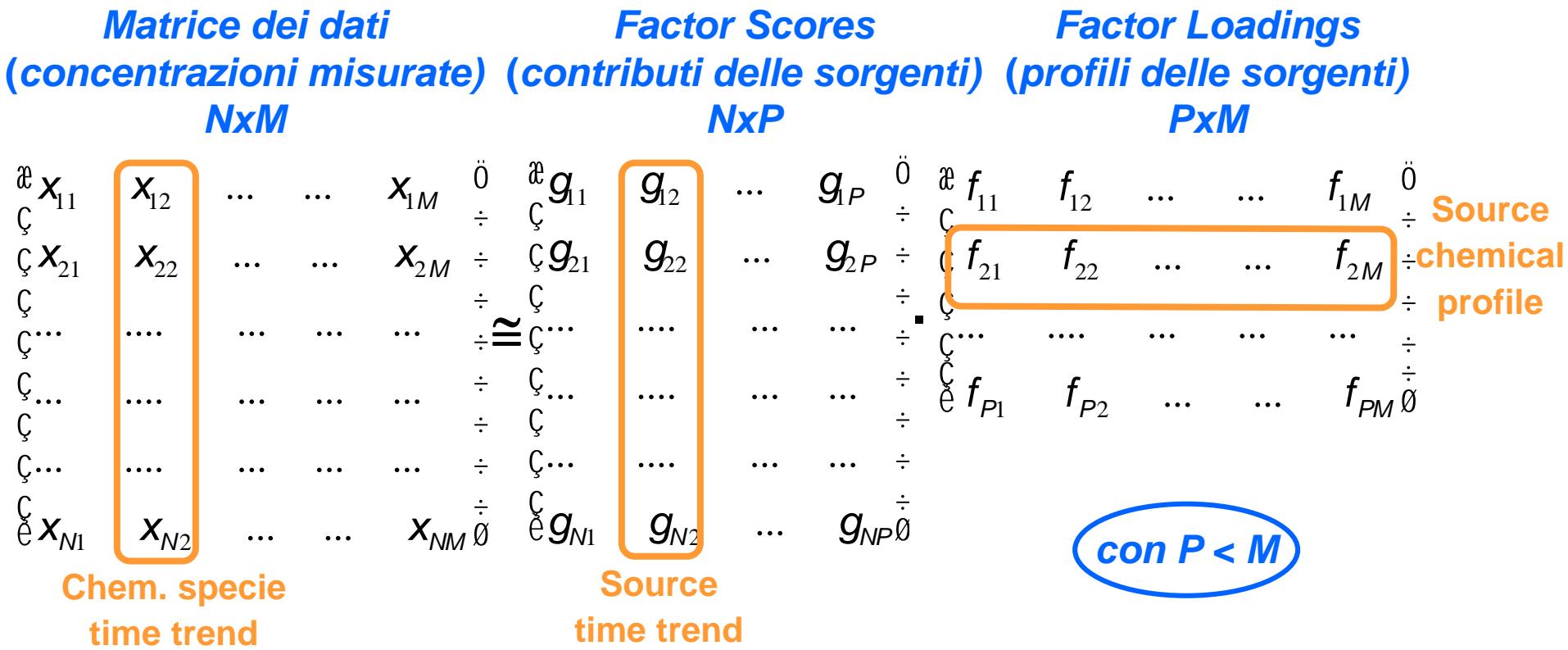
*M specie- P sorgenti
(M equazioni)*

- non ci sono abbastanza equazioni per determinare sia F che G: sono necessarie informazioni sui profili delle sorgenti (F)

- **Metodo dell'elemento tracciante (es. ^{14}C , levoglucosano)**
 - esempio: Biomass burning = f x [levoglucosano]
- **Chemical Mass Balance (CMB)**

Analisi Fattoriale

$$x_{ij} @ \sum_k g_{ik} \times f_{kj}$$



Le matrice dei dati è approssimata dal prodotto di 2 matrici “più piccole”: lo scopo è ridurre la dimensione del data-set senza ridurre troppo l’informazione in esso contenuta

*M variabili originarie:
concentrazioni misurate delle
specie chimiche*



*P < M nuove variabili
(combinazioni lineari delle vecchie variabili):
“componenti” o “fattori”*

Positive Matrix Factorisation (PMF)

- parte direttamente dall'equazione del bilancio di massa: $X = GF (+ E)$, ed utilizza concentrazioni misurate assolute (NON standardizzate o centrate)
- tramite un algoritmo iterativo determina la fattorizzazione con matrici **G** ed **F** positive tali da minimizzare lo scarto fra le concentrazioni misurate e quelle ricostruite dal modello.

METODO DEI MINIMI QUADRATI PESATI, CON VINCOLO DI NON NEGATIVITÀ'

Matrice dei dati

$$X = \{x_{ij}\}$$

Incertezze sperimentalistiche e minimi di rivelabilità

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

F e G non-negative t.c.

$$Q = \sum_{i=1}^N \sum_{j=1}^M \left(\frac{e_{ij}}{S_{ij}} \right)^2 \text{ sia minimo}$$

Matrici dei profili (F) e dei pesi (G) delle sorgenti

Incertezze su F e G

Parametri del modello (numero di fattori, ...)

Minimi quadrati pesati

$$Q = \sum_{i=1}^N \sum_{j=1}^M \frac{e_{ij}^2}{s_{ij}^2} = \sum_{i=1}^N \sum_{j=1}^M \frac{(x_{ij} - \sum_{k=1}^P g_{ik} f_{kj})^2}{s_{ij}^2}$$

Funzione da minimizzare

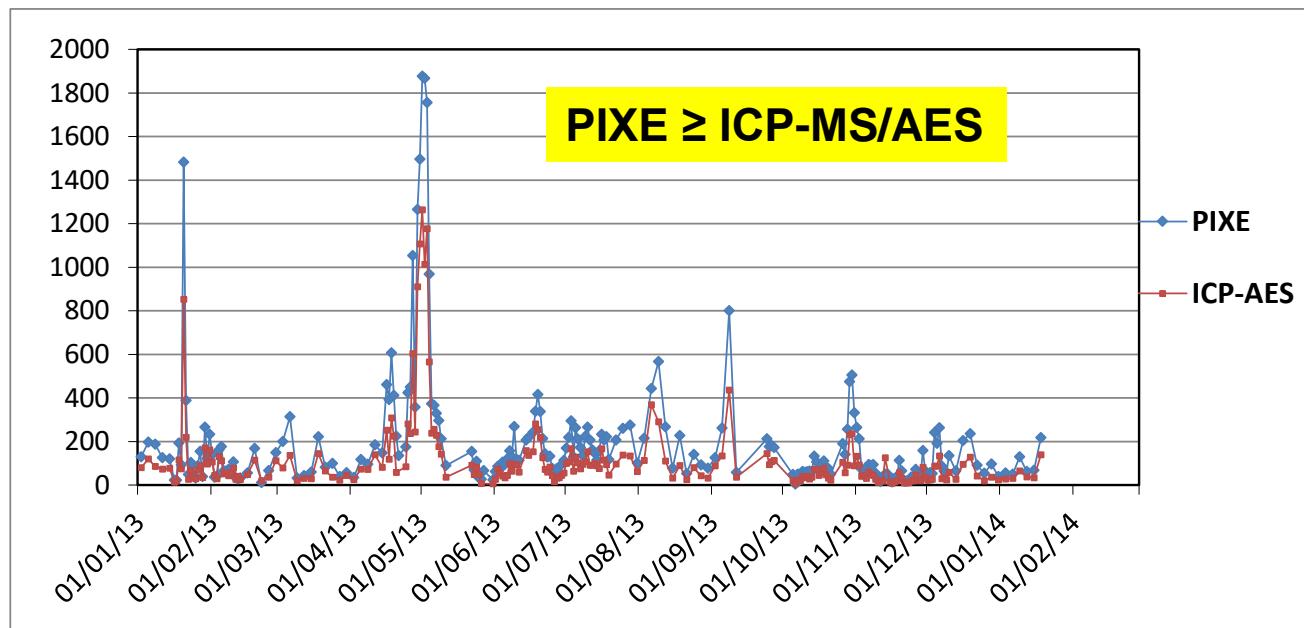
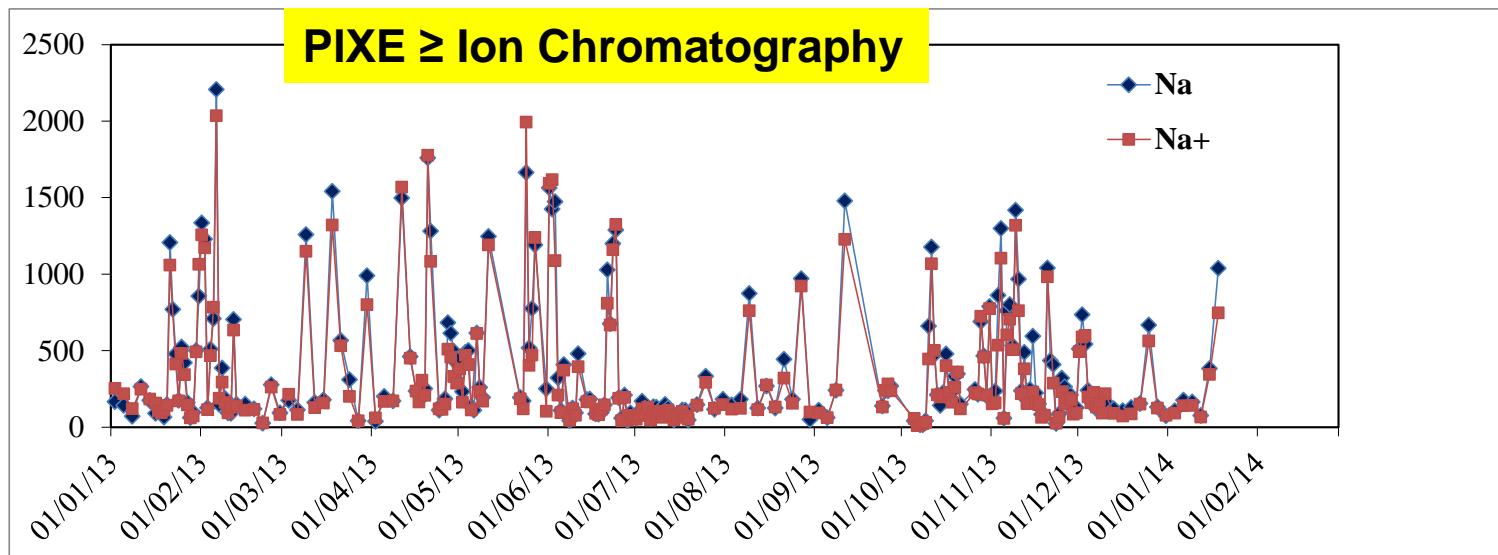
Ogni dato x_{ij} è pesato con uno specifico peso ($w_{ij} = 1/s_{ij}^2$), legato alla sua incertezza sperimentale

- Una formulazione dei minimi quadrati correttamente pesata produce un fit ottimale dei dati.
- La possibilità di inserire ciascun dato con un suo peso individuale permette di sfruttare correttamente ed efficacemente l'informazione contenuta nel data-set, consentendo anche l'inserimento di dati problematici (casi mancanti, casi minori o prossimi al minimo di rivelazione).
- I pesi rendono la funzione Q adimensionale e quindi invariante per cambiamenti di scala (optimal scaling).
- I modelli R e Q di analisi fattoriale comunemente utilizzati sono equivalenti a risolvere un problema di minimi quadrati pesati con errori irrealistici nei dati di input.

Controllo e validazione dei dati

- La preparazione dei dati di input (concentrazioni e incertezze/pesi) è un passaggio fondamentale nella PMF:
 - le concentrazioni devono essere accuratamente validate;
 - le incertezze devono essere valutate realisticamente, in modo da tener conto di tutti i contributi (non solo di analisi ma anche di campionamento)
- Le campagne di studio del particolato producono dataset molto estesi, ottenuti tramite diversi strumenti di campionamento e tecniche di analisi, le fonti di errore possono essere molte (e vengono spesso sottostimate):
 - se alcune specie vengono misurate con più tecniche è fondamentale confrontare i risultati (e scegliere con criterio quali usare nel modello);
 - è molto utile fare calcoli di ricostruzione della massa “mass closure” e di bilancio ionico (da verificare su tutti i singoli campioni);
 - le incertezze delle diverse tecniche/laboratori devono essere “armonizzate”;
 - se una stessa specie chimica viene analizzata con una tecnica/strumento su parte dei campioni e con un’altra tecnica/strumento sugli altri campioni occorre fare attenzione nel “raccordare” i dati.

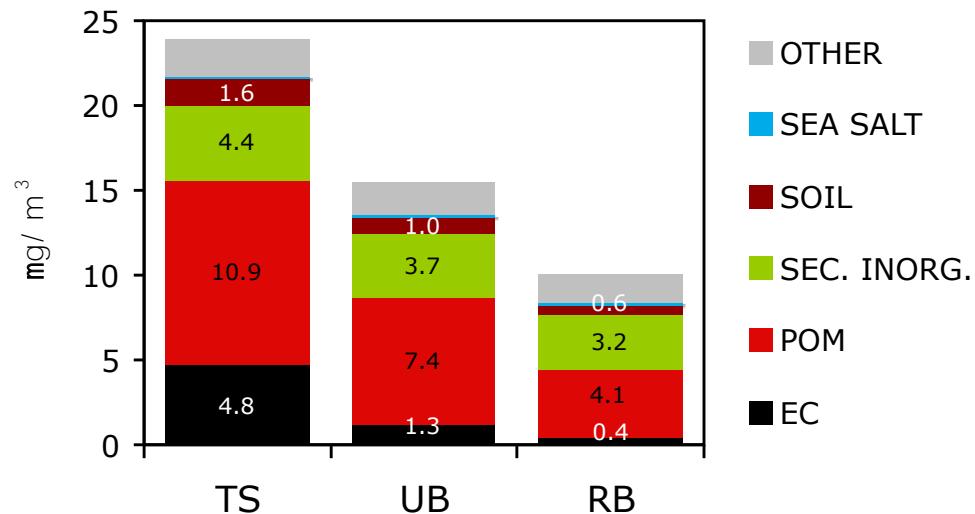
Confronto fra diverse tecniche



Mass closure e bilancio ionico

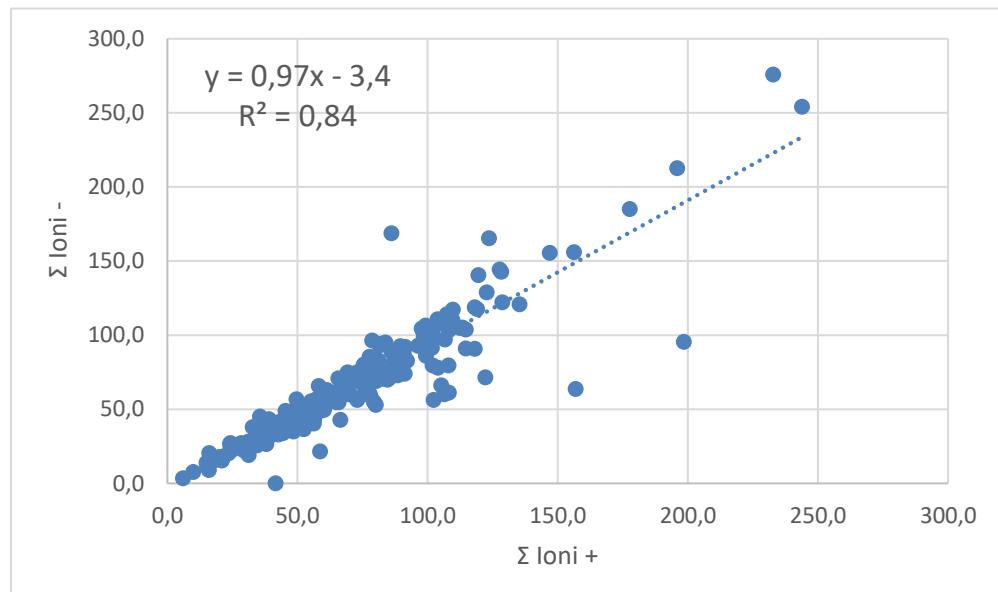
MASS CLOSURE

Utile come validazione (su ciascun campione la massa ricostruita non deve superare quella misurata e le ricostruzioni devono essere “ragionevoli”), e come confronto con risultati PMF



BILANCIO IONICO

Permette di identificare giorni con bilanci ionici “atipici” e quindi scoprire possibili dati non validi.



Numero minimo di campioni

Come tutte le tecniche multivariate, la PMF necessita di un adeguato numero di campioni e specie chimiche.

Il numero minimo di campioni necessari per individuare correttamente le variabili latenti e quantificare accuratamente il loro peso non può essere conosciuto a priori, perché dipende dalla quantità di informazione contenuta nel dataset: campioni con grande variabilità nei contributi relativi delle diverse sorgenti (inclusi campioni con contributi nulli o trascurabili di alcune) rinforzano notevolmente il modello, mentre l'aggiunta di campioni con similari contributi relativi delle sorgenti è di scarsa utilità (per questo sono molto utili campioni ad elevata risoluzione temporale dato che mostrano maggiori differenze nei contributi delle sorgenti).

Ci sono tuttavia delle indicazioni di massima che è importante osservare:

- In accordo con il manuale EPA, **almeno 100 campioni giornalieri e 20 specie chimiche** (Norris et al., 2008, Brown and Hafner, 2005);

Se sono disponibili dataset più piccoli relativi a siti della stessa area o a diverse frazioni di PM, la scelta migliore può essere quella di unirli in un unico input (questo significa ipotizzare stessi profili, i contributi possono essere molto diversi e si possono/devono calcolare per i singoli siti).

Software disponibili

	Pulling	Error estimate	User-friendly
EPA PMF v5			
EPA PMF v3			
PMF2			
ME-2 scripting			

- Programmi PMF3 e PMF5 dell'Environmental Protection Agency (EPA):
<https://www.epa.gov/air-research/positive-matrix-factorization-model-environmental-data-analyses>
- Programmi PMF2, PMF3 e ME sviluppati da P.Paatero (non sono gratuiti)

Rispetto alla precedente versione (PMF3), PMF5 permette:

- l'inserimento di vincoli sui profili e sui contributi delle sorgenti (PULLING),
- 2 ulteriori metodi di stima dell'incertezza sui risultati (DISP e BS-DISP).



EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide

RESEARCH AND DEVELOPMENT



J R C R E F E R E N C E R E P O R T S



European Guide on Air Pollution Source Apportionment with Receptor Models

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

2014

REPORT EUR 29000 EN

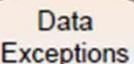
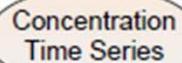
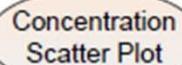
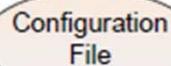
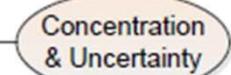
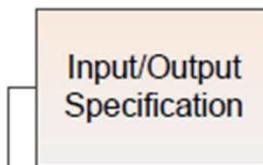


Per stima incertezze:

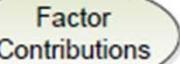
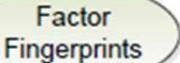
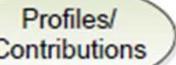
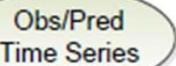
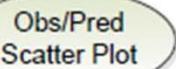
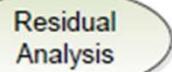
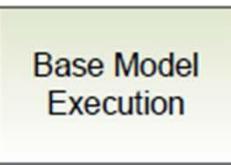
Brown, S.G., Eberly, S., Paatero, P., Norris, G.A. Methods for estimating uncertainty in PMF solutions: Examples with ambient air and water quality data and guidance on reporting PMF results (2015) Science of the Total Environment, 518-519, pp. 626-635.

EPA PMF5.0: struttura del programma

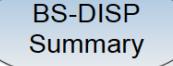
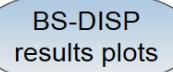
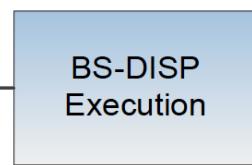
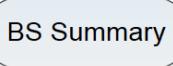
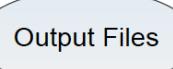
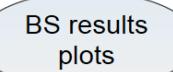
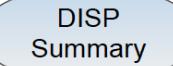
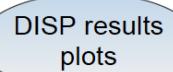
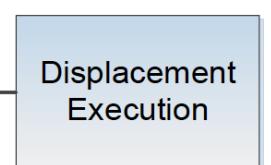
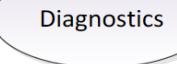
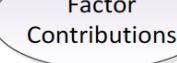
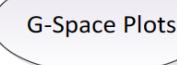
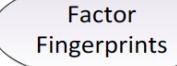
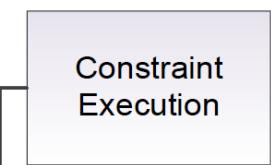
Model Data



Base Model



Rotational Tools (constraints/Peak) and error estimate



Input data

Matrici di input (tab delimited, comma-separated o excel workbooks): una per le concentrazioni e una per i pesi. Per entrambe non sono ammesse celle vuote e la matrice dei pesi può avere solo valori > 0

Concentration Data File: H:\scuola-IAS\prep-dati-PMF_AIRUSE_Firenze_PM10.xls
Concentration data table with parameter names in the first row. Optionally, the second row may contain units and the first column may contain date/time.

Concentration Worksheet: DATA-IAS

Browse Load

Uncertainty Data File: H:\scuola-IAS\prep-dati-PMF_AIRUSE_Firenze_PU.xls
Observation-based or equation-based uncertainty. Must match concentration data format.

Date/Time Column: date ID/Site Column:

Missing Value Indicator: -999 Exclude Entire Sample

La matrice dei pesi può anche essere ricostruita dal modello (“equation-based”). In questo caso occorre fornire una matrice con 3 righe: nomi, MDLs medi ed errori percentuali medi delle diverse specie.

Output Files

Output Folder: H:\scuola-IAS
Specify a destination folder for output files.

Output File Prefix: 9Fbase

Output File Type: Tab-Delimited Text (*.txt) Comma-Delimited Text (*.csv) Excel 97-03 Workbook (*.xls) Excel 07-10 Workbook (*.xlsx)

Output Only Selected Rows

Possibilità di sostituire automaticamente i casi mancanti con la mediana (ed errore 400%).

Save File Locations and Settings in a Configuration File: H:\scuola-IAS\scuolaIAS.config

Load Last Saved Save Save As... Edit

Reset All Exit Program

Diversi approcci di preparazione dei dati

CENSORED: i valori < MDL vengono sostituiti con 0 o con una frazione dell'MDL:
es. Criterio di Polissar [Polissar et al., JGR, 1998]

$$\text{dati noti : } x_{ij} = c_{ij}, \quad s_{ij} = \sigma_{ij} + \frac{MDL_{ij}}{3}$$

$$\text{dati < MDL : } x_{ij} = \frac{MDL_{ij}}{2}, \quad s_{ij} = \frac{\overline{MDL}_j}{2} + \frac{MDL_{ij}}{3}$$

$$\text{dati mancanti : } x_{ij} = c_j, \quad s_{ij} = 4\tilde{c}_j \quad \text{con } \tilde{c}_j \text{ media geometrica}$$

(c_{ij} concentrazione misurata e σ_{ij} incertezza sperimentale)

UNCENSORED: si utilizzano anche le concentrazioni < MDL, anche nel caso i valori misurati siano negativi.

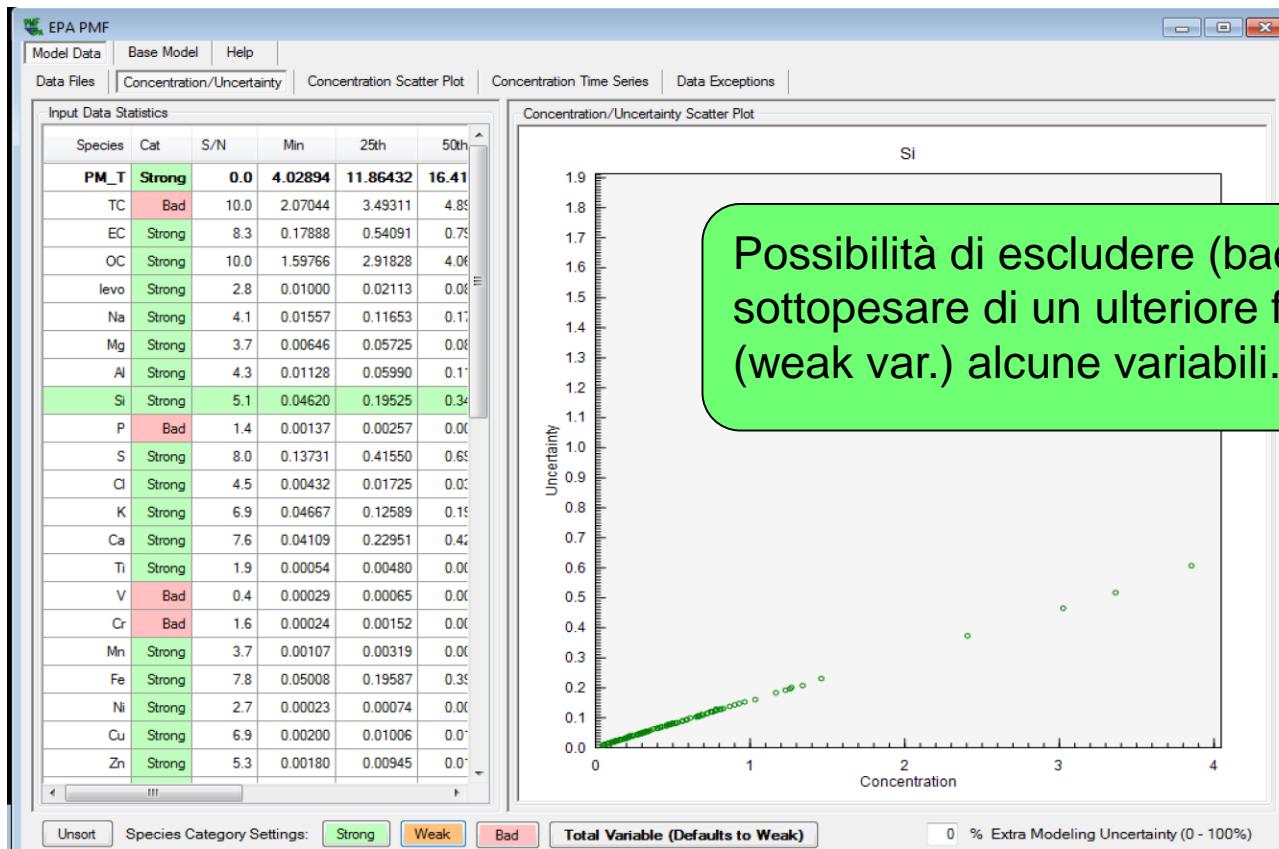
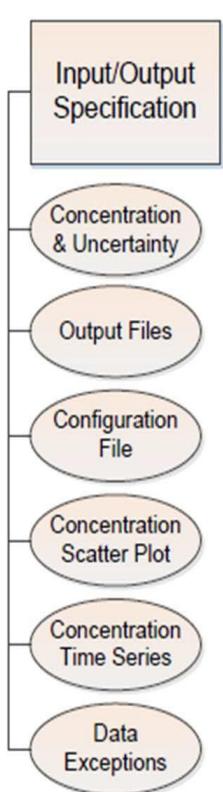
L'incertezza deve includere tutti i possibili contributi della misura (anche quelli del campionamento) e anche delle ipotesi del modello (fluttuazioni dei profili, trasformazioni durante il trasporto, etc.).

Fattori di incertezza costanti sulle colonne o righe di X non influenzano Q: possono bypassare il processo di fit ed essere inseriti a posteriori.

La presenza di molti casi inferiori al DL in diverse variabili può portare il modello a generare fattori "fantasma", prodotti proprio per fissare valori inferiori al DL che si presentano simultaneamente in queste variabili.

Selezione delle specie chimiche

- Evitare doppi conteggi (es. S-SO₄), a meno che non aggiungano informazione (es.K-K⁺)
- Escludere specie analiticamente poco robuste (valori anomali e presenza di molti dati vicini o inferiori al DL), ma attenzione a non eliminare informazione utile, dato che la presenza di più elementi, anche se non traccianti, può essere fondamentale per risolvere problemi di collinearità.



Possibilità di escludere (bad var.) o sottopesare di un ulteriore fattore 3 (weak var.) alcune variabili.

Rapporto segnale/rumore

Il rapporto S/N fra il segnale di interesse (S) e il rumore di fondo (N) può essere usato per valutare la robustezza delle variabili, ma **NON è assolutamente l'unico parametro da considerare e va stimato con attenzione.**

L'espressione utilizzata in EPA PMF3 (rapporto fra la somma in quadratura delle concentrazioni e delle incertezze) produceva stime di S/N artificiosamente alte o artificiosamente basse.

In EPA PMF5 è stata introdotta una nuova espressione:

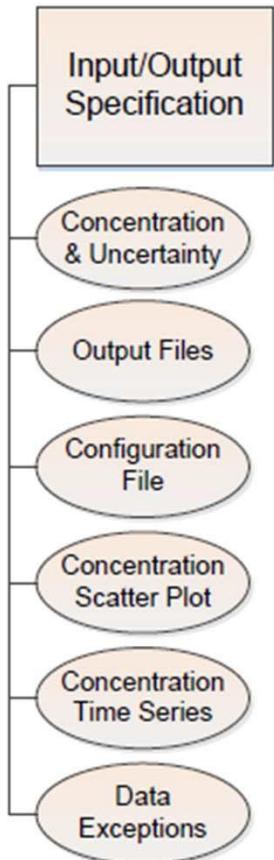
$$\left(\frac{S}{N} \right)_j = \frac{1}{n} \sum_{i=1}^n d_{ij} \quad \left\{ \begin{array}{ll} d_{ij} = \left(\frac{x_{ij} - s_{ij}}{s_{ij}} \right) & \text{if } x_{ij} > s_{ij} \\ d_{ij} = 0 & \text{if } x_{ij} \leq s_{ij} \end{array} \right. \quad \begin{array}{l} S/N < 0.5: \text{bad variable} \\ 0.5 < S/N < 1: \text{weak variable} \\ S/N > 1: \text{strong variable} \end{array}$$

E' anche importante valutare il rapporto fra la quantità di informazione superiore ed inferiore all'MDL:

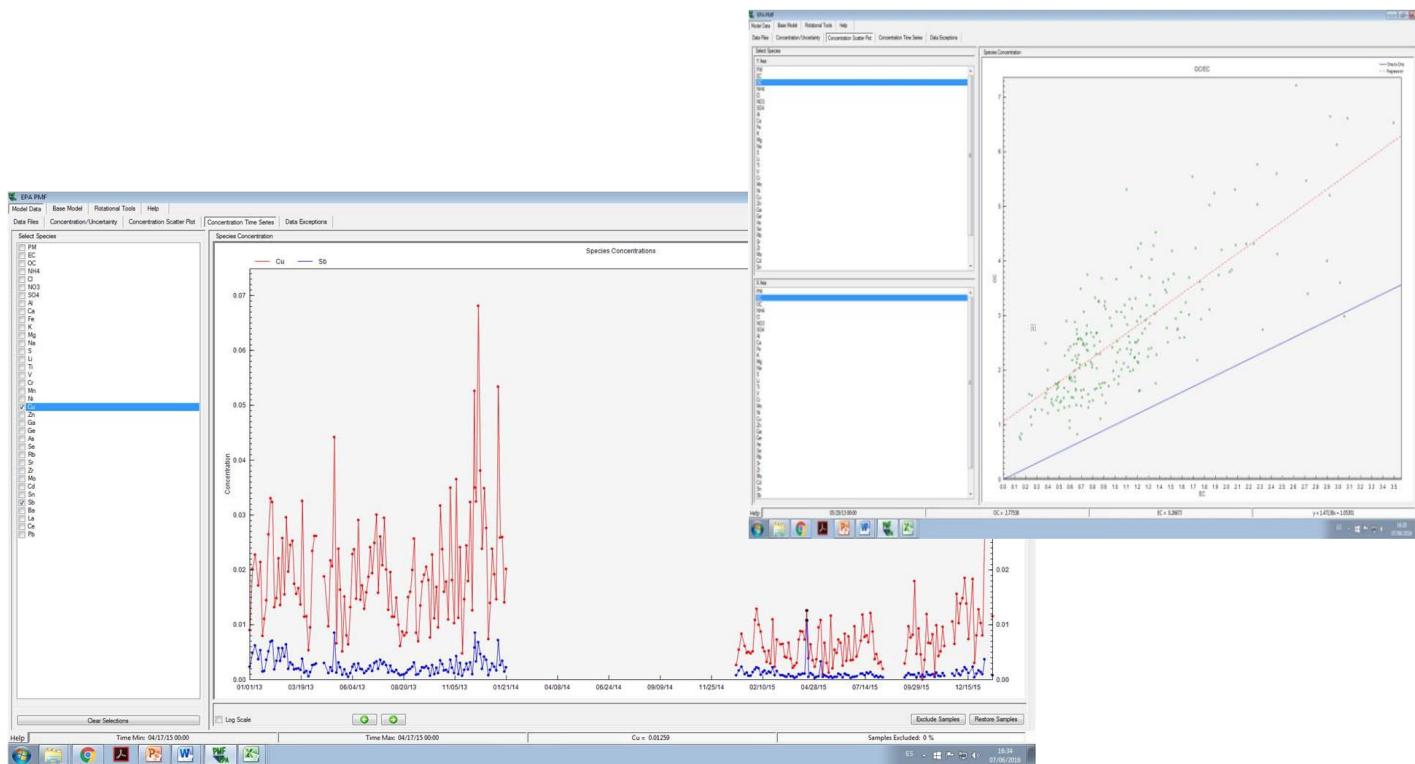
$$\text{Rapporto S/N} = \frac{\text{(somma di tutte le concentrazioni} > MDL)}{\langle MDL \rangle \times n^\circ \text{ di casi} < MDL}$$

Input data: elementi di diagnostica

Model Data



Il programma ha diversi elementi di diagnostica molto utili per una revisione ulteriore dei dati (scatterplot, andamenti temporali), con la possibilità di escludere singoli campioni.



Molto utile controllare anche scatterplot errore-concentrazione per trovare errori nella preparazione dei dati (la struttura del plot dipende infatti dal metodo di trattamento delle incertezze e dei casi mancanti o < MDL)

Outliers

Nei dati ambientali le distribuzioni sono generalmente non gaussiane e asimmetriche, con code verso i valori più alti.

Valori molto elevati possono essere causati da artefatti analitici, ma possono anche rappresentare qualcosa di reale (da valutare).

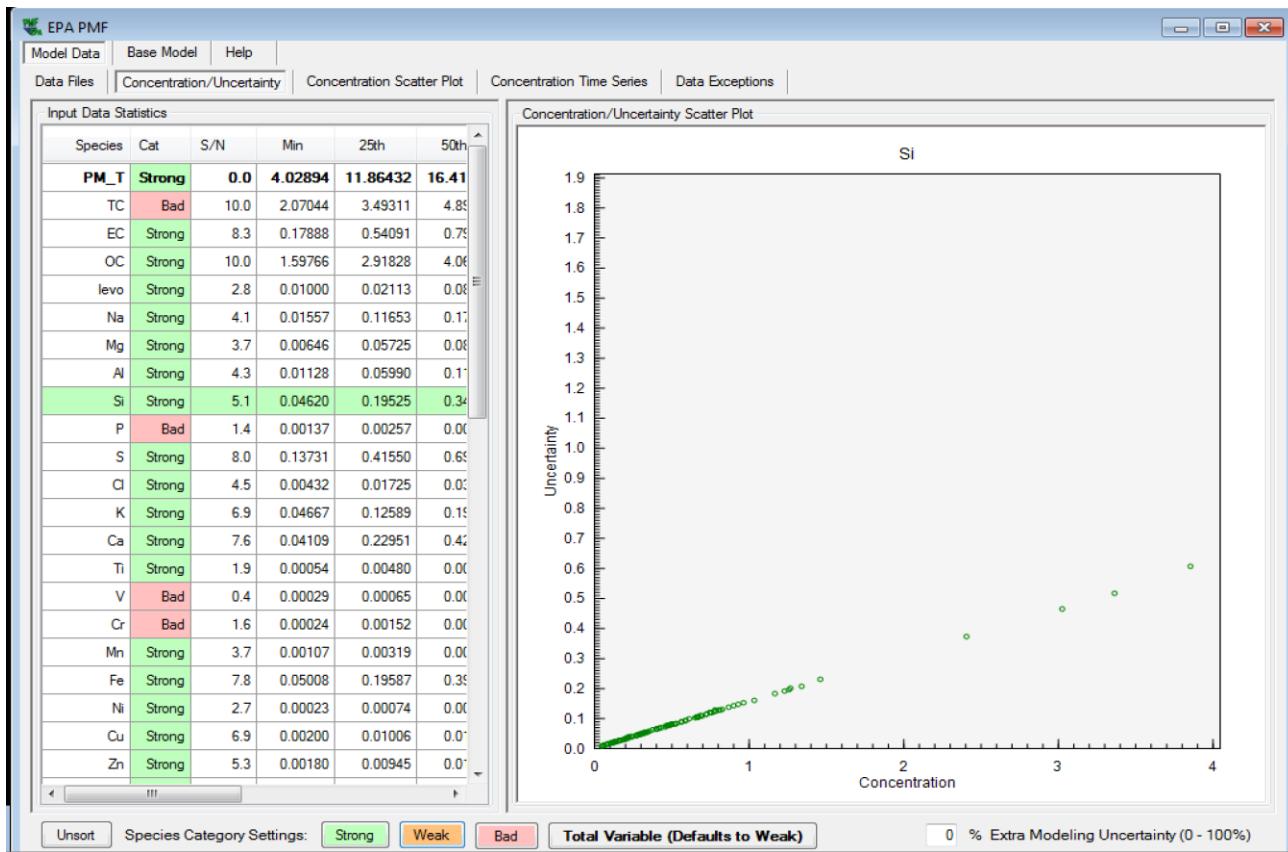
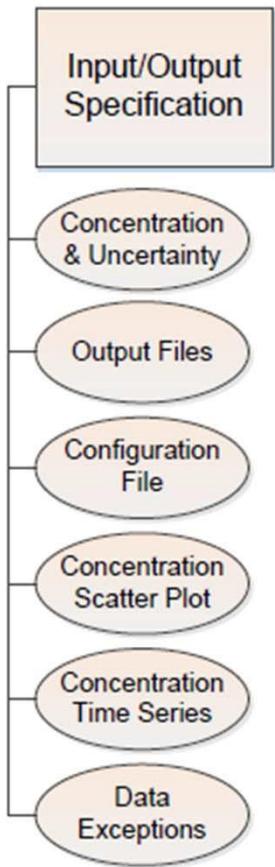
Questi valori disturbano l'analisi fattoriale tradizionale, ma possono essere gestiti bene nella PMF, utilizzando delle modalità "robuste", come ad esempio:

$$Q = \sum_{i=1}^N \sum_{j=1}^M \frac{e_{ij}}{h_{ij} s_{ij}} - \bar{Q}^2 \quad \text{con} \quad h_{ij} = \begin{cases} 1 & \text{se } |e_{ij}/s_{ij}| \leq a \\ \frac{|e_{ij}/s_{ij}|}{a} & \text{altrimenti} \end{cases}$$

La distanza "a" è solitamente impostata uguale a 4.0

Total variable

Model Data



La massa totale del PM può essere inserita come Total variable per avere matrici F e G assolute (attenzione: viene messa automaticamente come week, se avete già assegnato incertezza elevata ve rimessa strong).

I dati in uscita vengono presentati sia in unità assolute (le stesse della matrice di input) che con normalizzazione delle sorgenti a media 1.

Normalizzazione: G ed F assolute

$$x_{ij} \approx \sum_k g_{ik} \cdot f_{kj} = \sum_k (g_{ik} c_k) \cdot \left(\frac{f_{kj}}{c_k} \right) \equiv \sum_k \tilde{g}_{ik} \cdot \tilde{f}_{ki}$$

Output del modello

G ed F sono determinate a meno di un fattore di scala

Scalate per riprodurre pesi e profili fisici:
possibile solo conoscendo la massa del PM campionario
(come per APCFA)

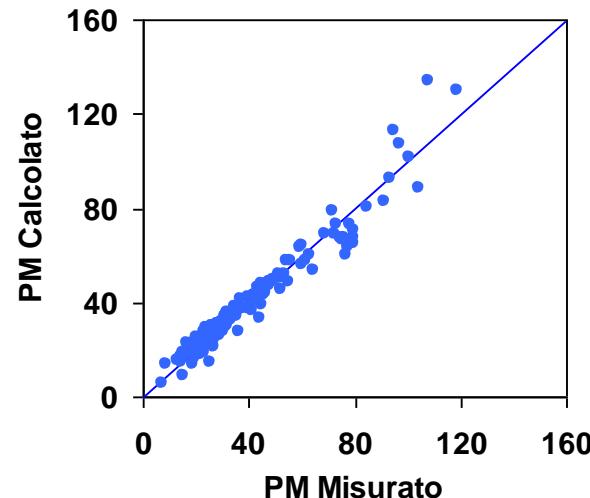
Metodo della massa interna
(conc. del PM introdotta come variabile, con peso molto basso per non guidare il fit)

$$\tilde{f}_{k,PM} = 1 \Rightarrow c_k = f_{k,PM}$$

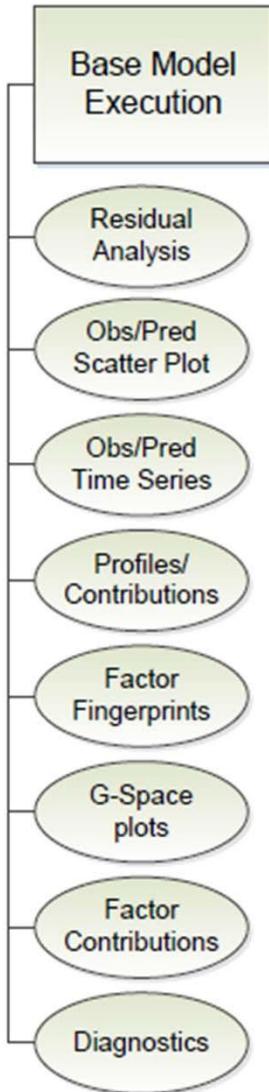
Metodo della regressione lineare

$$x_{i,PM} = \sum_{k=1}^P g_{ik} c_k + \varepsilon_i$$

Regressione lineare multipla della concentrazione del PM sui fattori (g_k): i coefficienti della regressione forniscono la normalizzazione



Esecuzione del modello BASE



Scelta del numero di fattori e del numero di inizializzazioni random (numero dei run)

EPA PMF

Model Data | Base Model | Rotational Tools | Help

Base Model Runs | Base Model Results

Base Model Runs

Number of Runs: 20 Number of Factors: 9
 Random Start Seed Number: 25

Error Estimation

Base Model Displacement Method
Selected Base Run: 15

Base Model Bootstrap Method
Selected Base Run: 15
Block Size: 3
Number of Bootstraps: 100
Min. Correlation R-Value: 0.6

Base Model BS-DISP Method

Displacement	Species	Cat.	S/N
<input type="checkbox"/>	PM_T	Strong	0.0
<input type="checkbox"/>	TC	Bad	10.0
<input type="checkbox"/>	EC	Strong	8.3
<input type="checkbox"/>	OC	Strong	10.0

Base Model Run Summary

Run Number	Q (Robust)	Q (True)	Converged
1	2877.3	2985.3	Yes
2	2877.3	2985.3	Yes
3	2877.3	2985.3	Yes
4	2678.8	2796.7	Yes
5	2678.8	2796.7	Yes
6	2877.3	2985.3	Yes
7	2877.3	2985.3	Yes
8	2877.4	2985.3	Yes
9	2877.4	2985.3	Yes
10	2678.8	2796.7	Yes
11	2678.8	2796.7	Yes
12	2877.3	2985.3	Yes
13	2877.4	2985.3	Yes
14	2678.8	2796.7	Yes
15	2678.8	2796.7	Yes
16	2678.8	2796.7	Yes
17	2678.8	2796.7	Yes
18	2877.4	2985.3	Yes
19	2877.3	2985.3	Yes
20	2877.2	2985.4	Yes

Problema delle soluzioni multiple

- a differenza della PCA, nella PMF la funzione Q può avere diversi minimi relativi (a causa dei pesi individuali e dei vincoli di non negatività)
 - A seconda del punto di partenza (F_0, G_0) l' algoritmo può convergere nell'uno o nell'altro, e non necessariamente nel minimo assoluto
- E' necessario effettuare più prove con diverse inizializzazioni pseudo-random:
- ✓ se con un numero sufficiente di prove (almeno 10, consigliabile 20) l' algoritmo converge sempre allo stesso Q , è probabile che esso sia il minimo assoluto
 - ✓ se si ottengono diversi valori di Q , occorre valutare quale sia quello a cui corrisponde la soluzione migliore
 - ✓ con un numero di fattori minore o maggiore rispetto a quello ottimale è più probabile ottenere soluzioni multiple

Esecuzione del modello BASE

Base Model Execution

Residual Analysis

Obs/Pred Scatter Plot

Obs/Pred Time Series

Profiles/Contributions

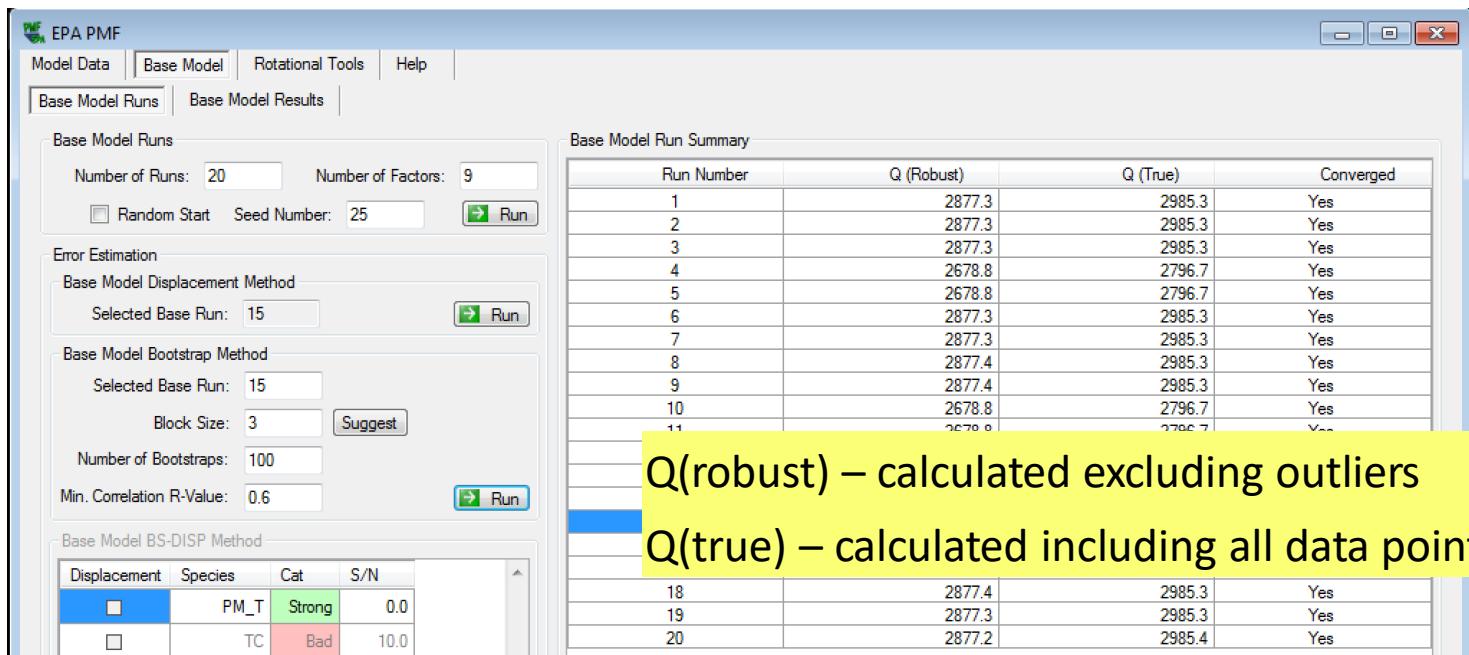
Factor Fingerprints

G-Space plots

Factor Contributions

Diagnostics

Il numero di fattori non è definito a priori: è importante eseguire il modello su un ampio intervallo di numero di fattori e poi vagliare tutti i risultati ottenuti



Q(robust) – calculated excluding outliers

Q(true) – calculated including all data points

$Q(\text{true}) \approx Q(\text{robust}) \approx Q(\text{theoretical/expected}) \rightarrow \text{good fit}$

$$Q_{\text{expected}} \approx (N - P) \times (M - P)$$

N: n° of elements/compounds
M: n° of samples
P: n° of factors

Q value e scelta del numero di fattori

Se la stima di s_{ij} è corretta, Q dovrebbe essere governata approssimativamente da una distribuzione chi-quadro (χ^2):

Il valore di aspettazione di Q è dato dal numero di gradi di libertà

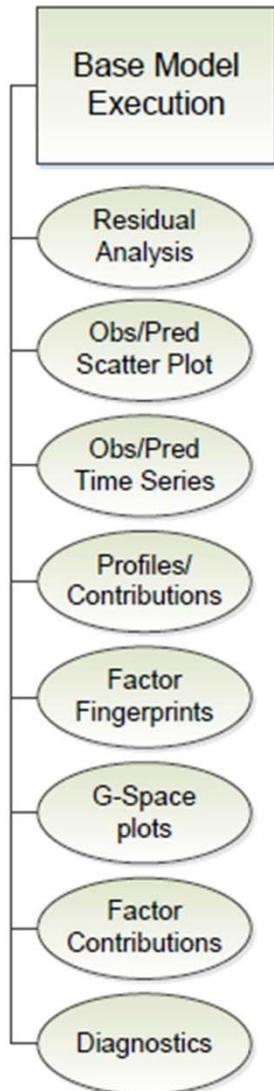
$$\langle Q \rangle = n^\circ \text{ of d.o.f.} = n^\circ \text{ of data} - n^\circ \text{ of constrains} =$$

$$= NxM - [NxP + MxP] + PxP = (N-P)(M-P)$$

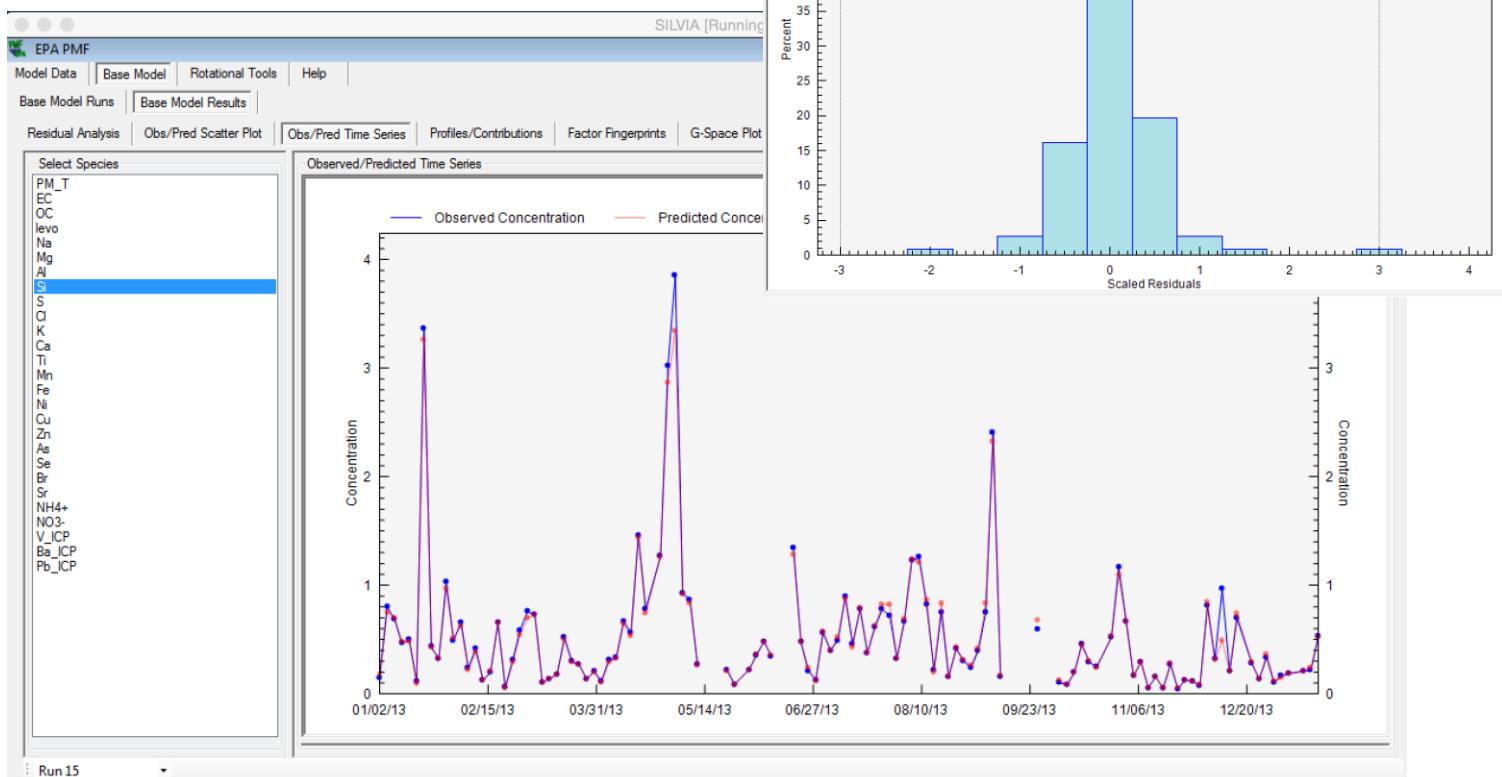
n° di dati (X) n° di vincoli (G, F) n° di gradi di libertà rotazionali

Per la scelta del numero di fattori, è utile vedere come cambia Q all'aumentare di P: dopo che un numero adeguato di sorgenti è stato incluso, fattori aggiuntivi non producono un significativo miglioramento di Q (attenzione: perché questo criterio sia valido è necessario che le incertezze siano state ben valutate!). E' anche importante guardare le distribuzioni dei residui delle singole specie chimiche e gli andamenti misurato vs ricostruito...

Risultati e diagnostica



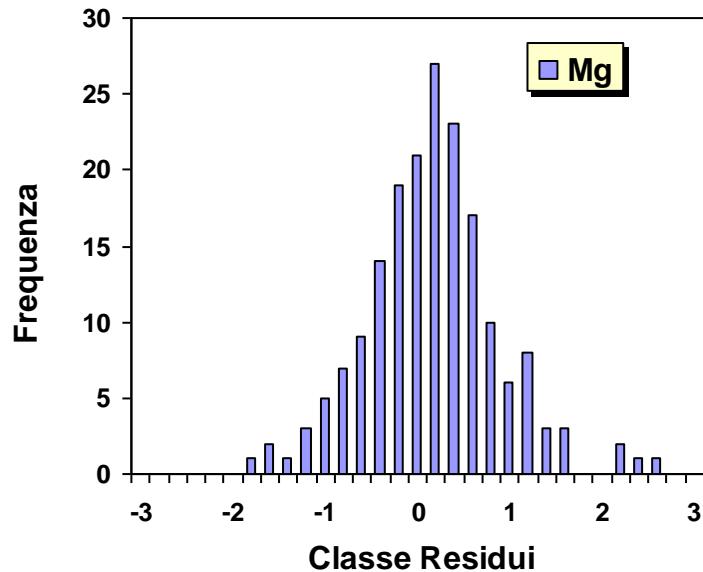
Controllare distribuzione residui
e andamenti misurato-ricostruito



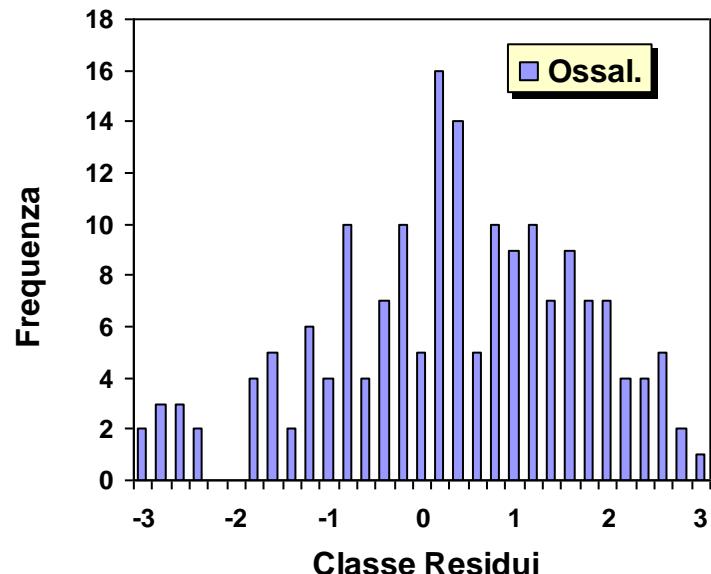
Valutare se aggiungere altre sorgenti e rivalutare dati di input
(scelta delle variabili, incertezze, casi da togliere)

Distribuzioni dei residui

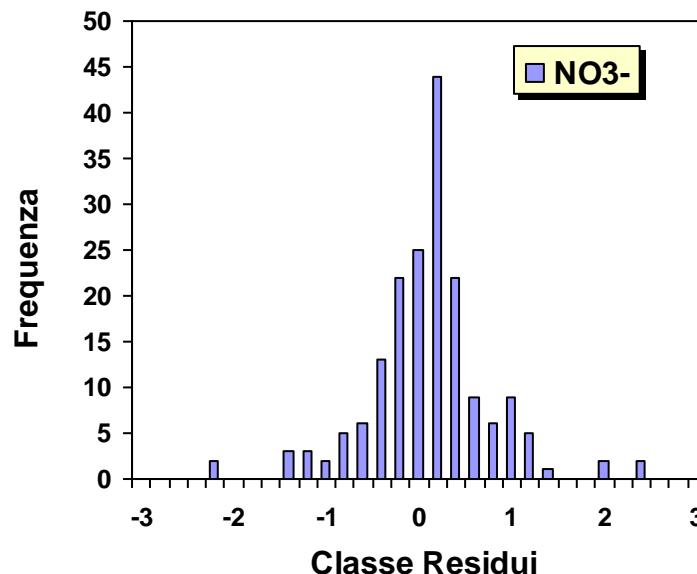
Se le incertezze sono state valutate correttamente e il modello è riuscito a riprodurre bene le concentrazioni di tutte le specie chimiche, le distribuzioni dei residui e_{ij}/s_{ij} dovrebbero essere gaussiane con valori compresi fra $\sim \pm 3$



Se vengono più larghe: incertezza sottostimata (se simmetriche) oppure è necessario aumentare il n° di fattori (se larghe e asimmetriche)



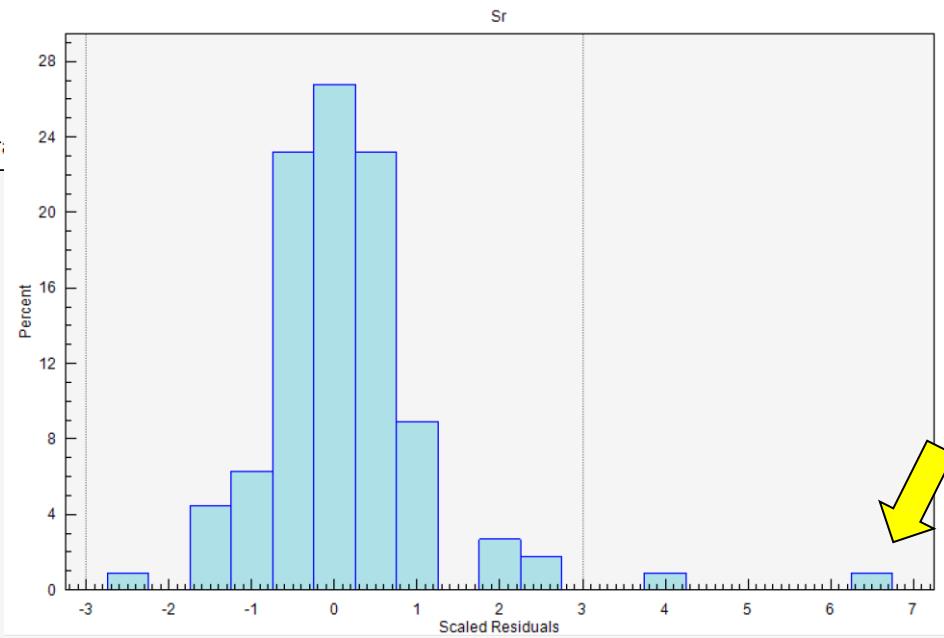
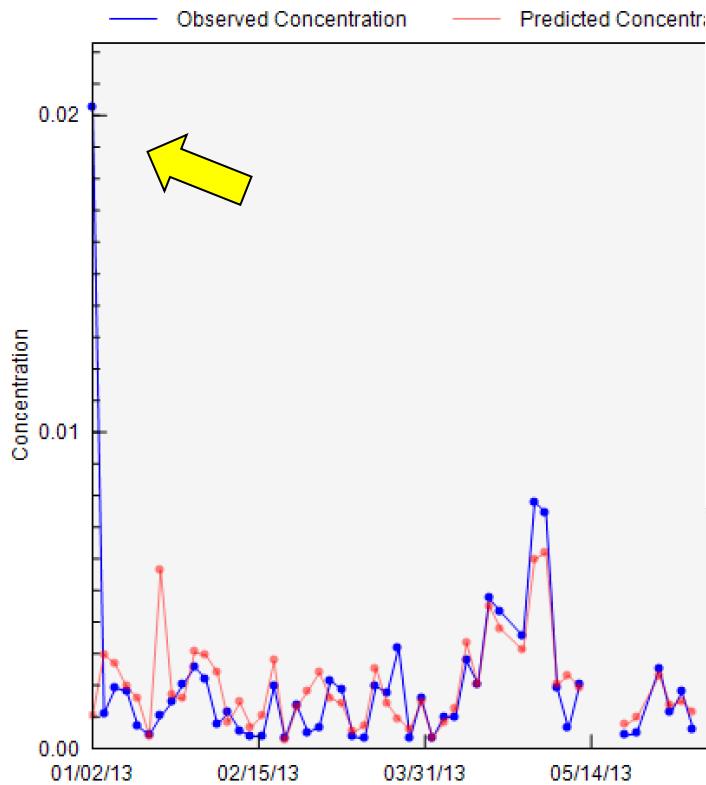
Se vengono molto strette: incertezza sopravvalutata, fattori unici reali o fattori unici fintizi (incertezza sottostimata!)



Andamenti misurato vs ricostruito

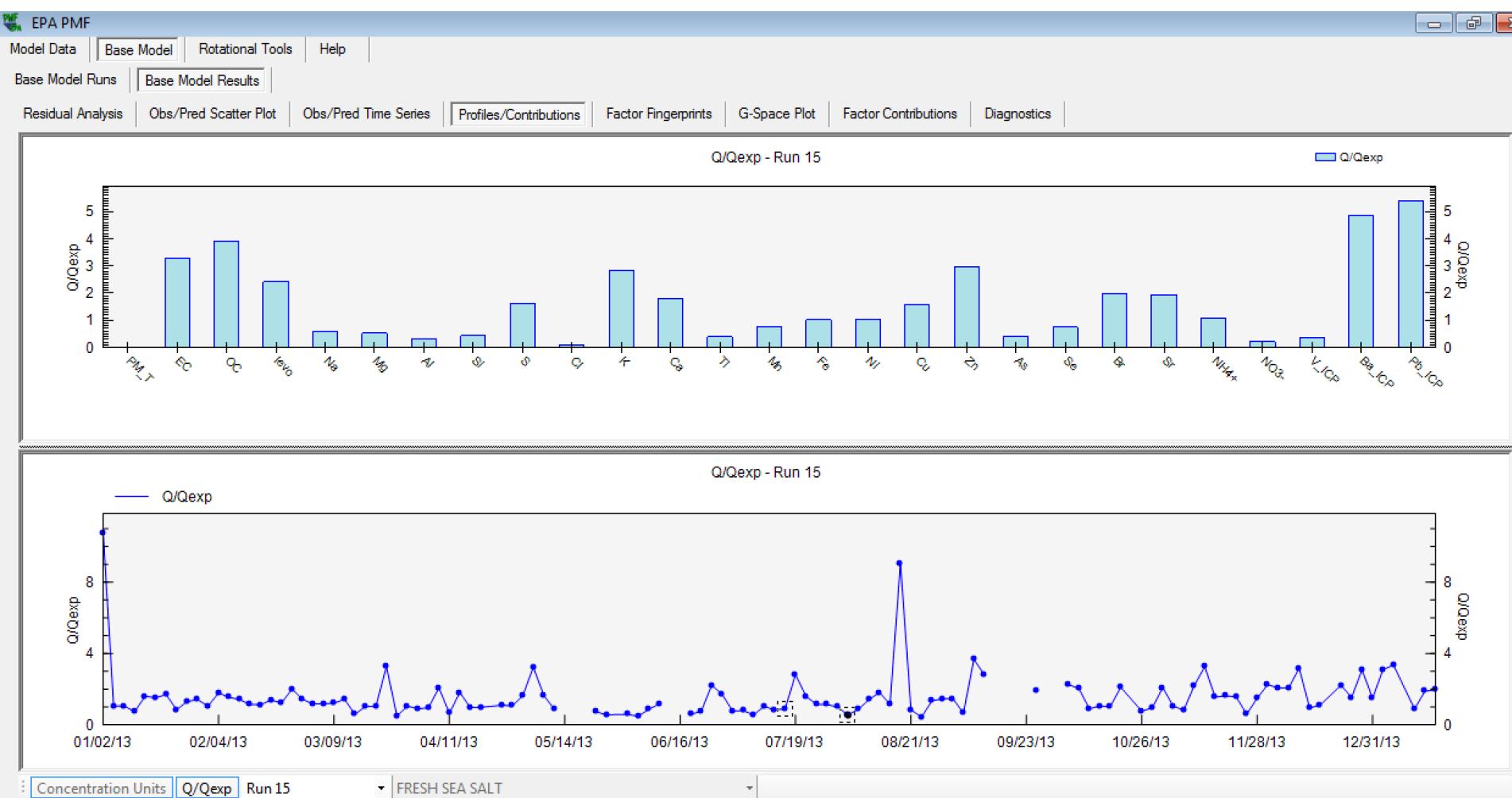
Molto importanti per capire quali sono le possibili cause di alti residui e decidere se aggiungere o meno altre sorgenti (o rimuovere il dato)

Es: picchi fuochi d'artificio (Sr)



Q/Qexpected

Possibilità di confrontare il Q ottenuto con quello atteso nelle diverse sorgenti, per ciascuna variabile e ciascun campione



Profili (F) e contributi (G)

Base Model Execution

Residual Analysis

Obs/Pred Scatter Plot

Obs/Pred Time Series

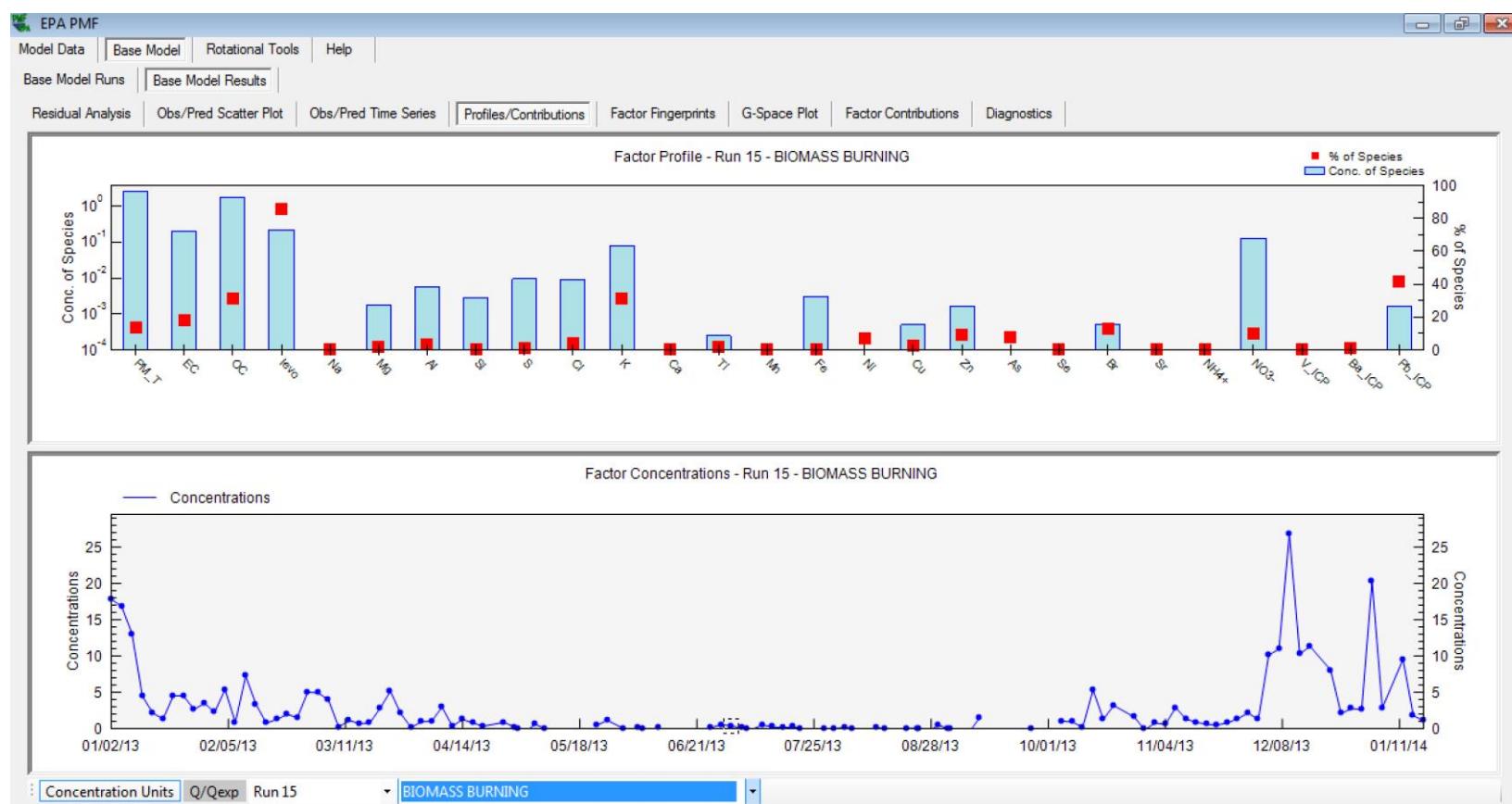
Profiles/ Contributions

Factor Fingerprints

G-Space plots

Factor Contributions

Diagnostics



STEP MOLTO IMPORTANTE:

Assegnazione dei fattori alle sorgenti e valutazione critica della consistenza e ragionevolezza fisica dei profili e dei contributi

Factor fingerprints

Base Model Execution

Residual Analysis

Obs/Pred Scatter Plot

Obs/Pred Time Series

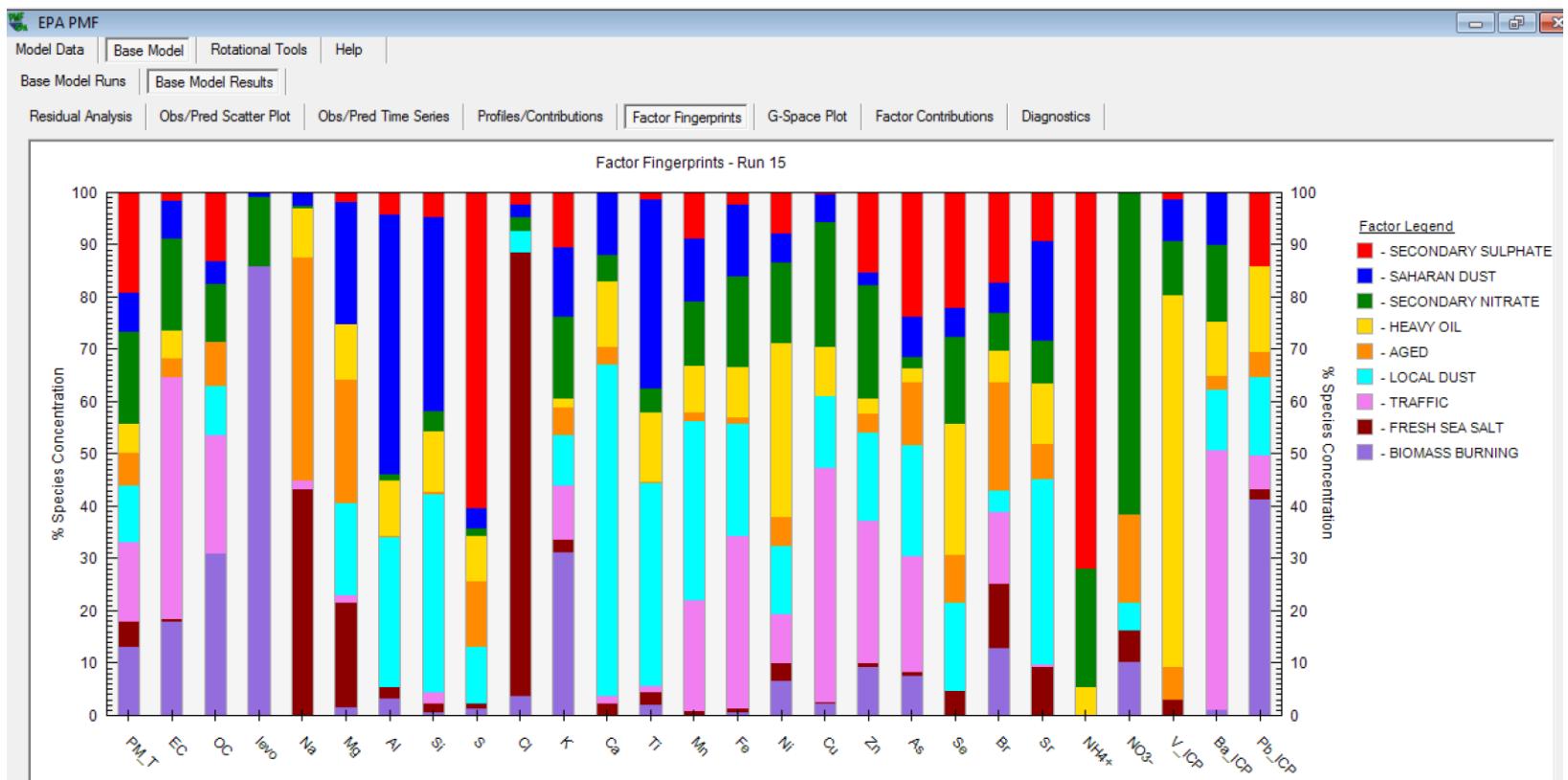
Profiles/ Contributions

Factor Fingerprints

G-Space plots

Factor Contributions

Diagnostics



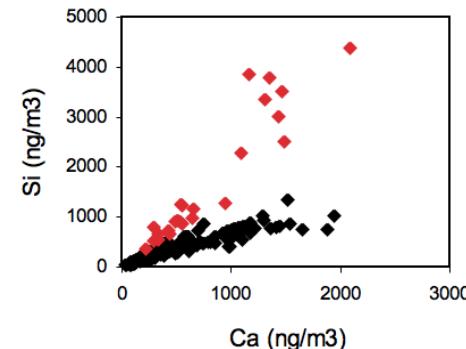
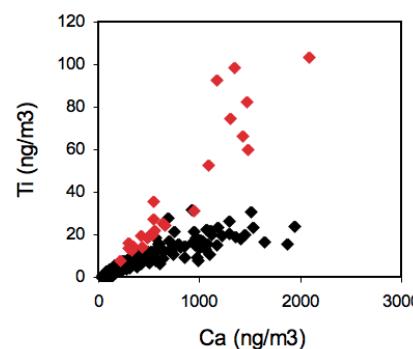
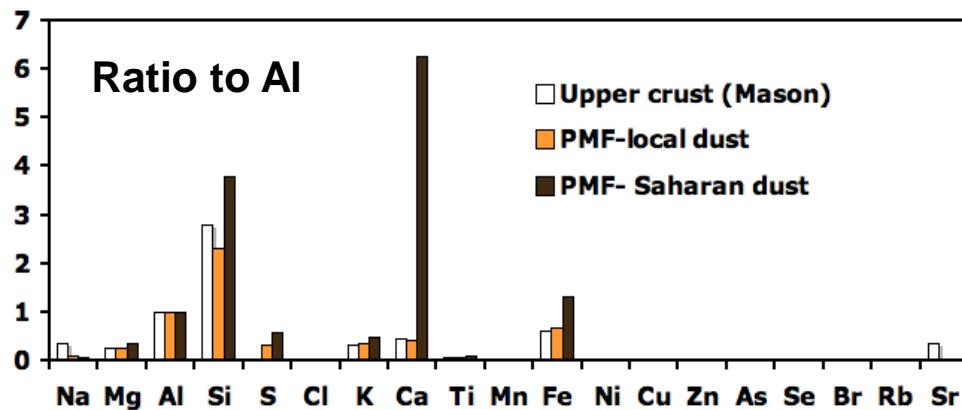
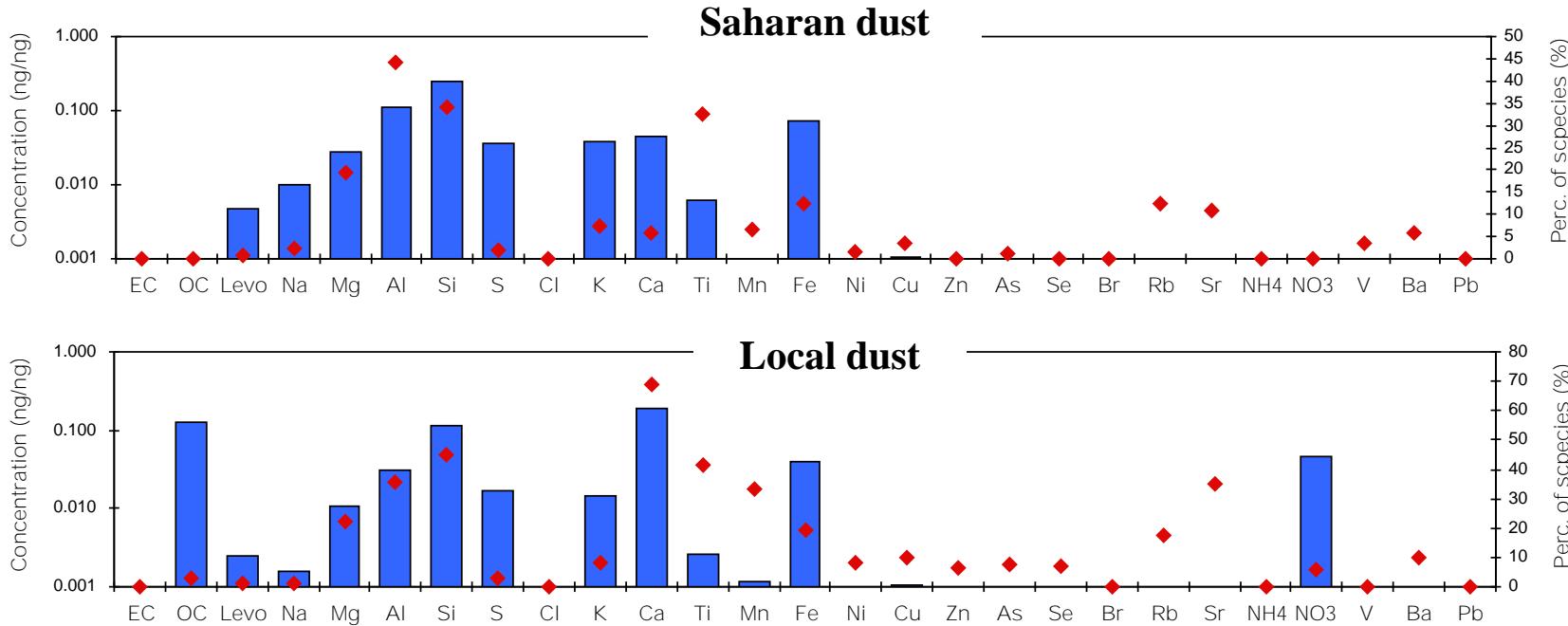
Contributi % medi delle sorgenti alla concentrazione delle diverse specie (e del PM)

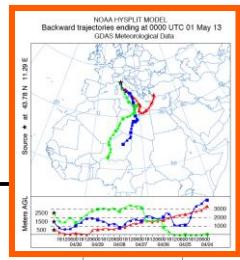
Attenzione: non includono i residui, sono chiusi a 100 sulla massa ricostruita (che può essere anche molto minore della misurata: controllare!)

Assegnazione dei fattori a specifiche sorgenti

- ✓ Considerare tutte le conoscenze a priori che si possono avere sull'area di studio (possibili sorgenti, inventario delle emissioni, flussi di traffico, metereologia, precedenti SA, etc.).
- ✓ Confrontare i profili ottenuti con i dati di letteratura e i database internazionali:
EPA: <https://www3.epa.gov/ttnchie1/software/speciate/>
EU: <http://source-apportionment.jrc.ec.europa.eu/Specieurope/index.aspx>
- ✓ Consentendo però una certa flessibilità dato che i processi di trasporto atmosferico possono alterare i profili dal punto di emissione al sito recettore e possono anche produrre sorgenti collineari che al recettore sono viste come un unico fattore con profilo misto.
- ✓ Anche l'andamento temporale è importante per determinare o sostenere l'assegnazione di un fattore ad una sorgente. Contano sia l'andamento delle emissioni (cicli produttivi, picchi del traffico), che le periodicità di tipo stagionale/settimanale/etc. e i fenomeni di trasporto.

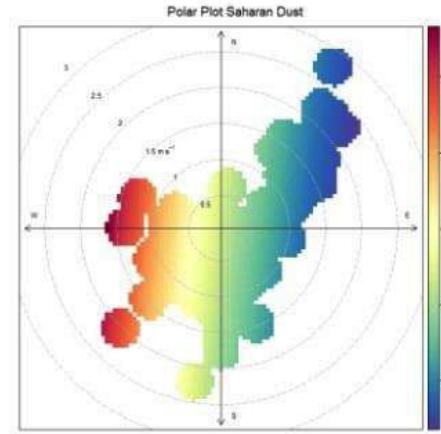
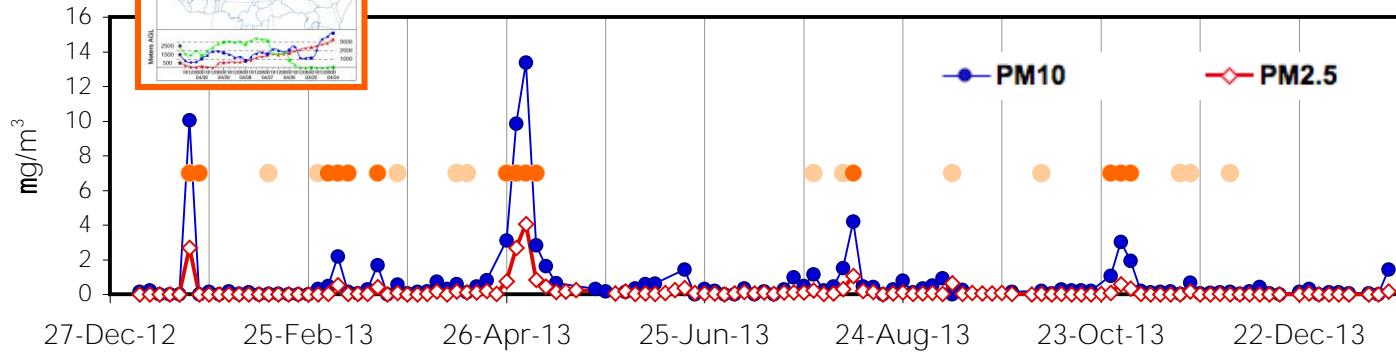
Separazione Dust locale e sahariano (Firenze, AIRUSE)





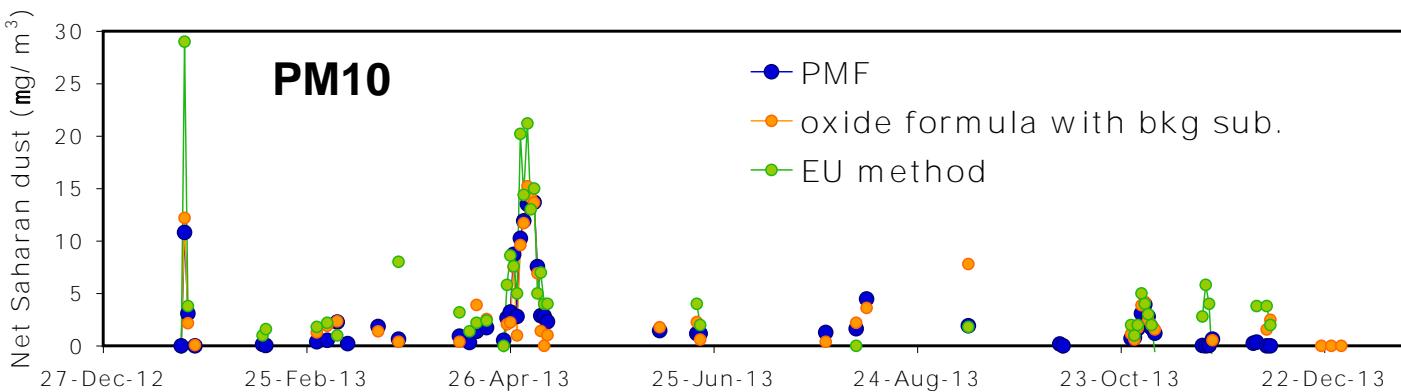
DUST (Florence): Saharan contribution

PMF - Saharan dust source



PM10

- PMF
- oxide formula with bkg sub.
- EU method



In agreement with the oxide formula approach [Nava et al., 2014]

EU method tends to overestimate the Saharan contribution

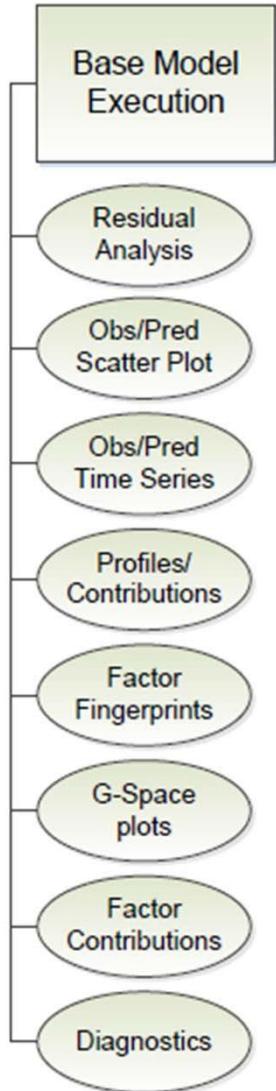
Controllo della ragionevolezza della soluzione

Ci sono vincoli fisici che devono essere sempre considerati per validare la soluzione trovata:

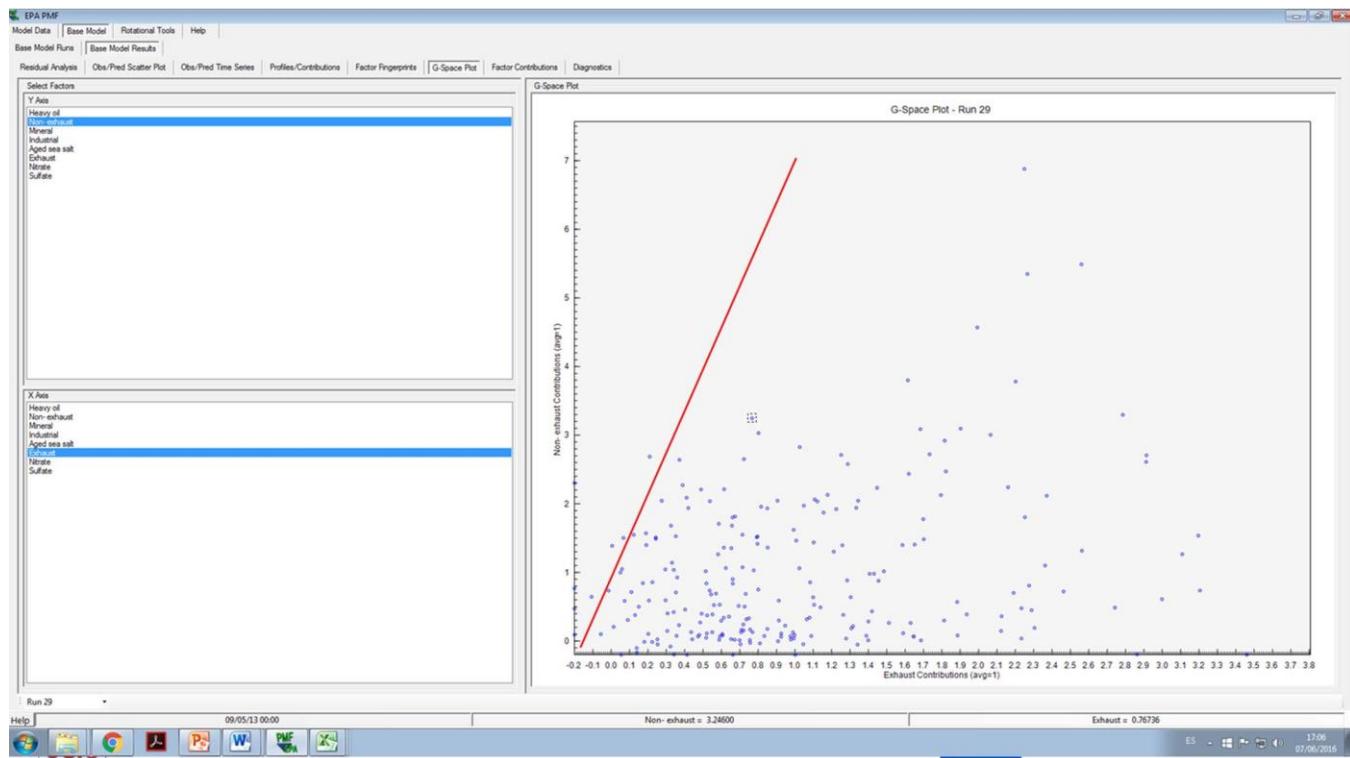
- la somma dei contributi di tutte le sorgenti ad un specie chimica deve essere minore della sua concentrazione (in particolare la somma dei contributi di tutte le sorgenti deve essere minore della concentrazione del PM), per tutti i singoli campioni;
- per definizione gli elementi di F devono essere < 1 ; anche la somma degli elementi di F di una sorgente deve essere < 1 (ma “non troppo” se tutte le principali specie sono state misurate).
- gli andamenti temporali delle sorgenti e i loro contributi devono essere “ragionevoli” (assenza/presenza di una sorgente in una determinata stagione o in determinati periodi/orari);
- i profili delle sorgenti devono essere ragionevoli rispetto a quanto noto sulla composizione delle emissioni (assenza/presenza di specie chimiche in taluni profili, rapporti elementali coerenti con quanto noto, ...);

**Questi criteri devono essere SEMPRE verificati,
nonché utilizzati per scegliere il numero di fattori ottimale
(e ridurre l’ambiguità rotazionale)**

G space plots



Scatterplots fra coppie di sorgenti



Valutare se parziali correlazioni hanno senso fisico.
Vedere come cambiano con il numero di fattori, con le rotazioni o imponendo vincoli.

Contributi delle sorgenti

Contributi % medi e andamenti temporali delle sorgenti

Base Model Execution

Residual Analysis

Obs/Pred Scatter Plot

Obs/Pred Time Series

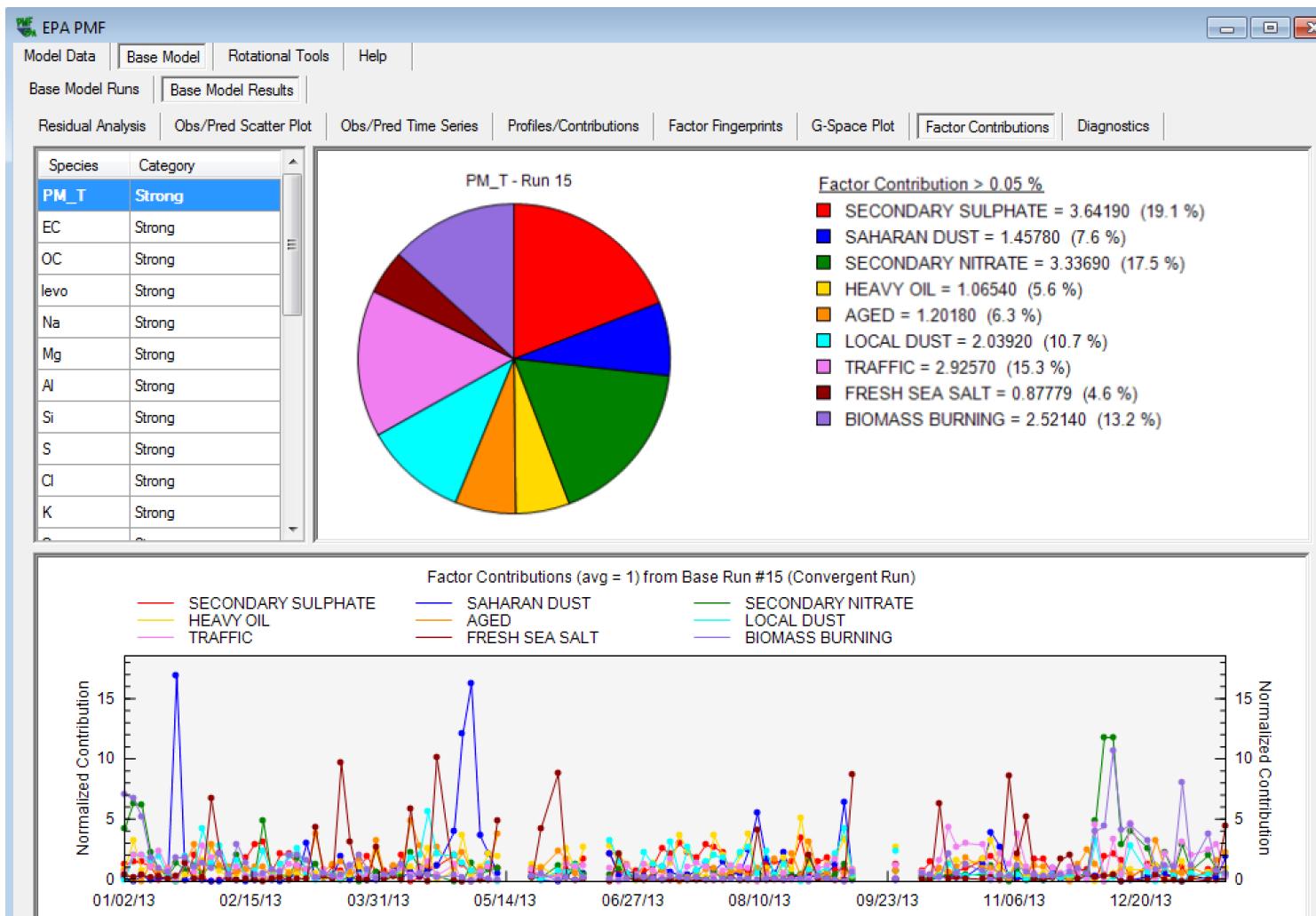
Profiles/ Contributions

Factor Fingerprints

G-Space plots

Factor Contributions

Diagnostics



Ambiguità rotazionale

Un problema di fattorizzazione ammette sempre più soluzioni, poiché ogni matrice T invertibile definisce una rotazione che non altera la matrice dei residui:

$$X = G F + E = G T T^{-1} F + E = (G T)(T^{-1} F) + E$$

Tutte le soluzioni ruotate fittano ugualmente bene i dati sperimentali!

Tutte le rotazioni possono essere rappresentate da una sequenza di rotazioni elementari come queste:

$$T = \begin{pmatrix} \cancel{1} & 0 & 0 & 0 \\ \cancel{0} & 1 & 0 & a \\ \cancel{0} & 0 & 1 & b \\ \cancel{0} & 0 & 0 & 1 \end{pmatrix}$$

Somma di una colonna di G ,
moltiplicata per “a”, ad
un’altra colonna di G

$$T^{-1} = \begin{pmatrix} \cancel{1} & 0 & 0 & 0 \\ \cancel{0} & 1 & 0 & -a \\ \cancel{0} & 0 & 1 & b \\ \cancel{0} & 0 & 0 & 1 \end{pmatrix}$$

Sottrazione di una riga di F ,
moltiplicata per “a”, da
un’altra riga di F

Scelta della rotazione ottimale

Nella PMF la condizione di non-negatività di G ed F riduce il numero di soluzioni possibili ma non elimina del tutto l'ambiguità rotazionale.

Fissato il numero di fattori, è quindi necessario selezionare la soluzione (G , F) ottimale in base a criteri di consistenza e ragionevolezza fisica:

- la somma dei contributi di tutte le sorgenti ad un specie chimica deve essere minore della sua concentrazione (in particolare la somma dei contributi di tutte le sorgenti deve essere minore della concentrazione del PM), per tutti i singoli campioni;
- gli andamenti temporali delle sorgenti e i loro contributi devono essere “ragionevoli” (assenza/presenza di una sorgente in una determinata stagione, ...);
- i profili delle sorgenti devono essere ragionevoli rispetto a quanto noto sulla composizione delle emissioni (assenza/presenza di specie chimiche in taluni profili, rapporti elementali coerenti con quanto noto, ...);
- per definizione gli elementi di F devono essere < 1 ; anche la somma degli elementi di F di una sorgente deve essere < 1 (ma “non troppo” se tutte le principali specie sono state misurate).

Questo può essere fatto tramite una valutazione a posteriori esplorando le possibili soluzioni (vedi parametro *FPEAK*) e/o imponendo dei vincoli a priori (vedi matrici F_{key} e G_{key} in PMF2 o vincoli in PMF EPA 5)

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

F and G zero values and Edges

Rotational ambiguity may be also reduced:

1) **locking to zero some elements of F and G** (use of F_{key} e G_{key} in PMF2 or constraints in PMF EPA5), on the basis of what is known on specific sources.

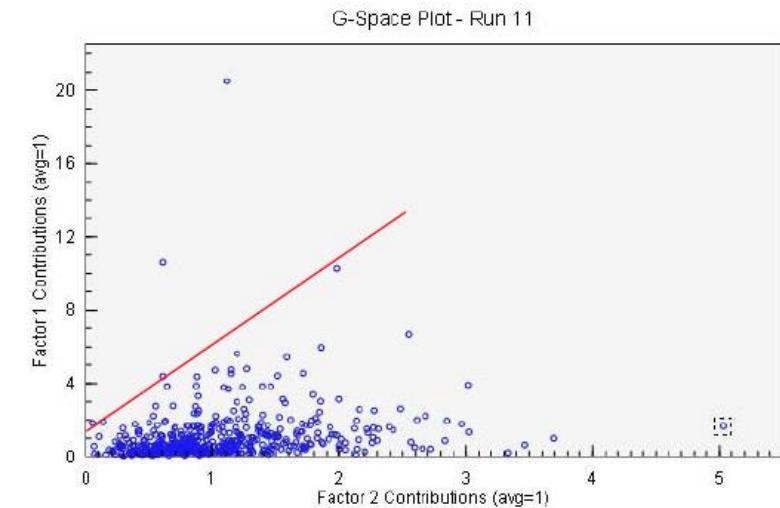
Actually, if all the columns of G and all the rows of F contain a significant number of zeros, then the result is unique. In other words, data points that have low or no impact from a source are very useful to the model. For this reason high time resolved data is very useful, as increasing the time resolution typically provides samples that have greater between-sample variability in the source contributions than samples integrated over longer time periods.

2) Looking at G-space plots.

if 2 factors are independent, points should completely fill the scatter plot:

with a correct rotation the limiting edges usually coincide with, or are parallel to, the axes.

However, a certain degree of correlation may be due to physical reasons (for ex. colocation od sources coupled with meteorology).

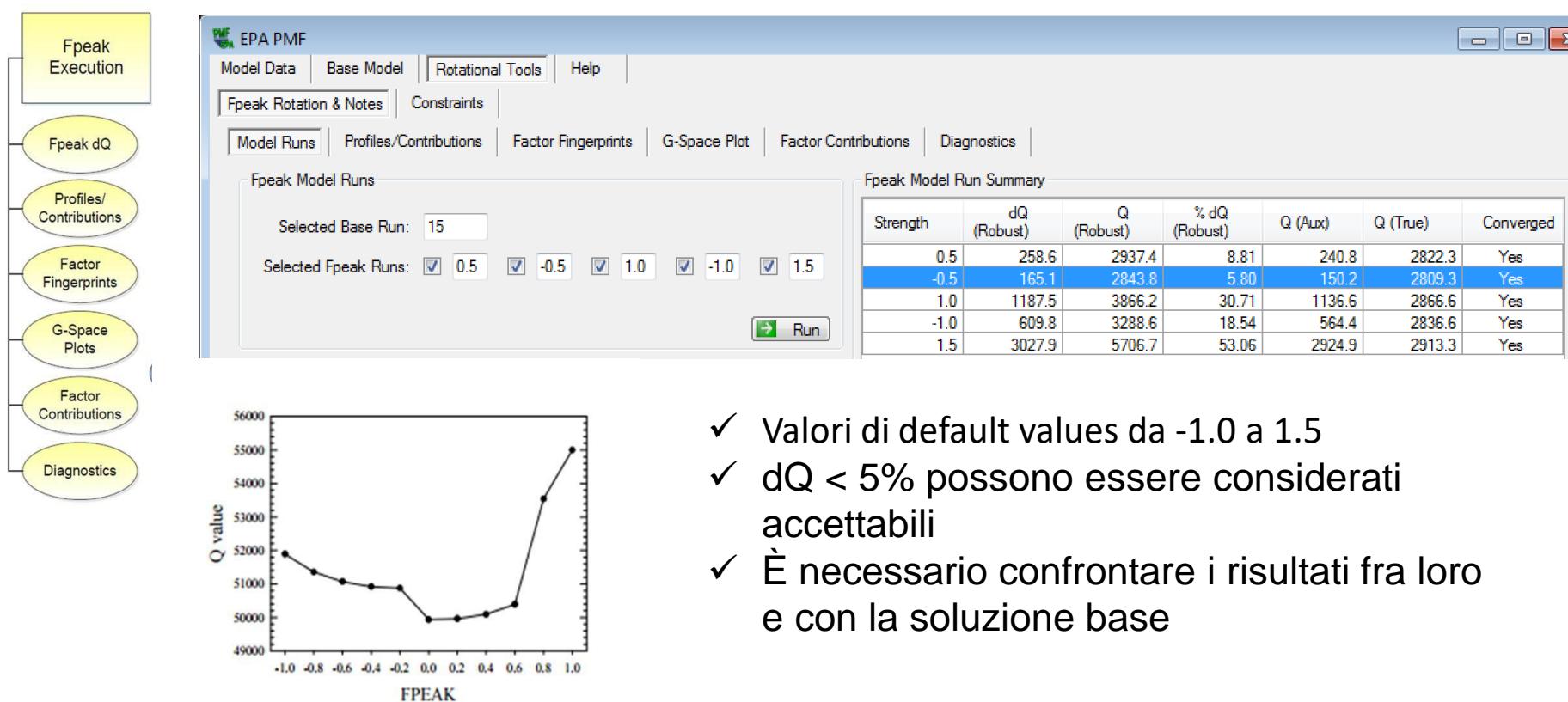


FPEAK

Il parametro FPEAK può essere utilizzato per esplorare le diverse soluzioni ruotate.

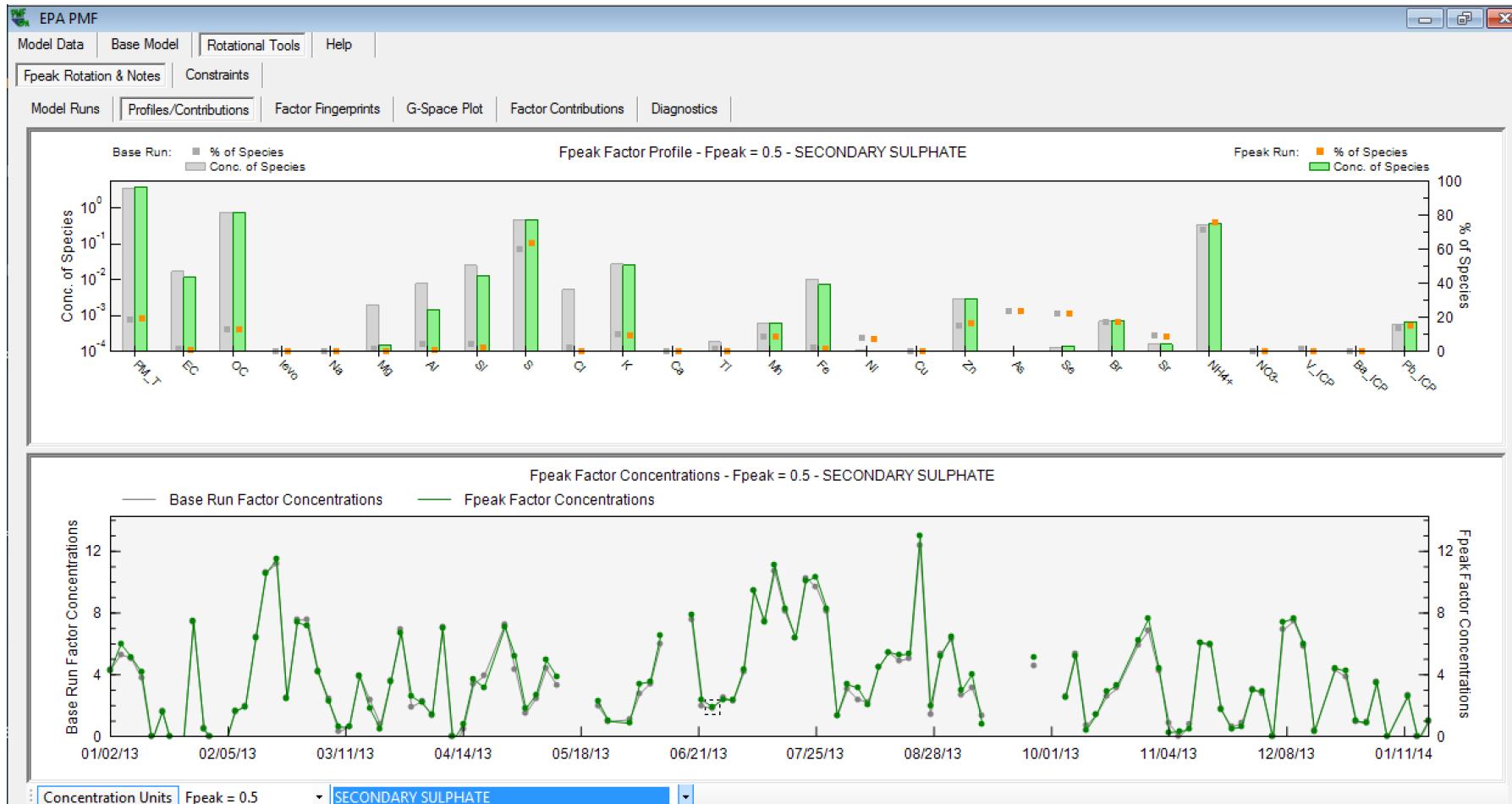
Agisce in modo simile ad una rotazione VARIMAX:

- Valori positivi amplificano le differenze all'interno dei profili e smussano gli andamenti
- Valori negativi riducono le differenze all'interno dei profili e producono andamenti temporali più variabili (raramente utili)



FPEAK

Confronto con la soluzione BASE



In assenza di miglioramenti significativi nell'interpretabilità dei profili/contributi o dei G space-plots, si consiglia di mantenere la soluzione a FPEAK nullo

Constraints

E' possibile inserire vincoli tramite espressioni (**Expression Builder**), ed anche a **partire dalle uscite del BASE RUN**

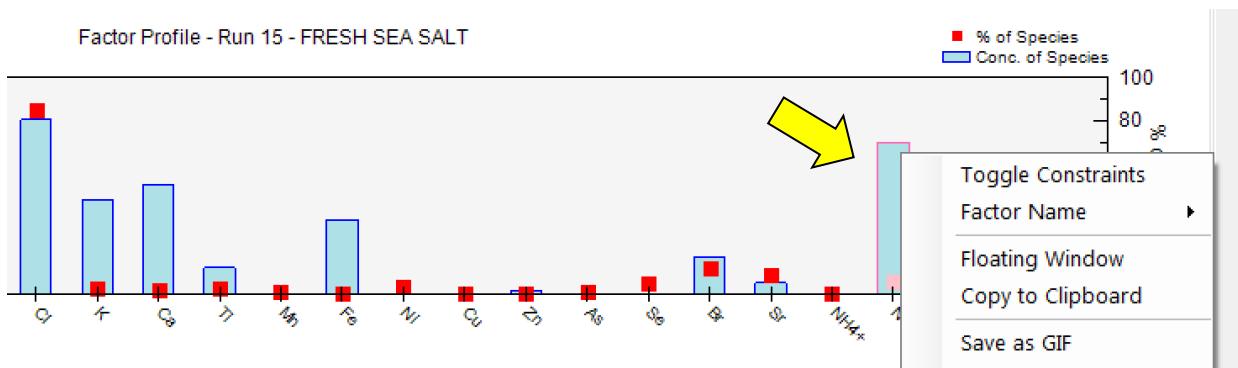
The screenshot shows the EPA PMF software interface. The top menu bar includes 'Model Data', 'Base Model', 'Rotational Tools', 'Help', 'Fpeak Rotation & Notes', and 'Constraints'. The 'Constraints' tab is selected. Below the menu is a toolbar with 'Model Runs' and 'Expressions'. The 'Expressions' section contains an 'Expression Builder' with three tabs: 'Ratio' (selected), 'Mass Balance', and 'Custom'. The 'Ratio' tab shows a 'Factor' list (SECONDARY SULPHATE, SAHARAN DUST, SECONDARY NITRATE, HEAVY OIL, AGED, LOCAL DUST, TRAFFIC, FRESH SEA SALT, BIOMASS BURNING) and two lists for 'Species (numerator)' and 'Species (denominator)' containing various chemical species like PM_T, EC, OC, levo, Na, Mg, Al, Si, S, Cl, K. A red box highlights this area. Below it is a table with columns 'Expression', 'dQ', and '% dQ'. Buttons for 'Add to Expressions', 'Remove Selected Expressions', and 'Remove All Expressions' are at the bottom. A purple box highlights this table. To the right, the 'Constrained Model Run' section includes a 'Run' button and a 'Selected Base Run: 15' dropdown. It has three tabs: 'dQ (Robust)', 'Q (Robust)', and '% dQ (Robust)'. Below this is the 'Error Estimation' section with 'Constrained Model Displacement Method' and 'Constrained Model Bootstrap Method' tabs. The 'Bootstrap' tab shows 'Number of Bootstraps: 6' and 'Minimum Correlation R-Value: 0.6'. A 'Run' button and 'Block Size: 3' are also present. At the bottom is the 'Constrained Model BS-DISP Method' section with a table:

Displacement	Species	Cat	S/N
□	PM_T	Strong	0.0
□	TC	Bad	10.0
□	EC	Strong	8.3
□	OC	Strong	10.0

N.B: anche il run di BASE della PMF include dei vincoli (non negatività)

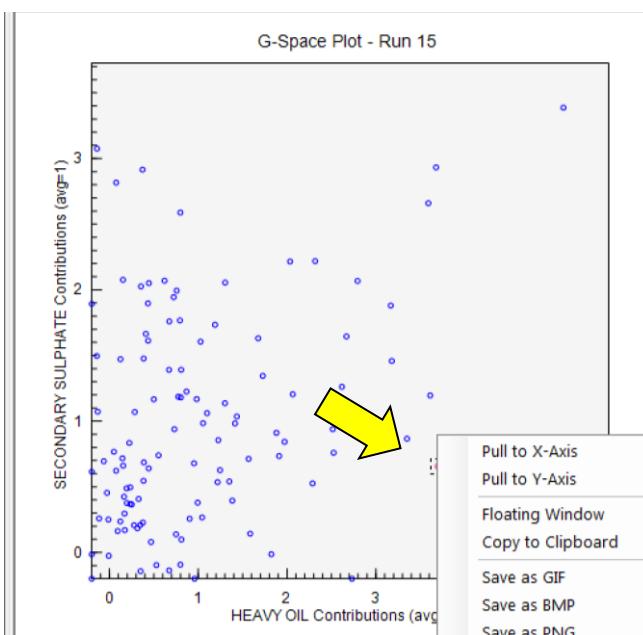
Constraints dal modello di base

Factor Profile - Run 15 - FRESH SEA SALT

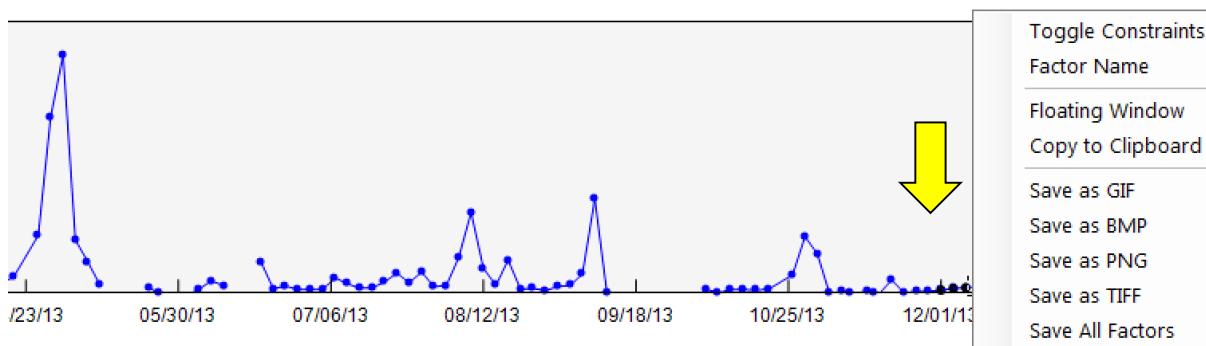


Si seleziona dai grafici le specie chimiche e/o i campioni su cui vogliamo applicare dei vincoli

- Pulling existing factor profile
- Pulling existing contributions
- Pulling G space plots



Factor Concentrations - Run 15 - SAHARAN DUST



Gli elementi selezionati vengono riportati in una tabella dove è possibile scegliere il tipo di vincolo.

FRESH SEA SALT	NO3-	Pull Down Maximally	NA
SAHARAN DUST	12/1/2013 12:00:00 AM	Pull Down Maximally	NA
SAHARAN DUST	12/4/2013 12:00:00 AM	Pull Down Maximally	NA
SAHARAN DUST	12/7/2013 12:00:00 AM	Pull Down Maximally	NA

Constraints dal modello di base

Strength of pulling:

- Soft: permitting a maximum dQ (default is 0.5%)
- Hard: regardless of Q increase: usare con estrema cautela e solamente in presenza di forti motivazioni

Soft options:

- Pull down maximally
- Pull up maximally
- Pull to value

Actor	Element	Type	Value	dQ	% dQ
A SALT	NO3-	Pull Down Maximally	NA	13.39	0.50
JRNING	levo	Pull Up Maximally	NA	13.39	0.50
DUST	AI	Pull to Value	0.08	13.39	0.50

Hard options:

- Forced to zero
- Forced to original value
- Forced to upper/lower limit

Actor	Element	Type	Value	dQ	% dQ
A SALT	NO3-	Set to Zero	0	NA	NA
JRNING	levo	Set to Original Value	0	NA	NA
DUST	AI	Define Limits	0.05/1	NA	NA

Controllare sempre l'aumento Q e gli effetti sui parametri forzati ma anche su tutto il resto (un vincolo su un elemento/sorgente può alterare anche altri profili e contributi e portare anche a scambi di sorgente)

Constraints: expression builder (soft)

1. Rapporti fra elementi all'interno di un fattore
2. Mass balance: equazioni fra elementi di F, anche di fattori diversi

Expressions

Expression Builder

Ratio Mass Balance Custom

Ratio

Factor: SECONDARY SULPHATE
SAHARAN DUST
SECONDARY NITRATE
HEAVY OIL
AGED
LOCAL DUST
TRAFFIC
FRESH SEA SALT
BIOMASS BURNING

Species (numerator): PM_T
EC
OC
levo
Na
Mg
Al
Si
S
Cl
K

Species (denominator): PM_T
EC
OC
levo
Na
Mg
Al
Si
S
Cl
K

Value: 0.12

Expression	dQ	% dQ
[FRESH SEA SALT Mg] - 0.12 * [FRESH SEA SALT Na] = 0	13.39	0.50

Es 1:

$Mg/Na = 0.12 \text{ in } FRESH \text{ SEA } SALT$

Expressions

Expression Builder

Ratio Mass Balance Custom

Mass Balance

=

Coefficient: 10 Factor: smelter
soil
secondary
traffic
marine
heavy oil Species: PM2.5
Aluminum
Ammonium Ion
Arsenic
Barium
Bromine
Calcium

Expression	dQ	% dQ
[soil Aluminum] - 10 * [traffic Aluminum] = 0	69.411	0.50

Es 2:

$AI(SOIL) = 10 * AI(TRAFFIC)$

Constraints: expression builder (soft)

3. Custom: equazioni fra elementi di F e di G, anche di fattori diversi (es.: contributi PM1<PM2.5<PM10)

Expressions

Expression Builder

Ratio Mass Balance Custom

Custom

Profiles Contributions

Factor: SECONDARY SULPHA
SAHARAN DUST
SECONDARY NITRATI
HEAVY OIL
AGED

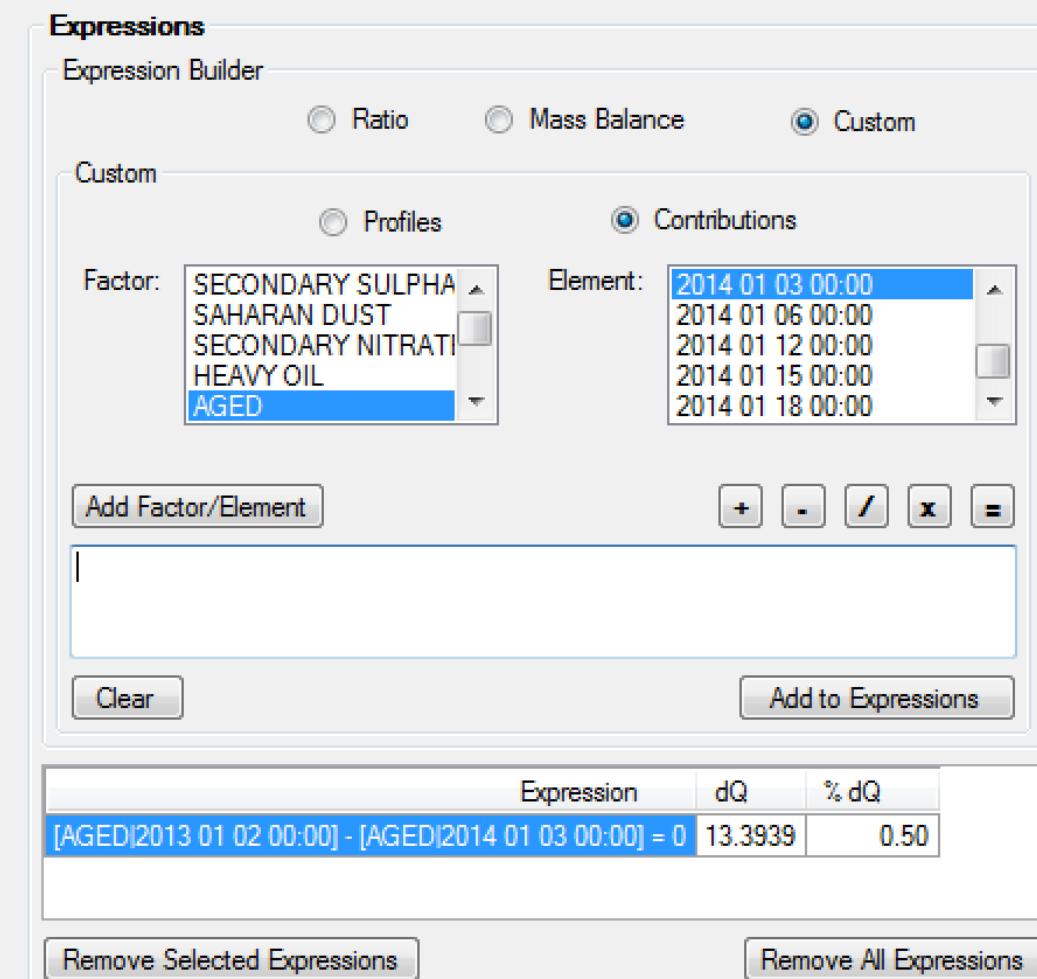
Element: 2014 01 03 00:00
2014 01 06 00:00
2014 01 12 00:00
2014 01 15 00:00
2014 01 18 00:00

Add Factor/Element + - / x =

Clear Add to Expressions

Expression	dQ	% dQ
[AGED 2013 01 02 00:00] - [AGED 2014 01 03 00:00] = 0	13.3939	0.50

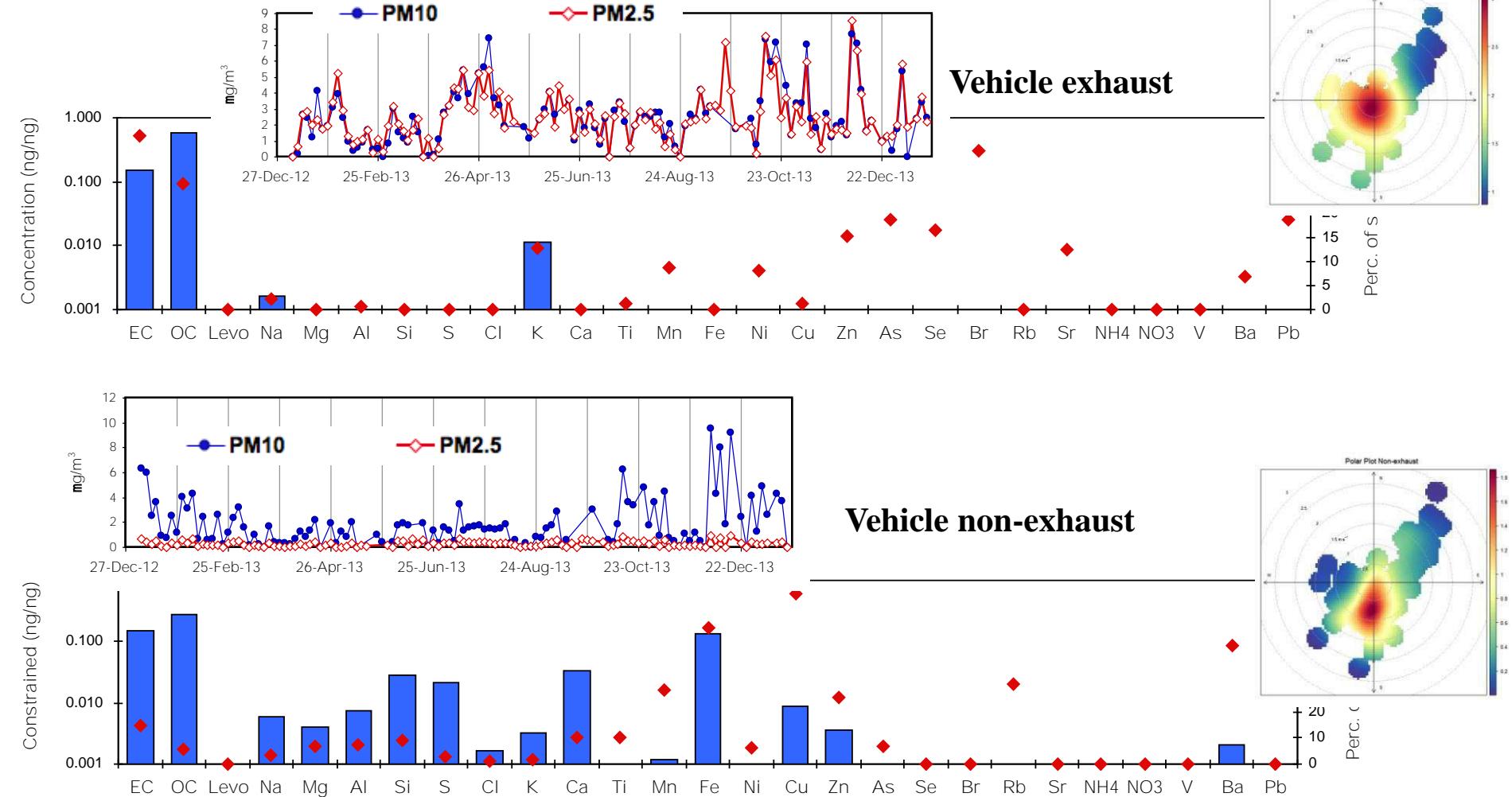
Remove Selected Expressions Remove All Expressions



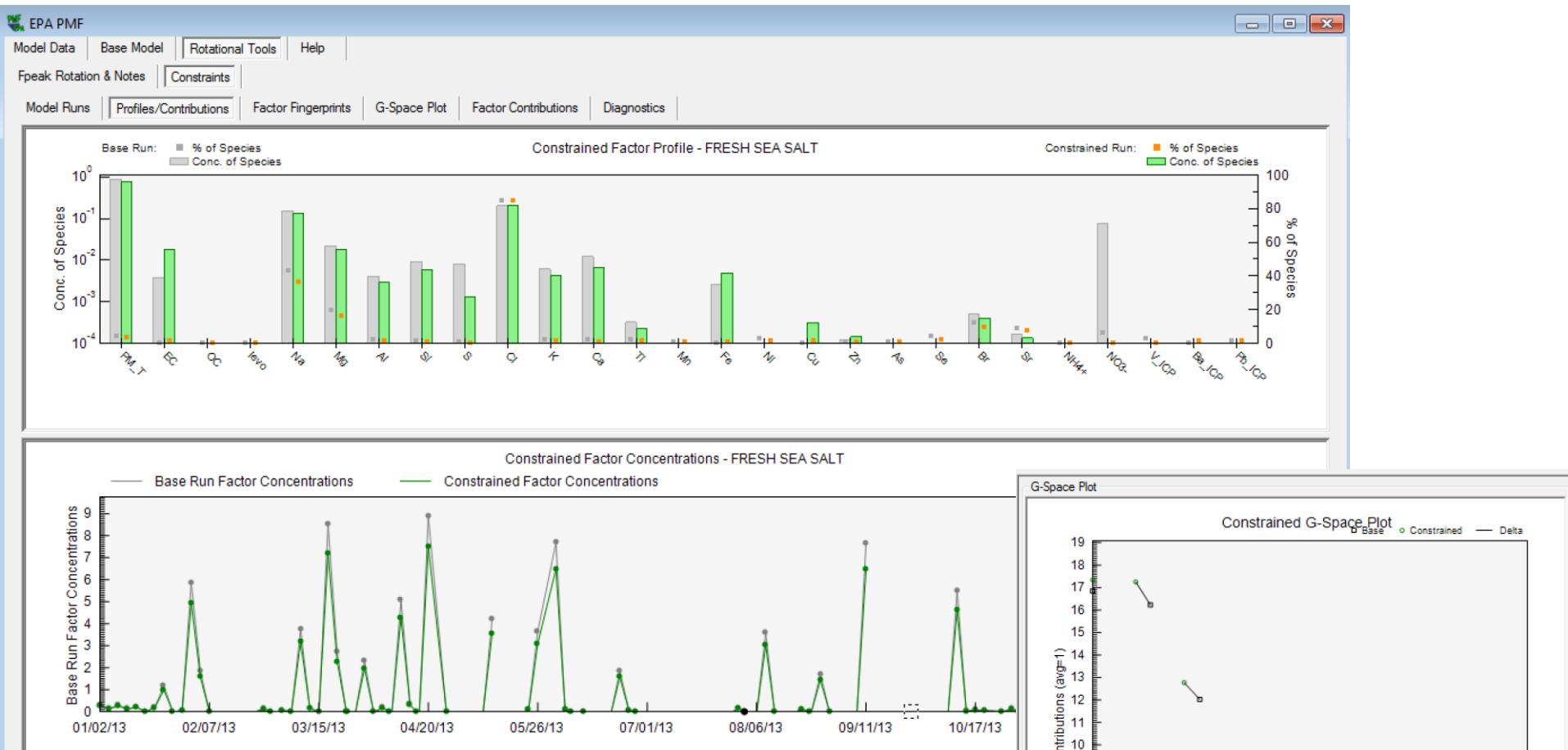
- PM10 and PM2.5 chemical speciation in 5 EU cities
- US-EPA PMF 5.0: best solution found pooling PM10 and PM2.5 daily samples in a single input matrix



AIRUSE
LIFE 11 ENV/ES/000584



Results of constrained PMF



Confronto con la soluzione BASE

Controllare dettagliatamente tutti i cambiamenti in F e G in tutte le sorgenti e le variazioni nei G-plots

Stima delle incertezze sui risultati

Le incertezze sui risultati derivano sia dai dati di input che dalle assunzioni del modello.

EPA PMF3 fornisce 3 procedure per la stima delle incertezze, basate sull'esecuzione di diversi run del modello e sull'analisi statistica delle soluzioni:

- **Bootstrapping (BS)**: metodo probabilistico (Montecarlo) che include la componente random delle incertezze, e, solo in parte, quella rotazionale.
Robusto rispetto a errate valutazioni delle incertezze di input.
- **Displacement (DISP)**: metodo basato sull'analisi di perturbazioni controllate degli elementi delle matrici F e G, per stimare l'incrtezza legata all'ambiguità rotazionale.
Sensibile a errate valutazioni delle incertezze di input.
- **Both (BS-DISP)**: include sia la componente random che quella rotazionale.
Meno sensibile a errate valutazioni delle incertezze di input rispetto al solo DISP.

Il solo utilizzo del solo BS può portare a:

- sottostimare l'errore, perché l'incertezza rotazionale può essere la componente dominante (soprattutto quando il numero di fattori è elevato);
- sovrastimare l'errore, se il data-set contiene pochi casi di contributi nulli (o quasi nulli), dato che questi ultimi limitano molto l'ambiguità rotazionale e tutte le volte che vengono eliminati dal resampling si ottengono risultati molto deviati.

Bootstrapping (BS)

L'analisi PMF viene fatta girare su un certo numero (scelto dall'operatore, consigliato almeno 100) di data-sets creati campionando in modo random blocchi di dati dalla matrice X originaria. Ad ogni run i fattori ottenuti sono confrontati con quelli del run di base e su di essi "mappati" in base al livello di correlazione (con una soglia specificata dall'operatore, tipicamente 0.6).

L'incertezza quindi valutata analizzando la statistica dei risultati ottenuti.

EPA PMF

Model Data | Base Model | Rotational Tools | Help |

Base Model Runs | Base Model Results | Base Model Bootstrap Results | Error Estimation Summary |

Base Model Runs

Number of Runs: 20 Number of Factors: 9

Random Start Seed Number: 25

Error Estimation

Base Model Displacement Method

Selected Base Run: 15

Base Model Bootstrap Method

Selected Base Run: 15
Block Size: 3
Number of Bootstraps: 100
Min. Correlation R-Value: 0.6

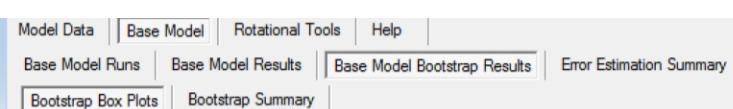
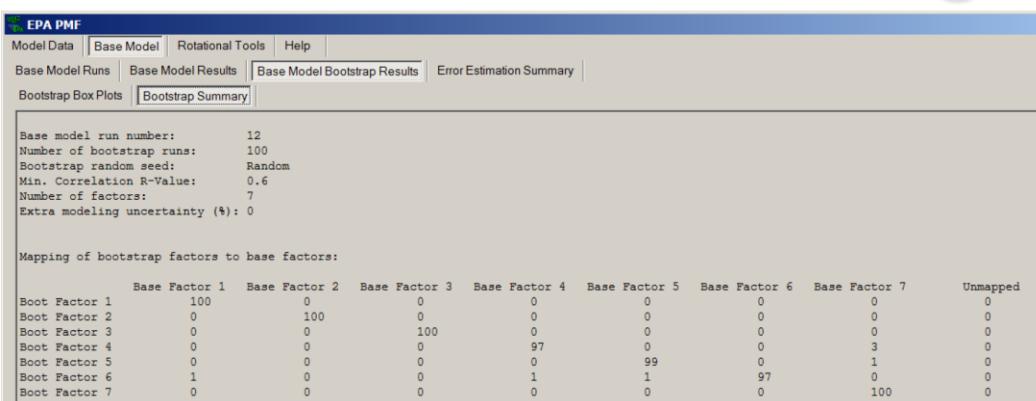
Base Model BS-DISP Method

Displacement	Species	Cat	S/N
<input type="checkbox"/>	PM_T	Strong	0.0
<input type="checkbox"/>	TC	Bad	10.0
<input type="checkbox"/>	SO	Good	0.0

Base Model Run Summary

Run Number	Q (Robust)
1	2877.3
2	2877.3
3	2877.3
4	2678.8
5	2678.8
6	2877.3
7	2877.3
8	2877.4
9	2877.4
10	2678.8
11	2678.8
12	2877.3
13	2877.4
14	2678.8
15	2678.8
16	2678.8
17	2678.8
18	2877.4
19	2877.3
20	2877.2

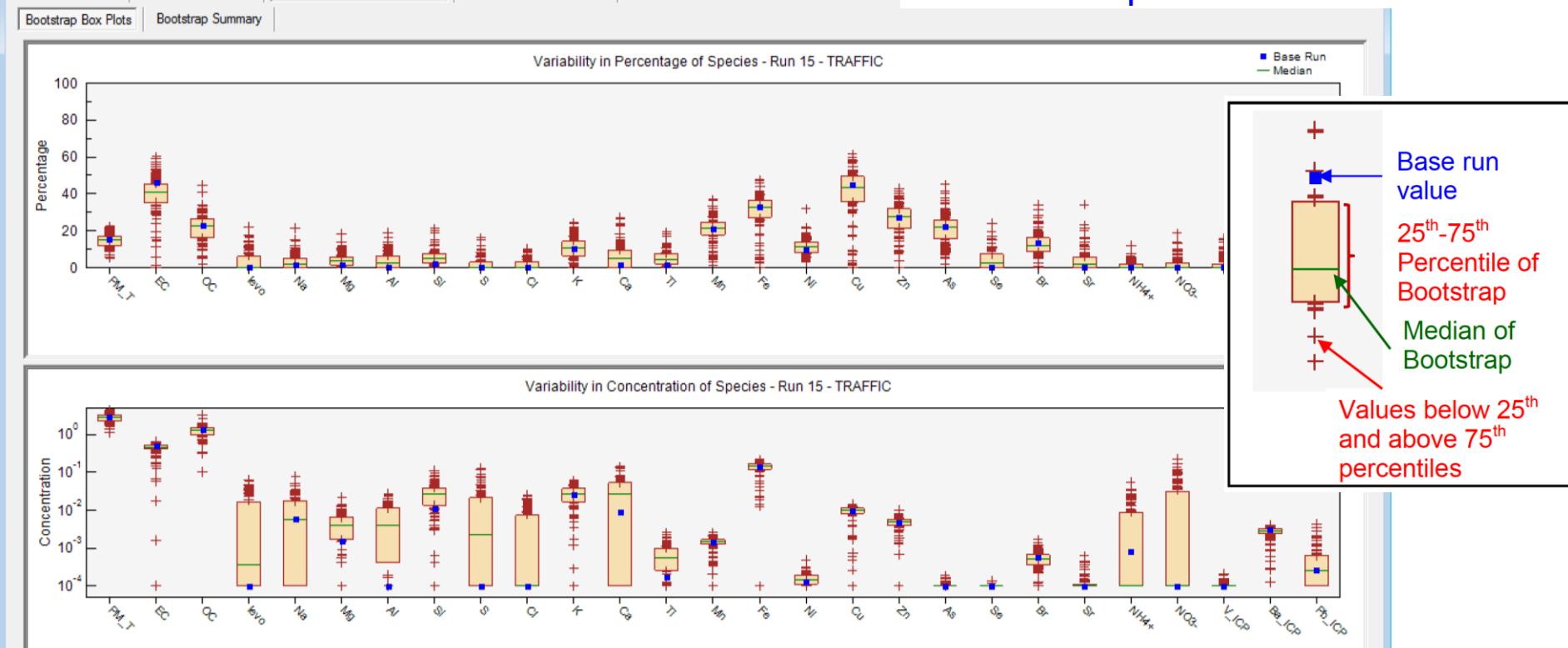
BS diagnostic



Controllare i fattori “unmapped”

Mapping >80%: i risultati sono stabili ed è legittimo usare le incertezze calcolate col BS

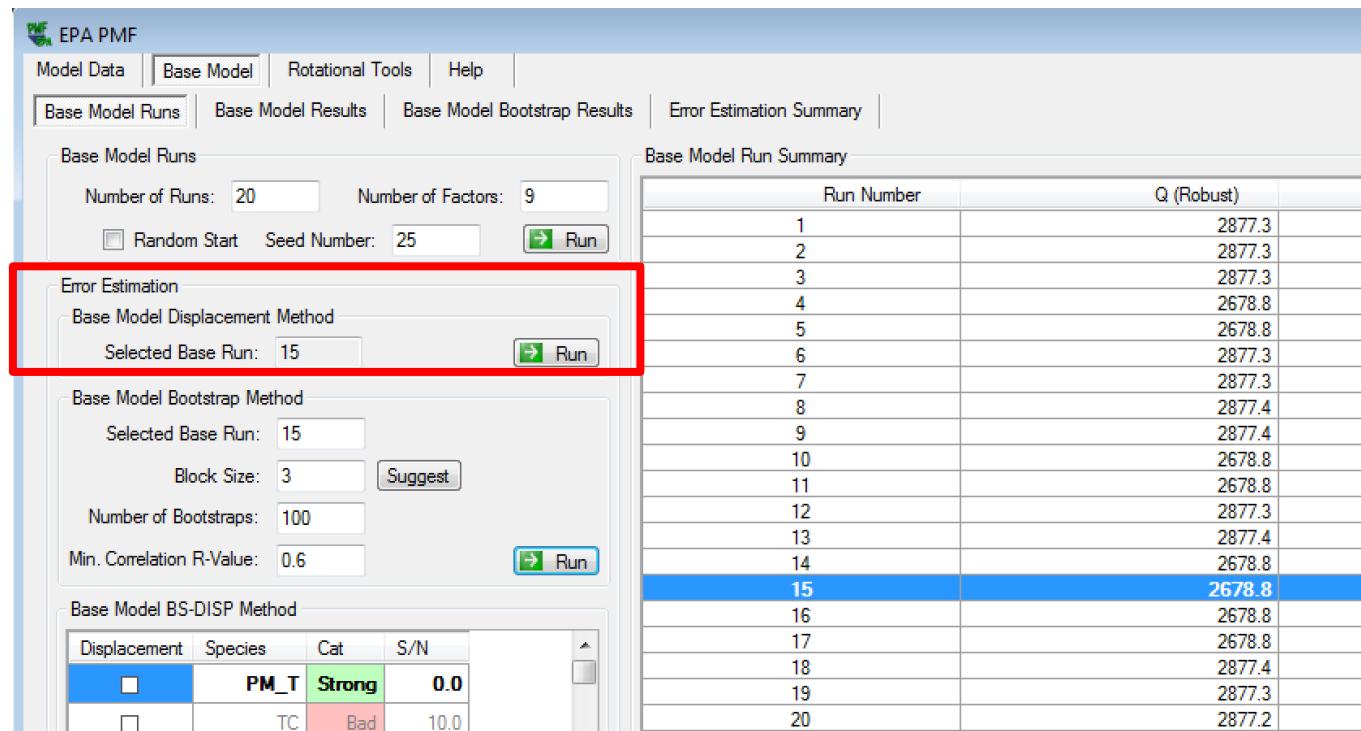
Controllare che i risultati del run di base (punti blu) cadano all'interno del box interquartile.



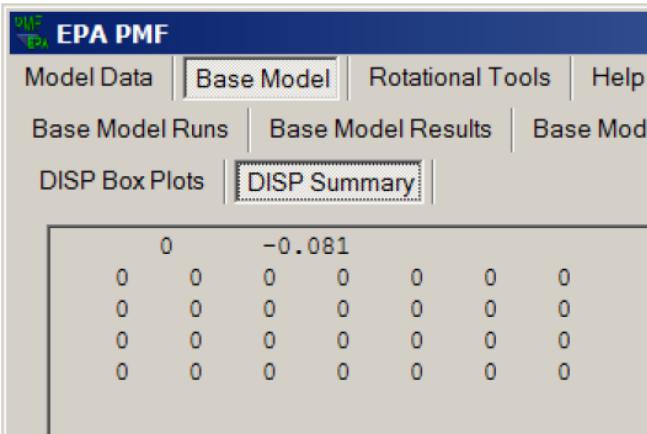
Displacement (DISP)

Esplora l'ambiguità rotazionale andando a determinare il più ampio intervallo nei valori degli elementi di F, tale da non produrre un aumento di Q rispetto alla soluzione base oltre una certa soglia (dQ_{max}). Fissato un dQ_{max} , ogni elemento di F è settato al valore massimo e minimo consentito e la relativa soluzione PMF è calcolata: il modello genera quindi un numero di uscite pari al numero di elementi di F per ciascun dQ_{max} .

Vengono generate soluzioni per $dQ_{max} = 4, 8, 15, 25$. Quelle relative a $dQ_{max} 4$ (le uniche graficate) sono generalmente considerate sufficienti per generare intervalli di errore robusti.



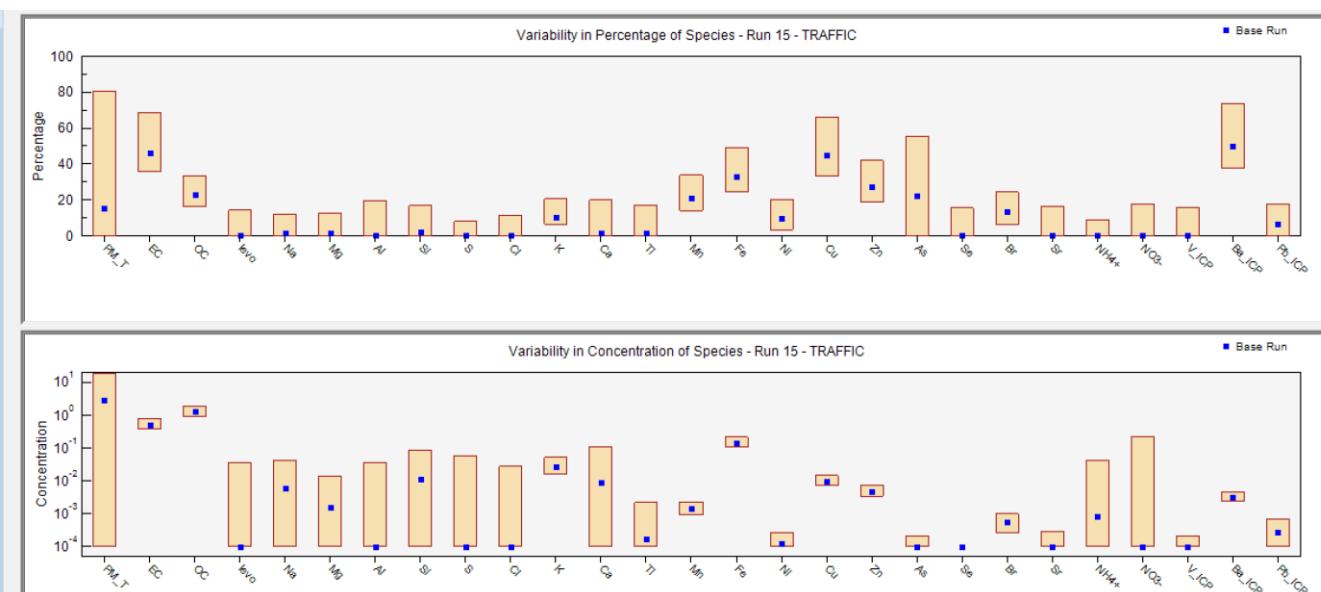
DISP diagnostic



- Primo numero: error code (0 = OK)
- Secondo valore: calo in Q (deve essere <1%, altrimenti significa che la soluzione di base non è min assoluto)
- Matrice con numero swaps nei diversi fattori (colonne) e per i diversi dQmax (righe): se ci sono swaps per dQmax 4 la soluzione non è sufficientemente robusta

Dipendenza da incertezze di input e numero fattori

Case 1. Small Errors	Case 2. Realistic Errors	Case 3. Simple Errors + 1 extra factor modeled
0 -0.111	0 0.000	0 -2.557
0 0 0 0	0 0 0 0	6 5 3 0 14
0 0 0 0	0 0 0 0	8 6 4 1 19
0 0 0 0	0 0 0 0	8 7 4 1 20
0 0 0 0	0 0 0 0	8 8 5 1 22



Grafici analoghi a quelli del BS, per tutti i fattori

BS-DISP

Combina BS e DISP (ogni campionamento del BS viene analizzato tramite DISP):

DISP copre lo spazio accessibile alle rotazioni, BS muove questo spazio in modo random nelle diverse direzioni: in questo modo si stima sia l'incertezza random che quella legata all'ambiguità rotazionale.

Richiede tempi di calcolo lunghi: conviene fare le prime prove con 50 run di BS ed effettuare il DISP solamente sulle specie più importanti e STRONG.

The screenshot shows the EPA PMF software interface with the following details:

- Top Navigation:** EPA PMF, Model Data, Base Model, Rotational Tools, Help.
- Sub-Navigation:** Base Model Runs, Base Model Results, Base Model DISP Results, Base Model Bootstrap Results, Error Estimation Summary.
- Base Model Runs Section:**
 - Number of Runs: 20, Number of Factors: 9.
 - Random Start, Seed Number: 25, .
 - Error Estimation, Base Model Displacement Method, Selected Base Run: 15, .
 - Base Model Bootstrap Method, Selected Base Run: 15, Block Size: 3, , Number of Bootstraps: 50, Min. Correlation R-Value: 0.6, .
- Base Model Run Summary Table:** A table showing Run Number and Q (Robust) values for 20 runs. The 15th run has a value of 2678.8, highlighted with a blue background.
- Base Model BS-DISP Method Section:** A table for the BS-DISP method with columns: Displacement, Species, Cat, S/N. The 'Species' column has rows for 'Si' and 'Strong', with 'Strong' being highlighted in green. The 'Cat' column has a value of 5.1. A red box highlights this section.

BS-DISP diagnostic

- Numero di run BS +1 (se non ci sono casi esclusi)
- Max calo in Q
- Numero di casi con calo di Q
- Numero di swaps nel best fit
- Numero di swaps in DISP
- Matrice swaps per i diversi dQmax (0.5, 1, 2, 4): per essere accettabile devono essere 0 per dQmax 0.5, altrimenti provare a ridurre n. fattori e/o vincoli

Model Data | Base Model | Rotational Tools | Help |

Base Model Runs | Base Model Results | Base Model Bootstrap Results | **Base Model BS-DISP Results** | Base Mo

BS-DISP Box Plots | **BS-DISP Summary**

101	-0.190	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

The five values in the first line are:

(1) k = # of cases in the file (includes both the full-data case and the accepted (not rejected) resamples). If no cases were excluded, k should be equal to the number of bootstraps * number of factors * number of species selected for BS-DISP.

(2) Largest decrease of Q. A large value is not alarming in itself, it only says that there was at least one resample where a deeper minimum appeared.

(3,4,5) # of cases with: /drop of Q / swap in best fit / swap in DISP/

BS-DISP: three examples (F. Amato)

51 -0.209 0 0 0	39 -21.149 8 0 4	11 -0.461 0 0 135
0 0 0 0	0 0 0 0	7 40 27 1 54
0 0 0 0	1 0 2 1	9 52 44 1 71
0 0 0 0	1 0 2 1	18 59 53 1 93
0 0 0 0	2 1 4 1	26 67 66 1 113

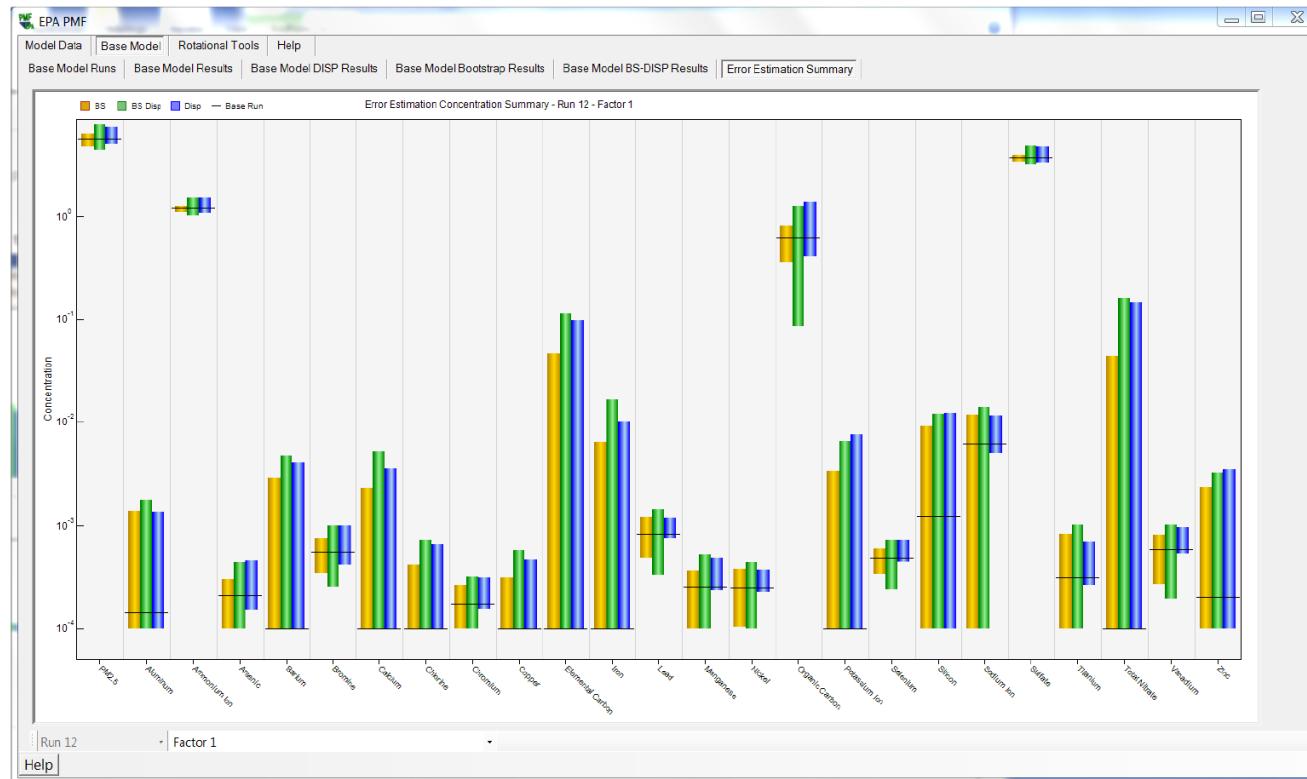


Error Estimation summary

I risultati ottenuti con questi 3 metodi (BS, DISP e BS/DISP) possono essere utilizzati per calcolare gli intervalli di confidenza sui profili e sui contributi delle sorgenti.

In particolare il modello fornisce tabelle (file _BaseErrorEstimationSummary) con:

- valori del run di base,
- mediana e percentili (5° e 95°) del BS
- media e percentili (5° e 95°) del BS-DISP
- media, min e max del DISP



What to report

- The specific chemical species used as input data (weak species)
- Total number of samples
- Q values, Q exp
- Missing/BDL values treatment
- Approach followed for the selection of optimal number of factors and the assignment of factor profiles to specific sources;
- Uncertainty estimate method
- BS mapping
- DISP swaps and dQ
- Constraints