

**Introduzione al
source
apportionment
tramite modelli a
recettore**



Source apportionment e modelli a recettore

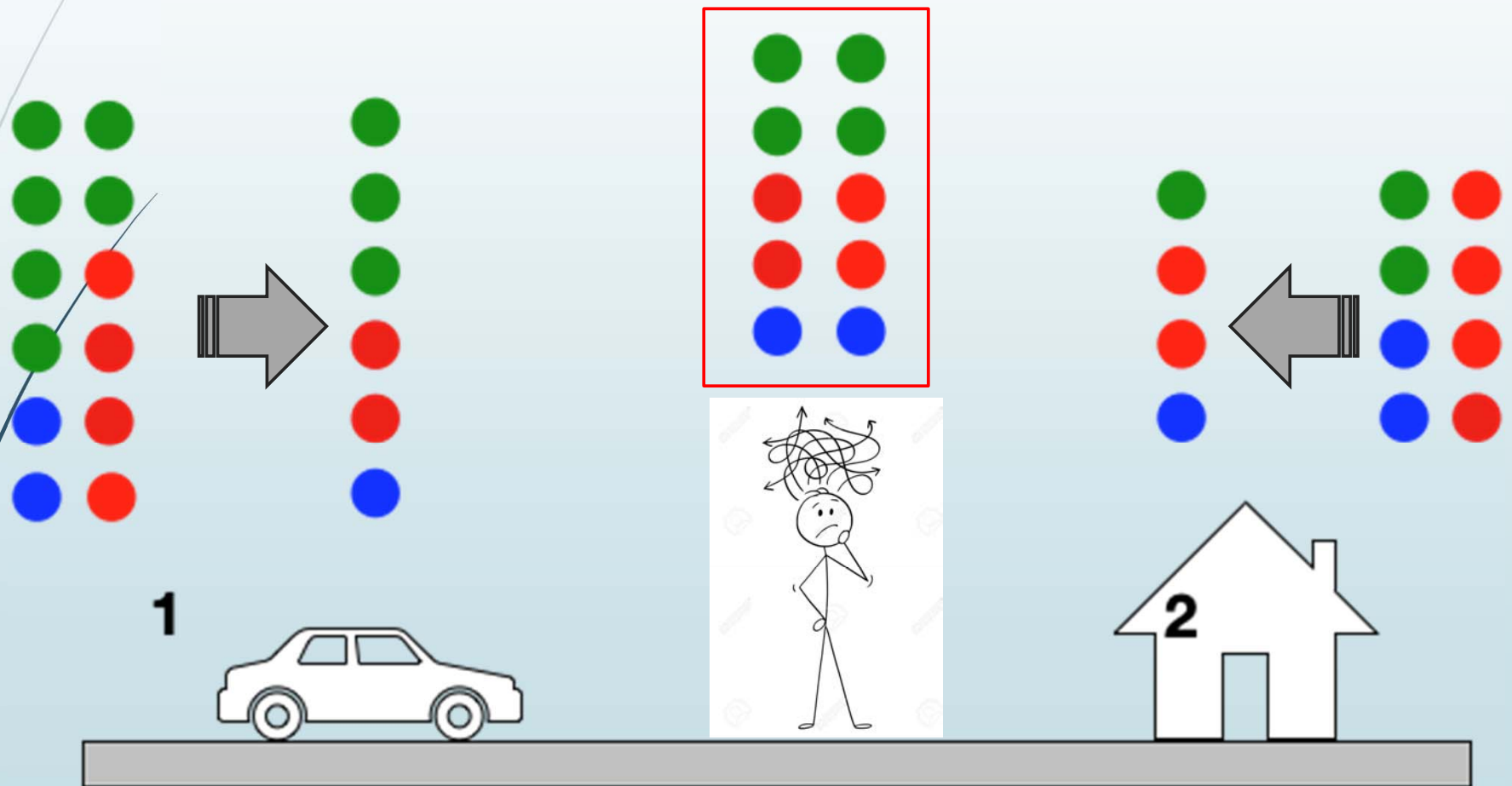
- Per Source Apportionment mediante modelli a recettore si intende **la ripartizione del contributo quantitativo in un punto, detto recettore, delle diverse categorie di sorgenti di un inquinante o di un insieme di inquinanti attraverso algoritmi che considerano le immissioni rilevate nel recettore, senza conoscere prioritariamente le quantità emesse dalle diverse sorgenti presenti nel territorio che possono influire direttamente o indirettamente.**

Source apportionment e modelli a recettore

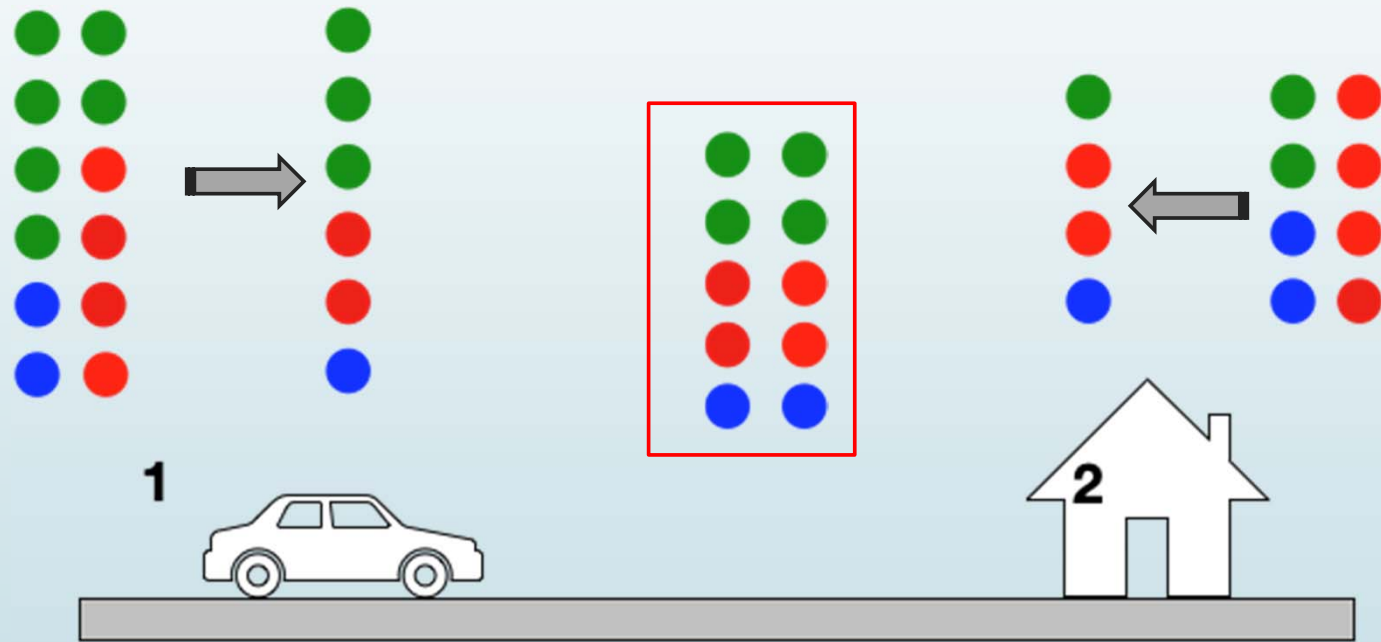
Perché servono?



Quali sono le sorgenti degli inquinanti osservati nel sito recettore?



Quali sono le sorgenti degli inquinanti osservati nel sito recettore?

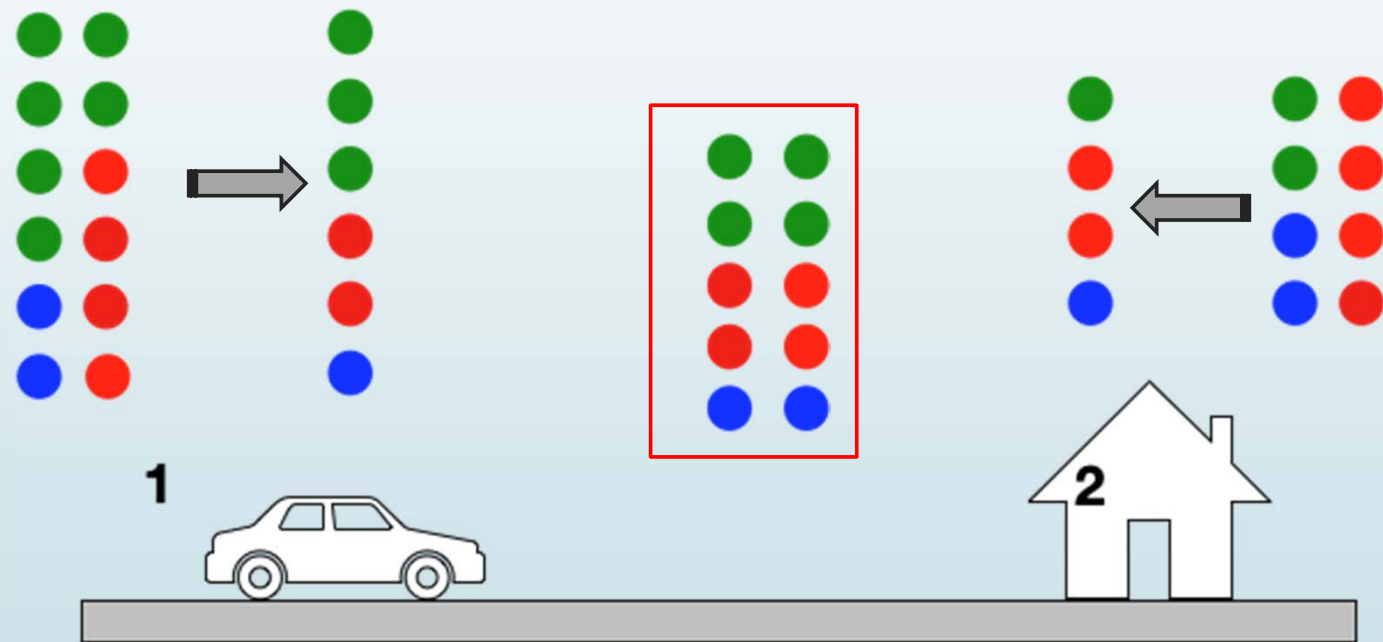


A numero di ●
B numero di ●
C numero di ●

$$\begin{aligned} [A] &= M_1 * fa_1 + M_2 * fa_2 \\ [B] &= M_1 * fb_1 + M_2 * fb_2 \\ [C] &= M_1 * fc_1 + M_2 * fc_2 \end{aligned}$$

Se conosco $fa_1, fb_1, fc_1,$
 $fa_2, fb_2, fc_2,$

Quali sono le sorgenti degli inquinanti osservati nel sito recettore?



A numero di ●
 B numero di ●
 C numero di ●

$$[A] = M_1 * fa_1 + M_2 * fa_2$$

$$[B] = M_1 * fb_1 + M_2 * fb_2$$

$$[C] = M_1 * fc_1 + M_2 * fc_2$$

Se conosco $fa_1, fb_1, fc_1,$
 $fa_2, fb_2, fc_2,$

**Sistema di 3 equazioni in
 2 incognite: risolvibile**

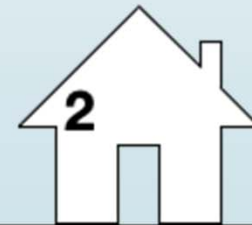
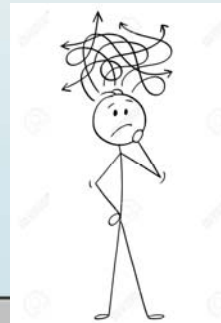
Quali sono le sorgenti degli inquinanti osservati nel sito recettore?



$$\begin{aligned} [A] &= 8 \\ [B] &= 6 \\ [C] &= 4 \end{aligned}$$



1



A numero di ●
B numero di ●
C numero di ●

$$\begin{aligned} M_1 &= ? \\ M_2 &= ? \end{aligned}$$

**Problema di source
apportionment risolvibile
con modello a recettore**



Source apportionment e modelli a recettore

- ▶ Alcuni modelli a recettore richiedono la conoscenza **quantitativa** delle emissioni delle categorie di **sorgenti** e, combinando queste informazioni con le immissioni, ne ricavano il relativo contributo.
- ▶ Il Chemical Mass Balance (**CMB**) è uno dei più noti tra i modelli a recettore appartenenti a questa prima classe

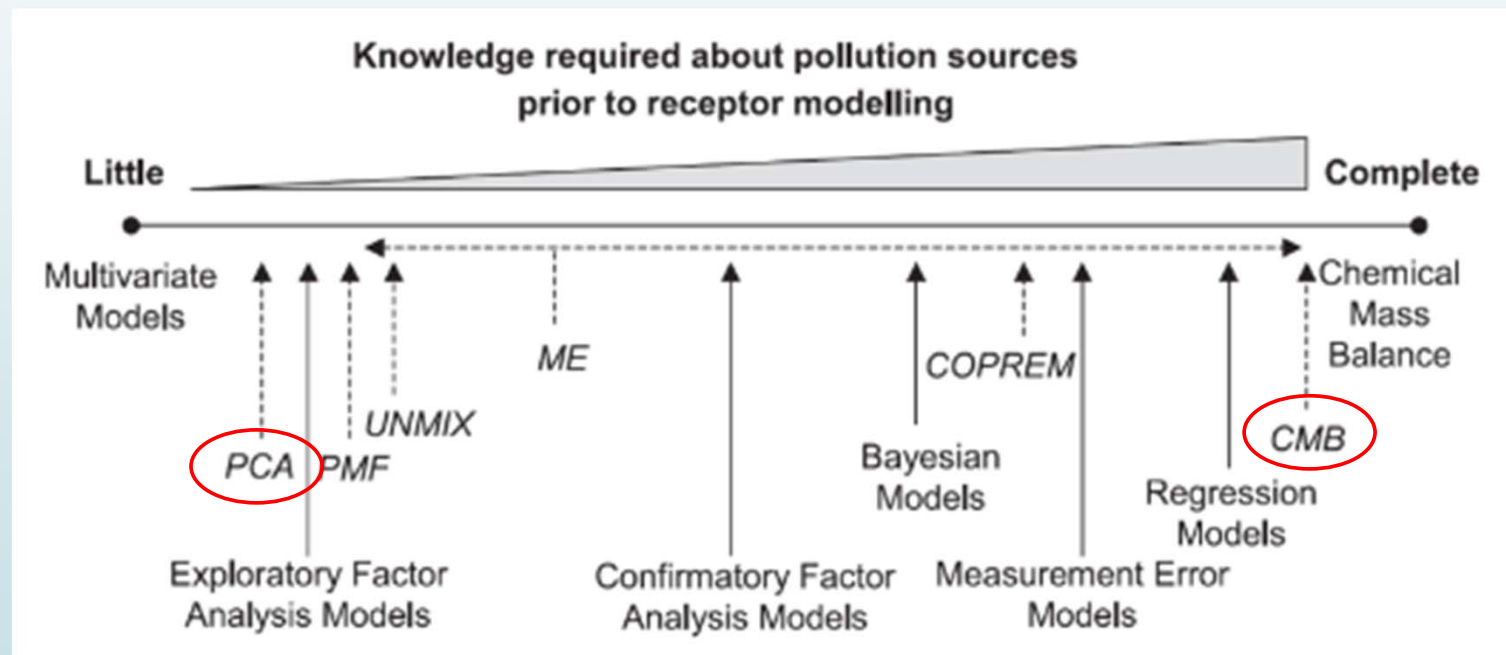


Source apportionment e modelli a recettore

- ▶ Alcuni modelli a recettore richiedono la conoscenza **quantitativa** delle emissioni delle categorie di **sorgenti** e, combinando queste informazioni con le immissioni, ne ricavano il relativo contributo.
- ▶ Il Chemical Mass Balance (**CMB**) è uno dei più noti tra i modelli a recettore appartenenti a questa prima classe
- ▶ Altri modelli, sfruttando la **variabilità intrinseca nel database** delle immissioni al recettore, stimano sia il contributo che la composizione quantitativa delle emissioni delle categorie di sorgenti.

Source apportionment e modelli a recettore

- La Positive Matrix Factorization (PMF) appartiene a questa seconda classe.



- Nella sua versione 5 (PMF5 USA-EPA) sono state implementate alcune opzioni che consentono di miscelare in parte le caratteristiche delle due classi citate di modelli a recettore.

A dark blue arrow points to the right at the top left. Below it, several thin, curved lines in shades of blue and grey sweep across the left side of the slide.

Source apportionment e modelli a recettore

- ▶ I modelli a recettore sono algoritmi di **natura statistica**, che non tengono conto delle trasformazioni che i vari inquinanti possono subire nel trasporto dalla sorgente al recettore.
- Equazioni che descrivono l'equilibrio chimico
- Equazioni che descrivono la cinetica di reazione
- Equazioni che descrivono i processi di diffusione
- Equazioni che descrivono i processi di rimozione



Source apportionment e modelli a recettore

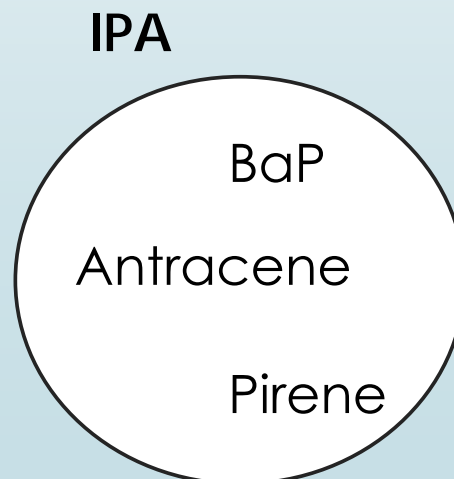
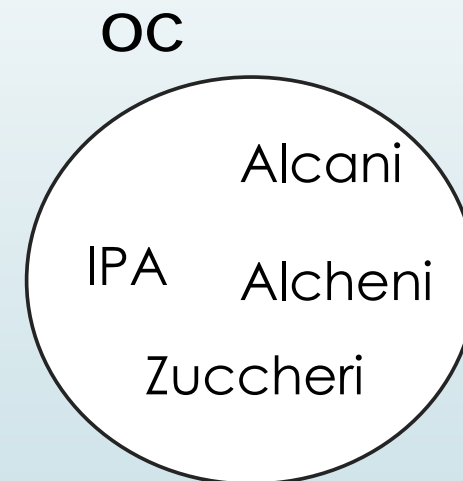
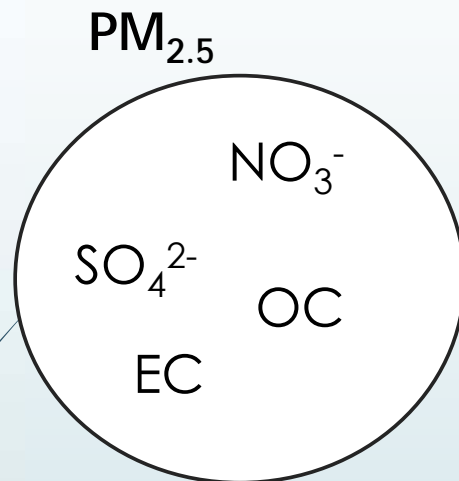
- I modelli a recettore sono algoritmi di **natura statistica**, che non tengono conto delle trasformazioni che i vari inquinanti possono subire nel trasporto dalla sorgente al recettore.
- Tuttavia, anzi proprio per tale motivo, richiedono specifiche competenze all'utilizzatore sia riguardanti i **processi chimico-fisici** che gli inquinanti possono subire, sia sulle **tecniche di campionamento** e di analisi, che una **conoscenza del territorio** oggetto di studio, al fine di poter adottare/modificare in modo efficace i parametri previsti dal modello, poter valutare adeguatamente i limiti del modello ed infine poter interpretare adeguatamente i risultati prodotti.



Nomenclatura

- ▶ Modello a recettore
- ▶ Source apportionment
- ▶ Variabile globale: insieme di specie chimiche differenti che nel contesto in studio possono essere considerate nel loro insieme come un unico inquinante

Variabile globale





Nomenclatura

- ▶ Modello a recettore
- ▶ Source apportionment
- ▶ Variabile globale: insieme di specie chimiche differenti che nel contesto in studio possono essere considerate nel loro insieme come un unico inquinante
- ▶ Categoria di sorgenti o sorgente: un insieme di sorgenti che emettono la **variabile globale** con rapporti tra le specie chimiche componenti che nel contesto in studio possono essere considerati costanti.

Categoria di sorgenti o sorgente



autoveicoli a benzina: ogni autoveicolo a benzina emette allo scarico composti chimici differenti. Considerando la flotta di veicoli nel contesto di studio, possiamo assumere costanti i rapporti tra le singole specie chimiche.

Biomass burning per riscaldamento domestico: ogni impianto emette composti chimici differenti. Nel contesto in studio possiamo assumere costanti i rapporti tra le singole specie chimiche.

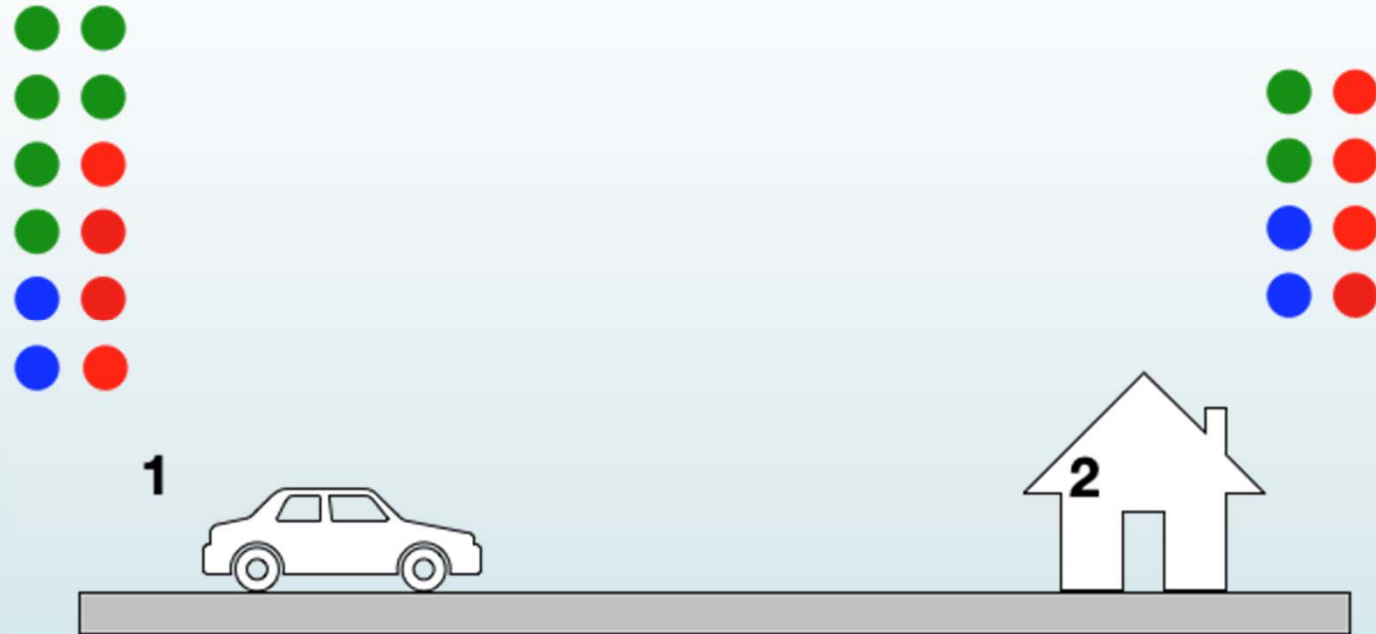




Nomenclatura

- ▶ Modello a recettore
- ▶ Source apportionment
- ▶ Variabile globale: insieme di specie chimiche differenti che nel contesto in studio possono essere considerate nel loro insieme come un unico inquinante
- ▶ Categoria di sorgenti o sorgente: un insieme di sorgenti che emettono la **variabile globale** con rapporti tra le specie chimiche componenti che nel contesto in studio possono essere considerati costanti
- ▶ Fingerprint di una sorgente: insieme dei rapporti f_j tra diverse specie chimiche presenti nella **variabile globale** emessa dalla **sorgente** considerata rispetto ad una od un insieme di esse.

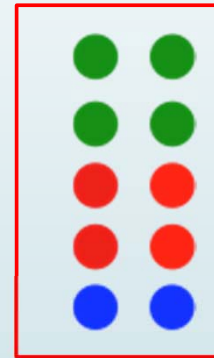
Fingerprint (o profili)



● $4/12 = 0.33$
● $2/12 = 0.17$
● $6/12 = 0.50$

$4/8 = 0.50$
 $2/8 = 0.25$
 $2/8 = 0.25$

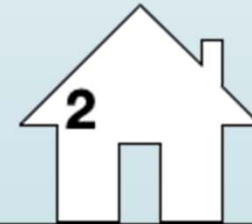
PMF: Positive Matrix Factorization



1



2



A numero di ●
 B numero di ●
 C numero di ●

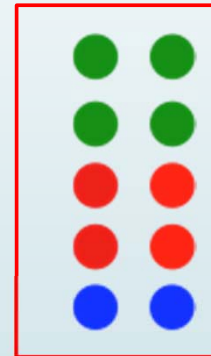
$$[A] = M_1 * fa_1 + M_2 * fa_2$$

$$[B] = M_1 * fb_1 + M_2 * fb_2$$

$$[C] = M_1 * fc_1 + M_2 * fc_2$$

Numero di incognite?

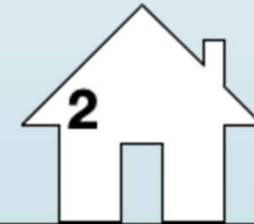
PMF: Positive Matrix Factorization



1



2



A numero di ●
B numero di ●
C numero di ●

$$\begin{aligned} [A] &= M_1 * fa_1 + M_2 * fa_2 \\ [B] &= M_1 * fb_1 + M_2 * fb_2 \\ [C] &= M_1 * fc_1 + M_2 * fc_2 \end{aligned}$$

**Sistema di 3 equazioni
in 8 incognite:
problema non
risolvibile**



PMF: Positive Matrix Factorization

Al tempo t_1 :

$$[A]_{t_1} = M_{1t_1} * fa_1 + M_{2t_1} * fa_2$$

$$[B]_{t_1} = M_{1t_1} * fb_1 + M_{2t_1} * fb_2$$

$$[C]_{t_1} = M_{1t_1} * fc_1 + M_{2t_1} * fc_2$$

**variabilità intrinseca nel database
delle immissioni la recettore**

T1:3 equazioni 8 incognite

PMF: Positive Matrix Factorization

Al tempo t1:

$$[A]_{t1} = M_{1t1} * fa_1 + M_{2t1} * fa_2$$

$$[B]_{t1} = M_{1t1} * fb_1 + M_{2t1} * fb_2$$

$$[C]_{t1} = M_{1t1} * fc_1 + M_{2t1} * fc_2$$

T1:3 equazioni 8 incognite

Al tempo t2:

$$[A]_{t2} = M_{1t2} * fa_1 + M_{2t2} * fa_2$$

$$[B]_{t2} = M_{1t2} * fb_1 + M_{2t2} * fb_2$$

$$[C]_{t2} = M_{1t2} * fc_1 + M_{2t2} * fc_2$$

T2:6 equazioni 10 incognite

Al tempo t3:

$$[A]_{t3} = M_{1t3} * fa_1 + M_{2t3} * fa_2$$

$$[B]_{t3} = M_{1t3} * fb_1 + M_{2t3} * fb_2$$

$$[C]_{t3} = M_{1t3} * fc_1 + M_{2t3} * fc_2$$

T3:9 equazioni 12 incognite

T4:12 equazioni 14 incognite

T5:15 equazioni 16 incognite

T6:18 equazioni 18 incognite

PMF: Positive Matrix Factorization

Con n istanti temporali (o campioni), 2 sorgenti e 3 variabili

$3n$ è il numero di equazioni
 $6 + 2n$ è il numero di incognite

$$3n > 6 + 2n$$

$$n > 6$$

T1:3 equazioni 8 incognite

T2:6 equazioni 10 incognite

T3:9 equazioni 12 incognite

T4:12 equazioni 14 incognite

T5:15 equazioni 16 incognite

T6:18 equazioni 18 incognite

PMF: Positive Matrix Factorization

Con n campioni, p sorgenti e m variabili, all'istante t_n possiamo scrivere:

$$\begin{aligned} [x_1]_{t_n} &= M_{1 t_n} * fa_1 + M_{2 t_n} * fa_2 + \dots + M_{k t_n} * fa_p \\ [x_2]_{t_n} &= M_{1 t_n} * fb_1 + M_{2 t_n} * fb_2 + \dots + M_{k t_n} * fb_p \\ [x_3]_{t_n} &= M_{1 t_n} * fc_1 + M_{2 t_n} * fc_2 + \dots + M_{k t_n} * fc_p \\ &\dots \\ &\dots \\ [x_m]_{t_n} &= M_{1 t_n} * fm_1 + M_{2 t_n} * fm_2 + \dots + M_{k t_n} * fm_p \end{aligned}$$

$$x_{i,j} = \sum_{k=1}^p g_{i,k} \cdot f_{k,j} + \varepsilon_{i,j}$$

i-esimo campione (1-n)
j-esima specie (1-m)
k-esima sorgente (1-p)

PMF: Positive Matrix Factorization

Fattorizzazione di una matrice nota \mathbf{X} in due matrici non note \mathbf{G} ed \mathbf{F} ad elementi tutti positivi

$$\mathbf{X} = \mathbf{G} \times \mathbf{F}$$

$$\mathbf{X} = [x_{i,j}], \quad \mathbf{G} = [g_{i,k}], \quad \mathbf{F} = [f_{k,j}]$$

$$x_{i,j} \geq 0 \quad \forall i, j; \quad g_{i,k} \geq 0 \quad \forall i, k; \quad f_{k,j} \geq 0 \quad \forall k, j$$

$x_{i,j}$ concentrazione della specie j -esima nel campione i -esimo

$g_{i,k}$ contributo della sorgente k -esima al campione i -esimo

$f_{k,j}$ concentrazione della specie j -esima nel profilo sorgente k -esima

PMF: Positive Matrix Factorization

Fattorizzazione di una matrice nota X in due matrici non note G ed F ad elementi tutti positivi

$$X = G \times F$$

$$X = [x_{i,j}], G = [g_{i,k}], F = [f_{k,j}]$$

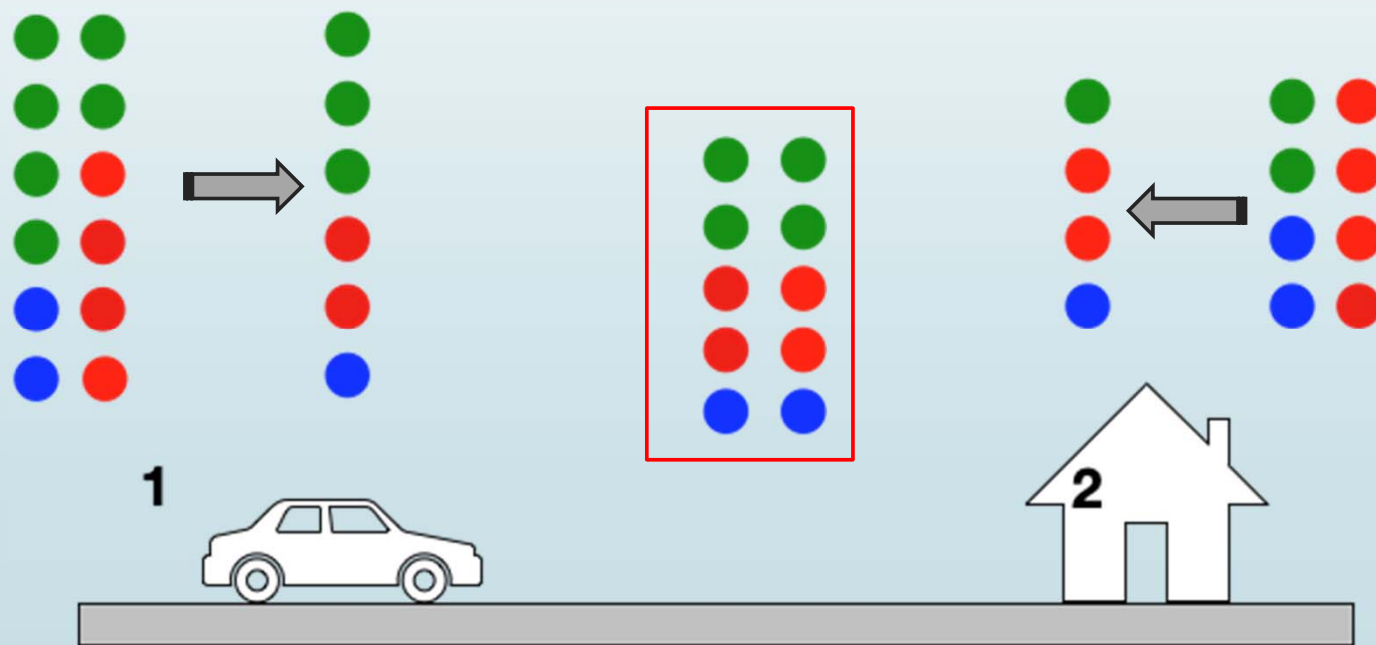
n numero campioni
 m numero di variabili
 p numero di sorgenti

Numero equazioni $n \times m >$ Numero incognite $p \times (m + n)$

PMF: Positive Matrix Factorization

Assunzione fondamentale:

Le specie chimiche emesse dalle sorgenti non subiscono trasformazioni chimiche nel tragitto dal punto di emissione al recettore.



PMF: Positive Matrix Factorization

Assunzione fondamentale:

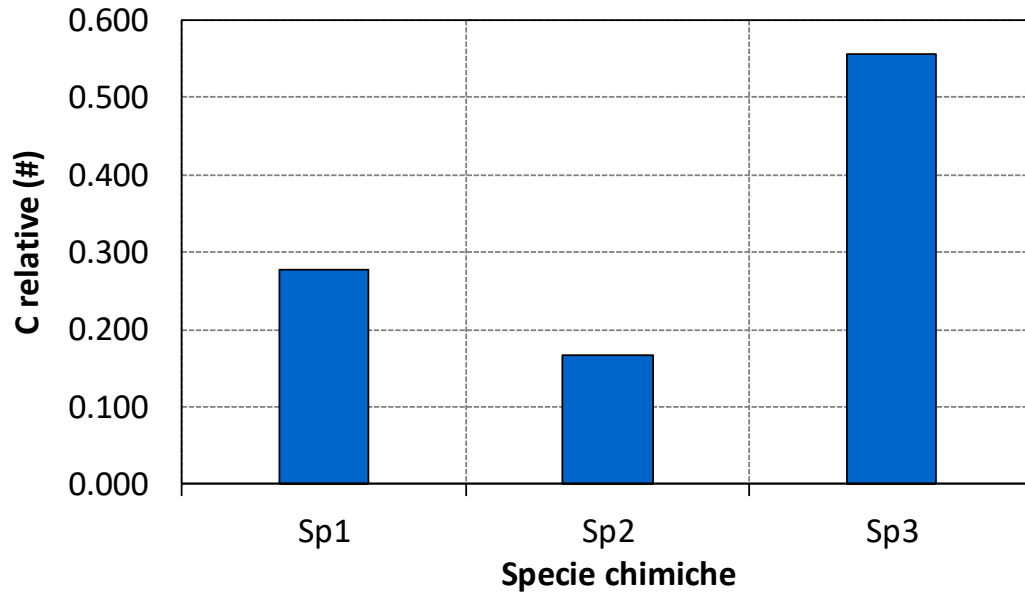
Le specie chimiche emesse dalle sorgenti non subiscono trasformazioni chimiche nel tragitto dal punto di emissione al recettore.

Dati p fattori (sorgenti), considerato l' i -esimo campione, allora la concentrazione $x_{i,j}$ della j -esima specie chimica in un punto recettore può essere scritta come:

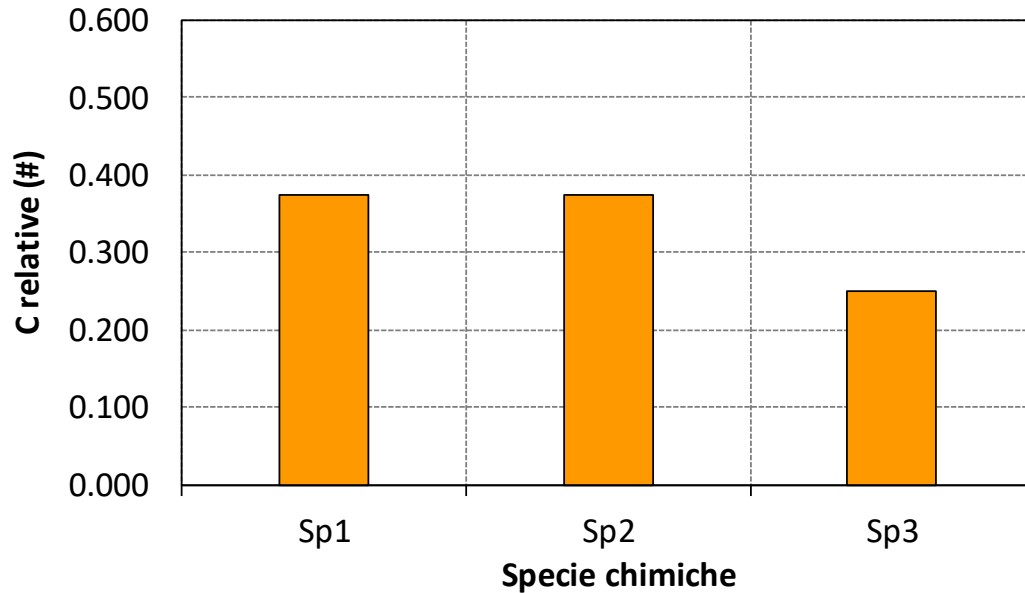
$$x_{i,j} = \sum_{k=1}^p g_{i,k} \cdot f_{k,j} + \varepsilon_{i,j}$$

$g_{i,k}$ è il contributo all' i -esimo campione della k -esimo fattore e $f_{k,j}$ la j -esima componente del profilo del k -esimo fattore e $\varepsilon_{i,j}$ rappresenta il residuo per ciascun campione e specie

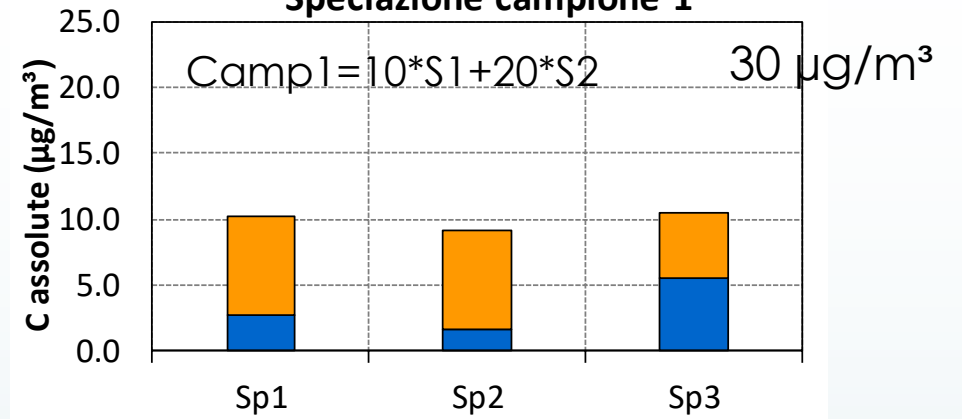
Sorgente S1: profilo



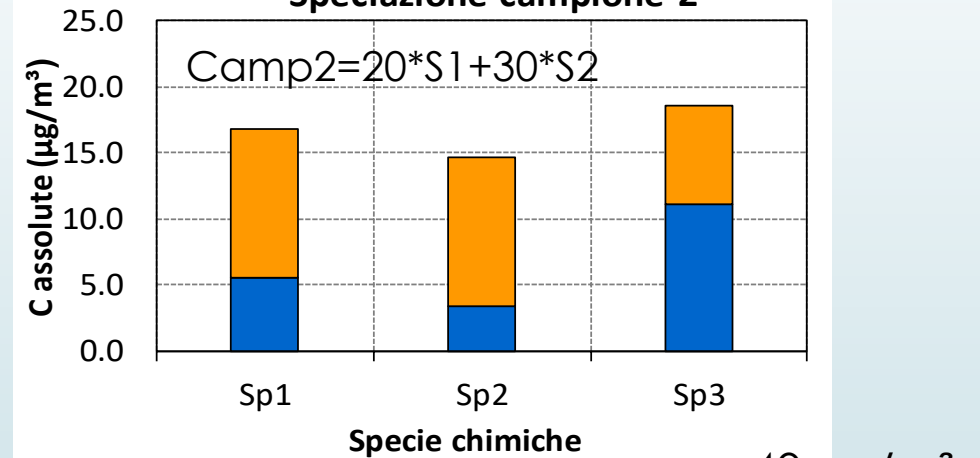
Sorgente S2: profilo



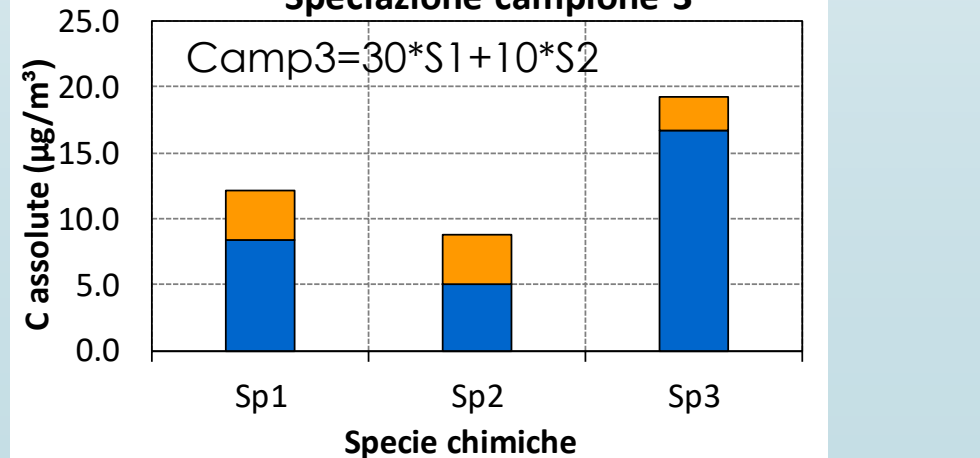
Speciazione campione 1



Speciazione campione 2



Speciazione campione 3



PMF: Positive Matrix Factorization

Variabile globale

$$\begin{bmatrix} 30 \\ 50 \\ 40 \end{bmatrix}$$

=

$$\begin{bmatrix} 10.3 & 9.2 & 10.6 \\ 16.8 & 14.6 & 18.6 \\ 12.1 & 8.8 & 19.2 \end{bmatrix}$$

=

$$\begin{bmatrix} 10 & 20 \\ 20 & 30 \\ 30 & 10 \end{bmatrix}$$

x

$$\begin{bmatrix} 0.28 & 0.17 & 0.56 \\ 0.38 & 0.38 & 0.25 \\ & & \end{bmatrix}$$

$x_{i,j}$

$g_{i,k}$

$f_{k,j}$



Esempio pratico

Source Apportionment of
PM_{2.5} in Delhi, India Using PMF
Model (Bull Environ Contam
Toxicol (2016) 97:286–293)

$n = 140$ campioni

$m = 23$ variabili

$p = 7$ sorgenti

E' risolvibile?

Esempio pratico

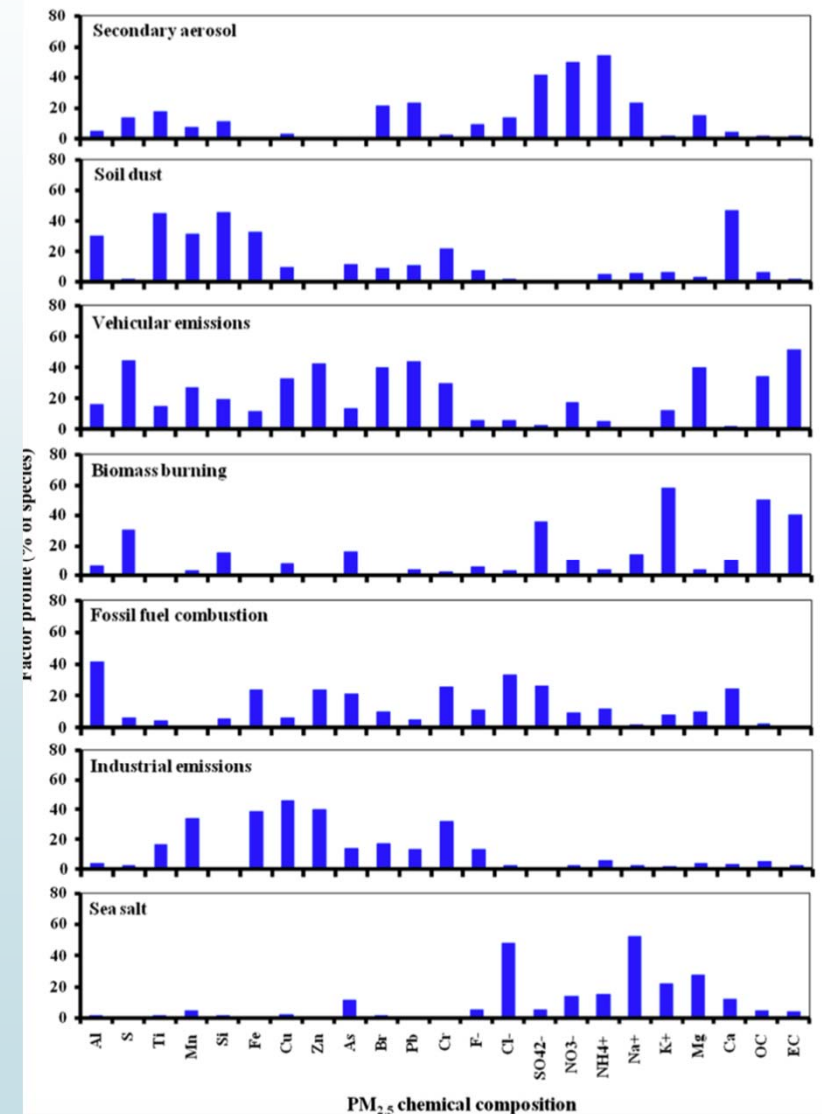
Source Apportionment of PM_{2.5} in Delhi, India Using PMF Model (Bull Environ Contam Toxicol (2016) 97:286–293)

n = 140 campioni
m = 23 variabili
p = 7 sorgenti

Numero equazioni $n \times m$
Numero incognite $p \times (n + m)$

$$3360 > 7 \times (163)$$

Il problema è sovradeterminato (ammette in generale infinite soluzioni) e non è risolvibile analiticamente



Come l'algoritmo della PMF risolve il problema del SA?

- Minimizzando il **fattore di merito** (o **funzione obiettivo**):

$$Q = \sum_i^n \sum_j^m \left[\frac{(x_{i,j} - \sum_k^p g_{i,k} f_{k,j})}{u_{i,j}} \right]^2$$

n : numero di campioni; m : numero di specie della variabile globale; p : numero di fattori

$u_{i,j}$: incertezza attribuita alla j -esima specie $x_{i,j}$ del campione i e comprende

- **l'incertezza analitica**
- **l'incertezza del campionamento**
- ma anche l'incertezza relativa alla validità **dell'assunzione** fondamentale del modello

Come l'algoritmo della PMF risolve il problema del SA?

- Minimizzando il **fattore di merito** (o **funzione obiettivo**):

$$Q = \sum_i^n \sum_j^m \left[\frac{(x_{i,j} - \sum_k^p g_{i,k} f_{k,j})}{u_{i,j}} \right]^2$$

Q_{robusto} = calcolato escludendo punti per i quali
l'incertezza scalata per i residui è maggiore di 4 ($Q_{ij} > 4$)

Q_{robusto} non è influenzato da outlier

Input del modello

- Le concentrazioni misurate $x_{i,j}$
- Le incertezze $u_{i,j}$

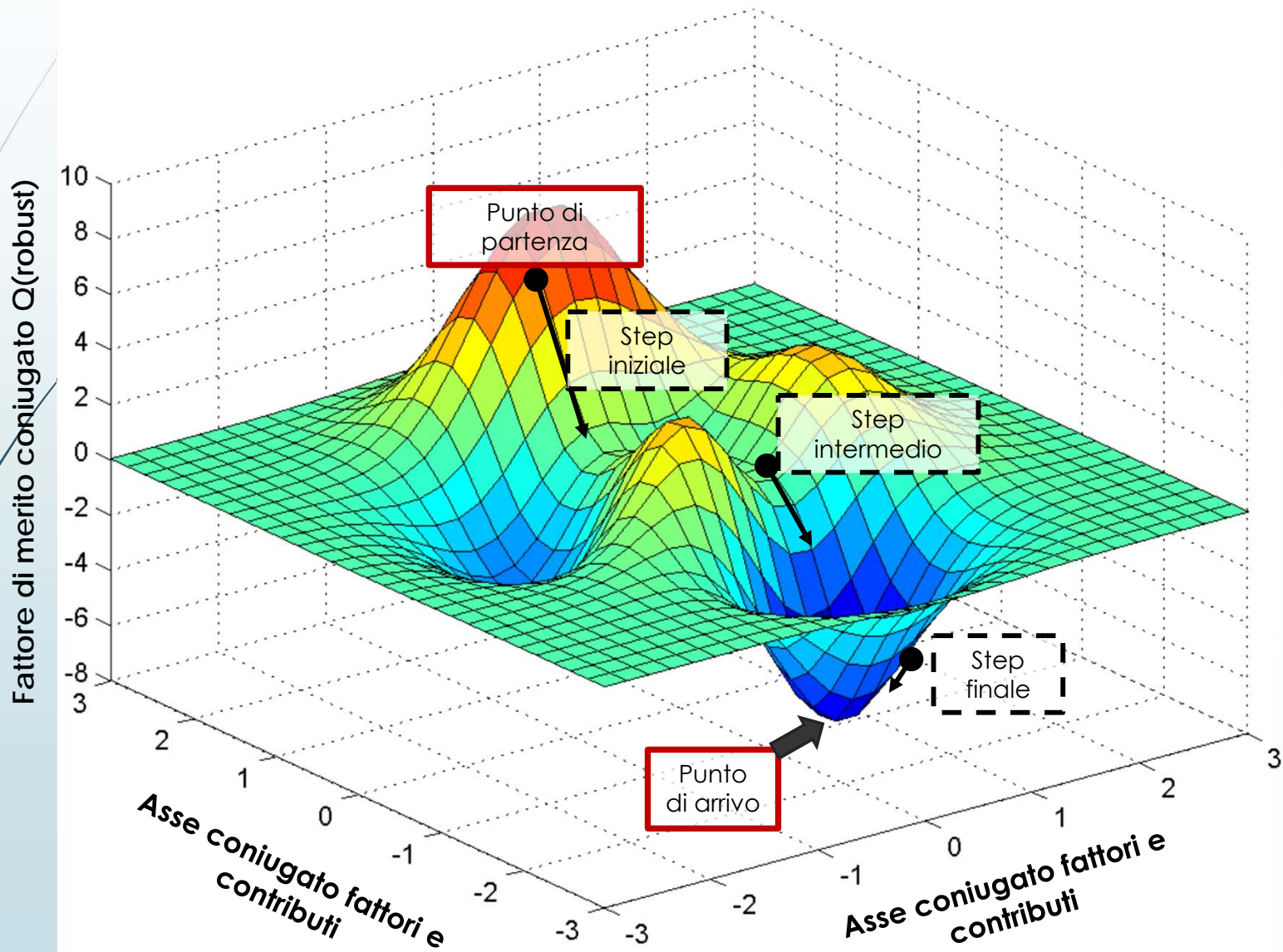
- Se $x_{i,j} < DL$ $u_{i,j} = 5/6 \times DL$

- Se $x_{i,j} > DL$ $u_{i,j} = \sqrt{(\varepsilon \cdot x_{i,j})^2 + \left(\frac{1}{2} \cdot mdl_j\right)^2}$

oppure calcolata dall'utente
(error propagation equations)

Avendo introdotto le incertezze, valori negativi per le concentrazioni, i contributi e i fattori sono ammessi, purché non significativamente differenti da zero

Ricerca della soluzione



Ricerca della soluzione

Partendo da un punto casuale nello spazio di $p \times m$ dimensioni (spazio coniugato), l'algoritmo ricerca il minimo di $Q(\text{robust})$ attraverso iterazioni con tre livelli di **iterazioni**: iniziale, intermedio, finale.

1° livello - iniziale: $dQ < 0.1$ (su 20 passi consecutivi in meno di 800 passi)

2° livello - intermedio: $dQ < 0.005$ (su 50 passi consecutivi in meno di 2000 passi)

3° livello - finale: $dQ < 0.0003$ (su 100 passi consecutivi in meno di 5000 passi)

Se nessuna non viene trovata alcuna soluzione che rispetta i requisiti precedenti si dice che la soluzione non converge.

Per trovare minimo assoluto vengono effettuati diversi *run* (20-100), ciascuno a partire da un punto di partenza differente

Valore di Q atteso

$$Q = \sum_i^n \sum_j^m \left[\frac{(x_{i,j} - \sum_k^p g_{i,k} f_{k,j})}{u_{i,j}} \right]^2$$

Q_{expected} è uguale al numero di gradi di libertà del sistema

$Q_{\text{expected}} = \text{numero di dati} - \text{numero di vincoli}$

$$Q_{\text{expected}} = m \times n - p \quad (n+m)$$

Q/Q_{expected} deve tendere a 1*

*se incertezze dimensione db sono corretti



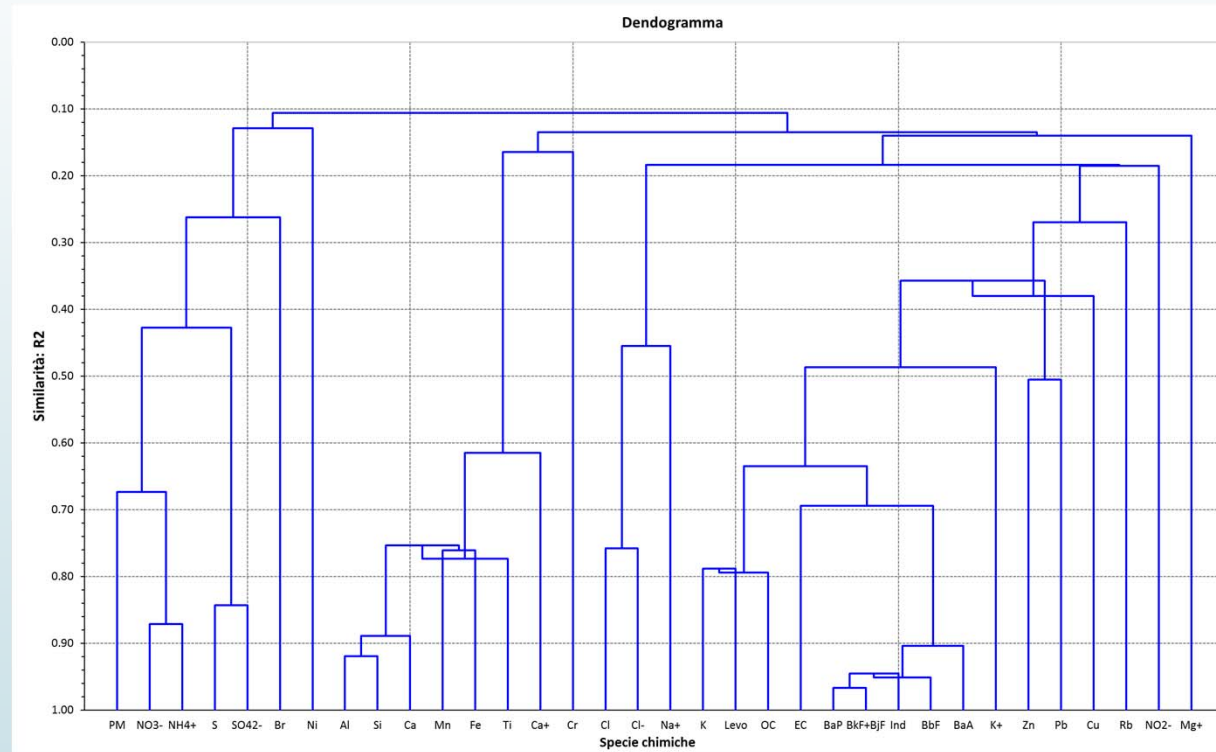
Quale è la dimensione minima del database (X)

- E' consigliato fortemente che il numero dei campioni sia almeno il triplo rispetto al numero di specie che vengono utilizzate nei calcoli.

Si tenga presente che:

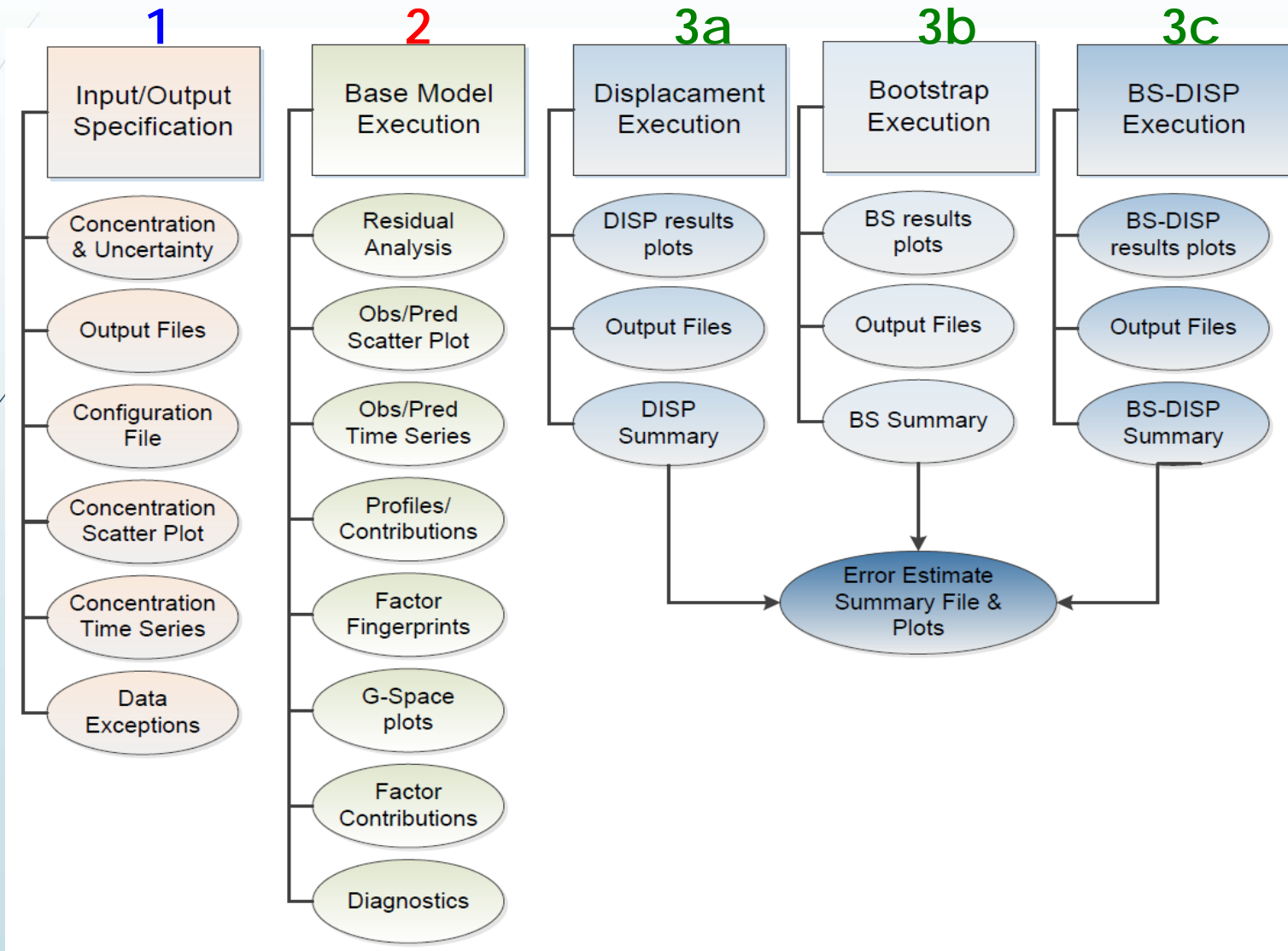
- la PMF5 consente all'utilizzatore di escludere dai calcoli alcune variabili (specie) presenti nel database
- Se due variabili sono fortemente correlate tra loro e lo scarto dalla linearità può essere considerato 'rumore' allora le due variabili portano la stessa informazione; una delle due può quindi essere trascurata.

Quale è la dimensione minima del database (X)

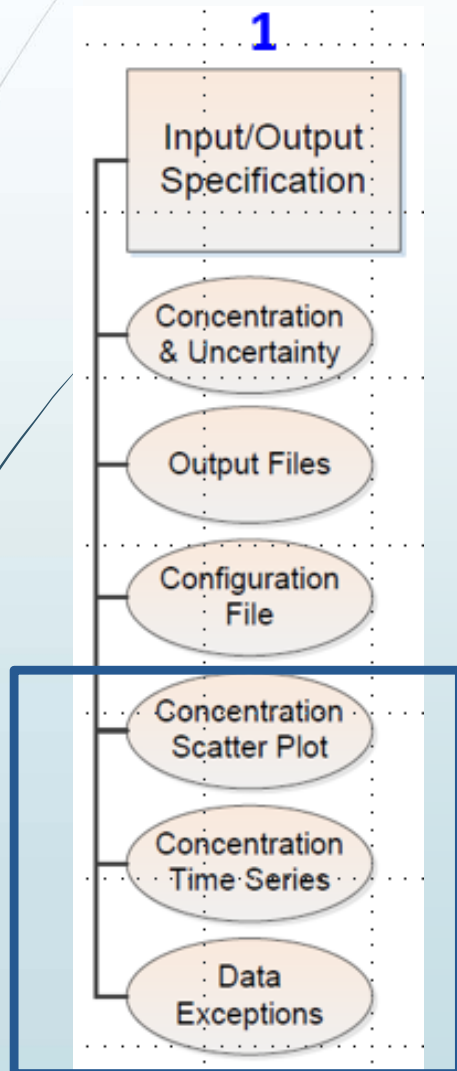


L'analisi a cluster con **indice di similarità R2** consente, tra altro, di eliminare nelle elaborazioni le variabili tra loro molto correlate:

Esecuzione del modello

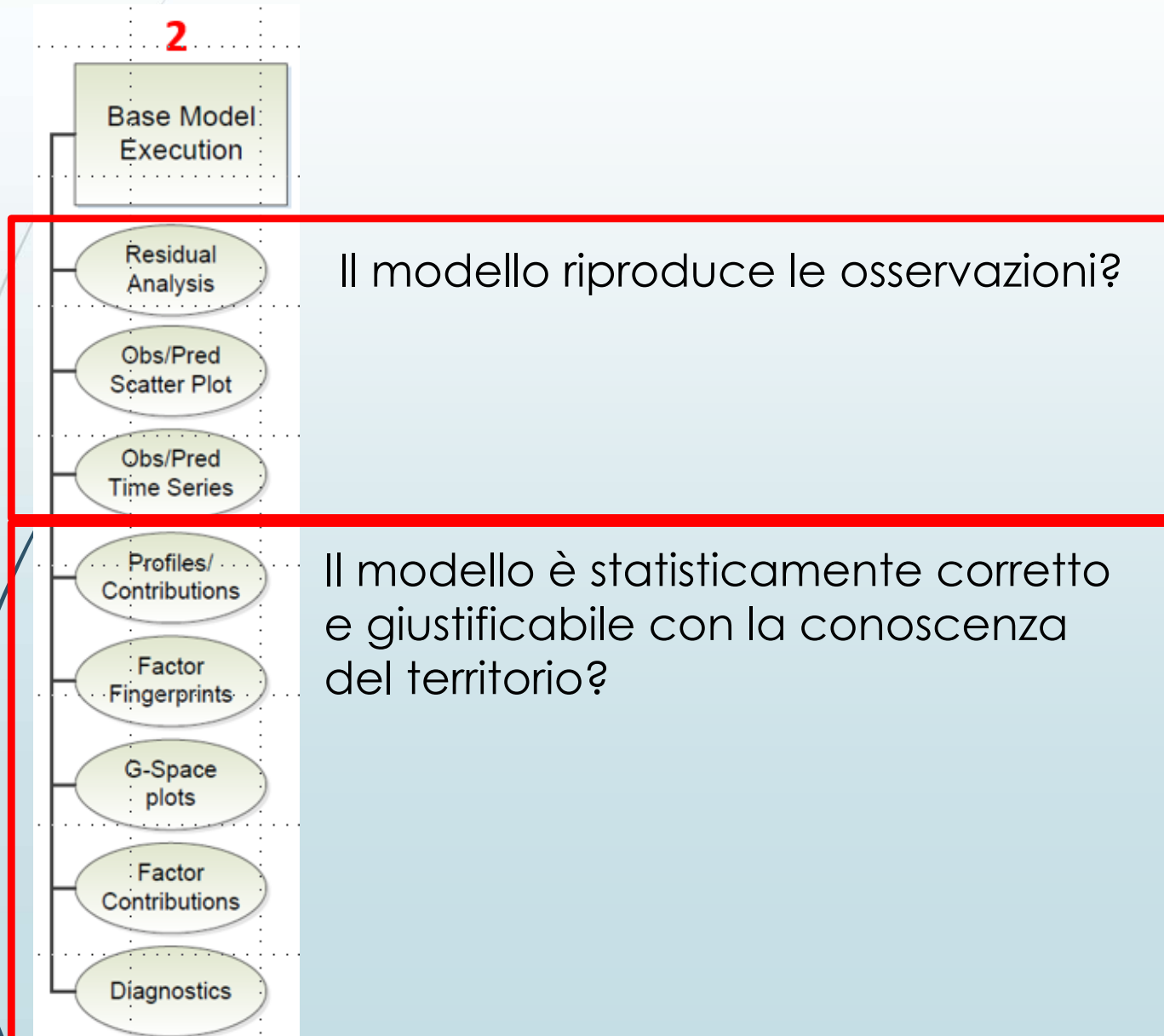


Esecuzione del modello



Offre strumenti per esplorare il database, verificando l'esistenza di eventuali correlazioni tra variabili, la presenza di campioni outlier o di stagionalità nelle serie temporali delle variabili, con la possibilità di eliminare campioni dal calcolo, non dal database, ecc.

Esecuzione del modello



Base Model Runs

Number of Runs: 20 Number of Factors: 6

Random Start Seed Number: 79

Error Estimation

Base Model Displacement Method

Selected Base Run: 5

Base Model Bootstrap Method

Selected Base Run: 5

Block Size: 22

Number of Bootstraps: 20

Min. Correlation R-Value: 0.6

Base Model BS-DISP Method

Displacement	Species	Cat	S/N
<input checked="" type="checkbox"/>	PM2.5	Strong	9.0
<input type="checkbox"/>	Aluminum	Strong	0.1
<input type="checkbox"/>	Ammonium Ion	Strong	8.9
<input type="checkbox"/>	Arsenic	Strong	0.1

Factor Names

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Run 5	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Run 6	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Run 7	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6

Run Progress

Help HAVE Concentration Data HAVE Uncertainty Data HAVE Base Results NO DISP Results NO Bootstrap Results NO BS-DISP Results

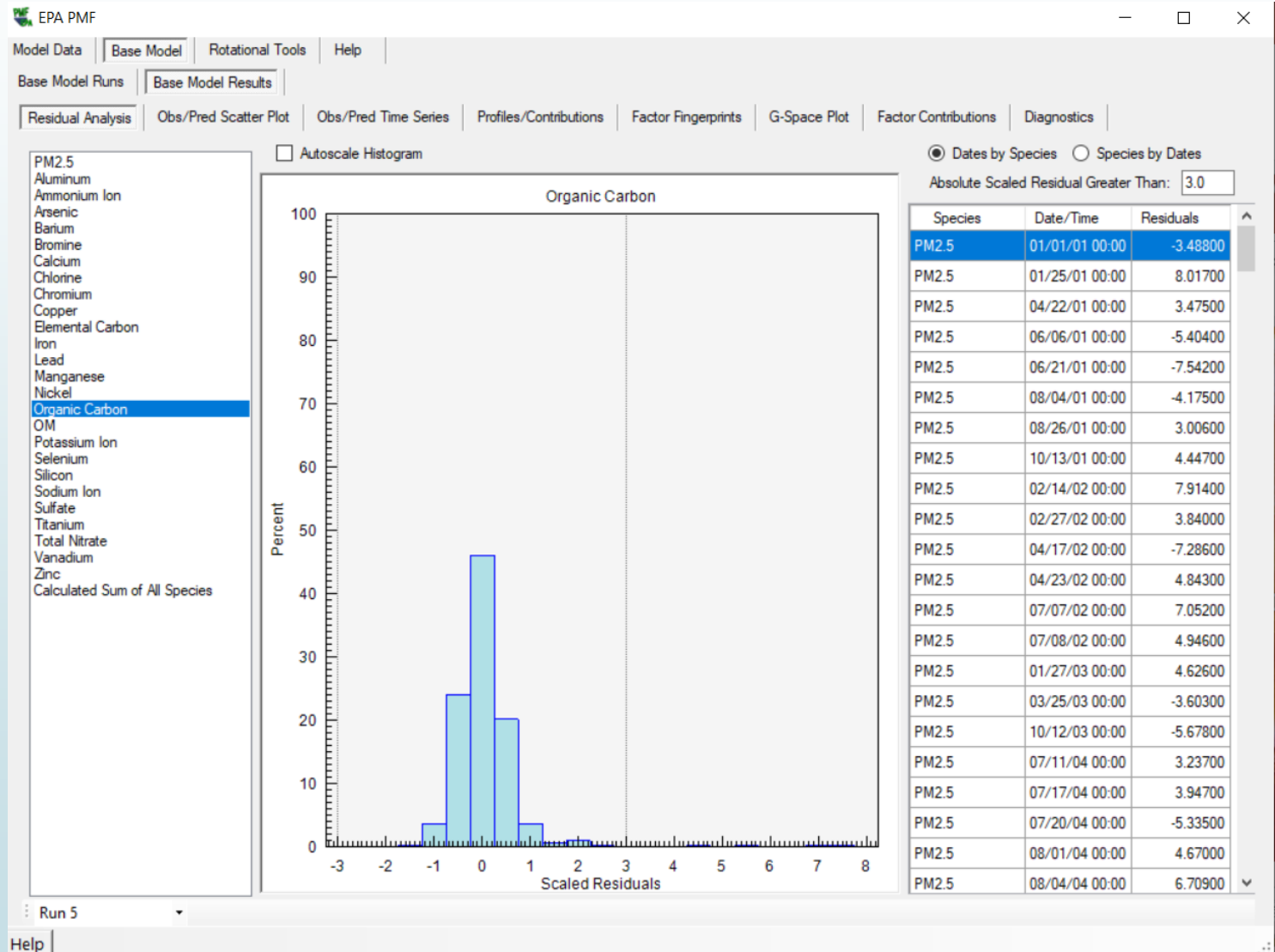
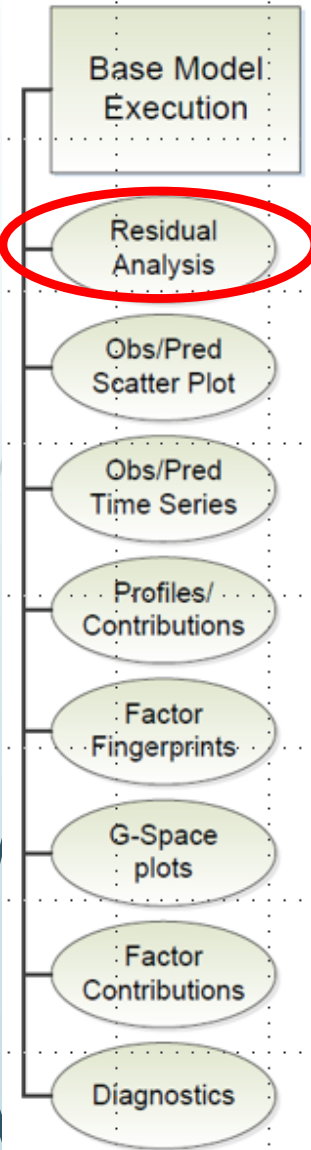
All'avvio sono consigliati 20 run. Dopo le analisi critiche e quindi rivalutati e decisi in modo definitivo i parametri del Base Model, è consigliabile utilizzare 100 run per la ricerca della soluzione 'finale'.

Per valutare l'effetto delle variazioni dei parametri è necessario utilizzare un Seed qualunque ma fisso. Per la ricerca della soluzione finale invece è opportuno utilizzare un Seed casuale.

Dopo una analisi critica dei primi risultati l'utente deciderà eventuali variazioni (incrementi o decrementi del numero di fattori).

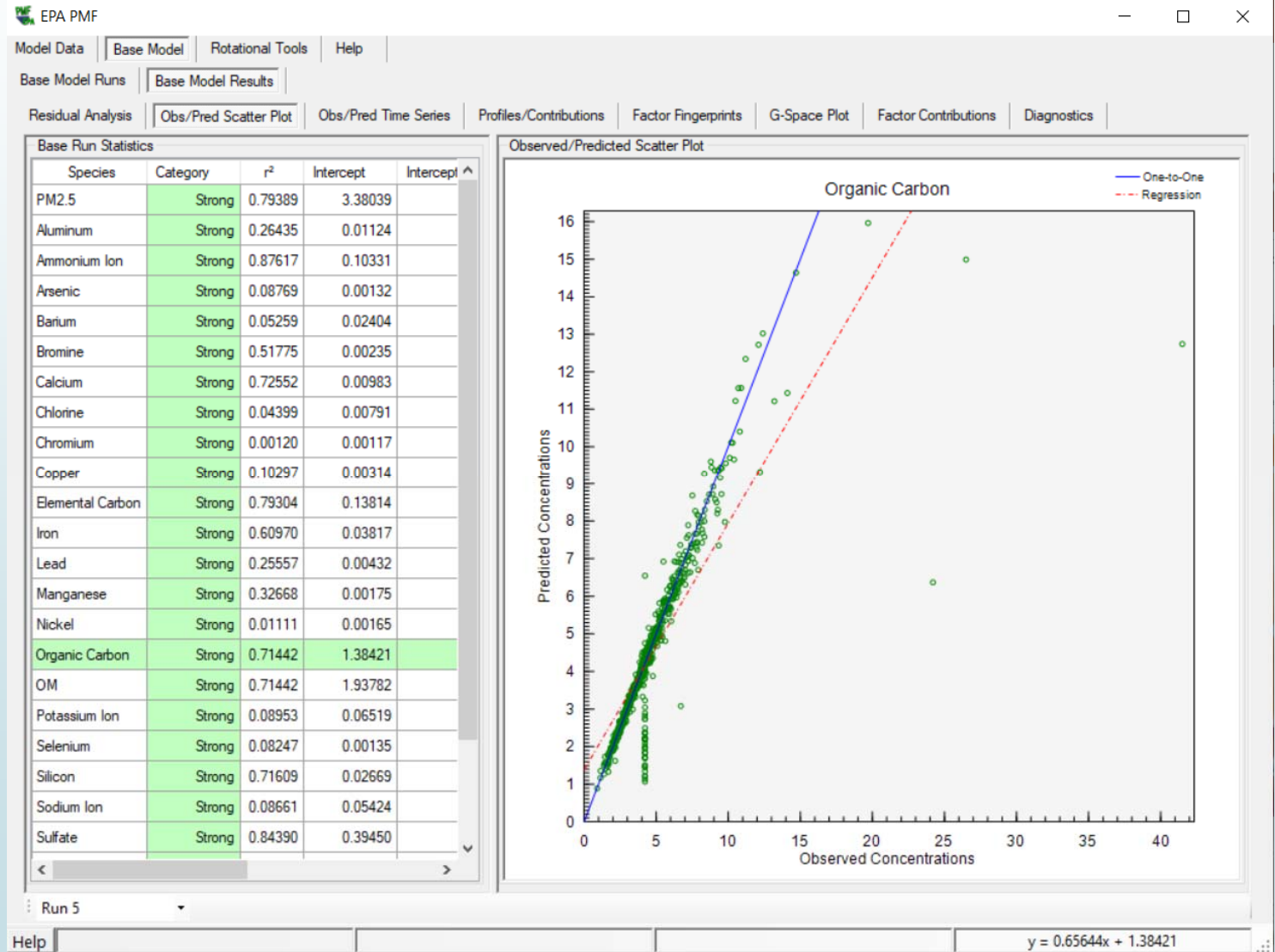
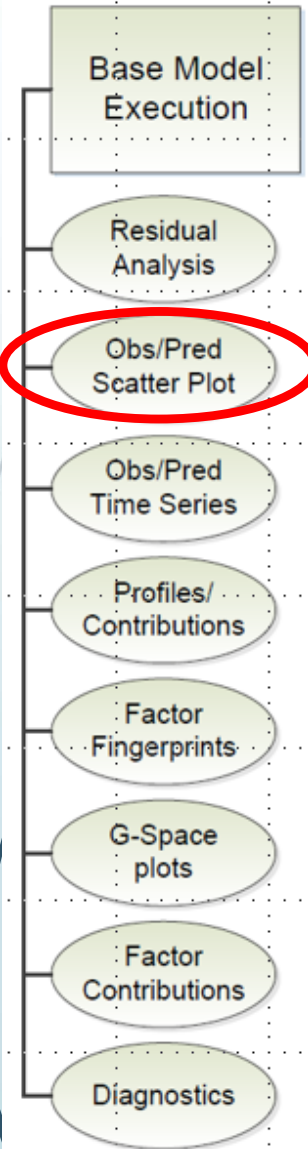
Esecuzione del modello

2



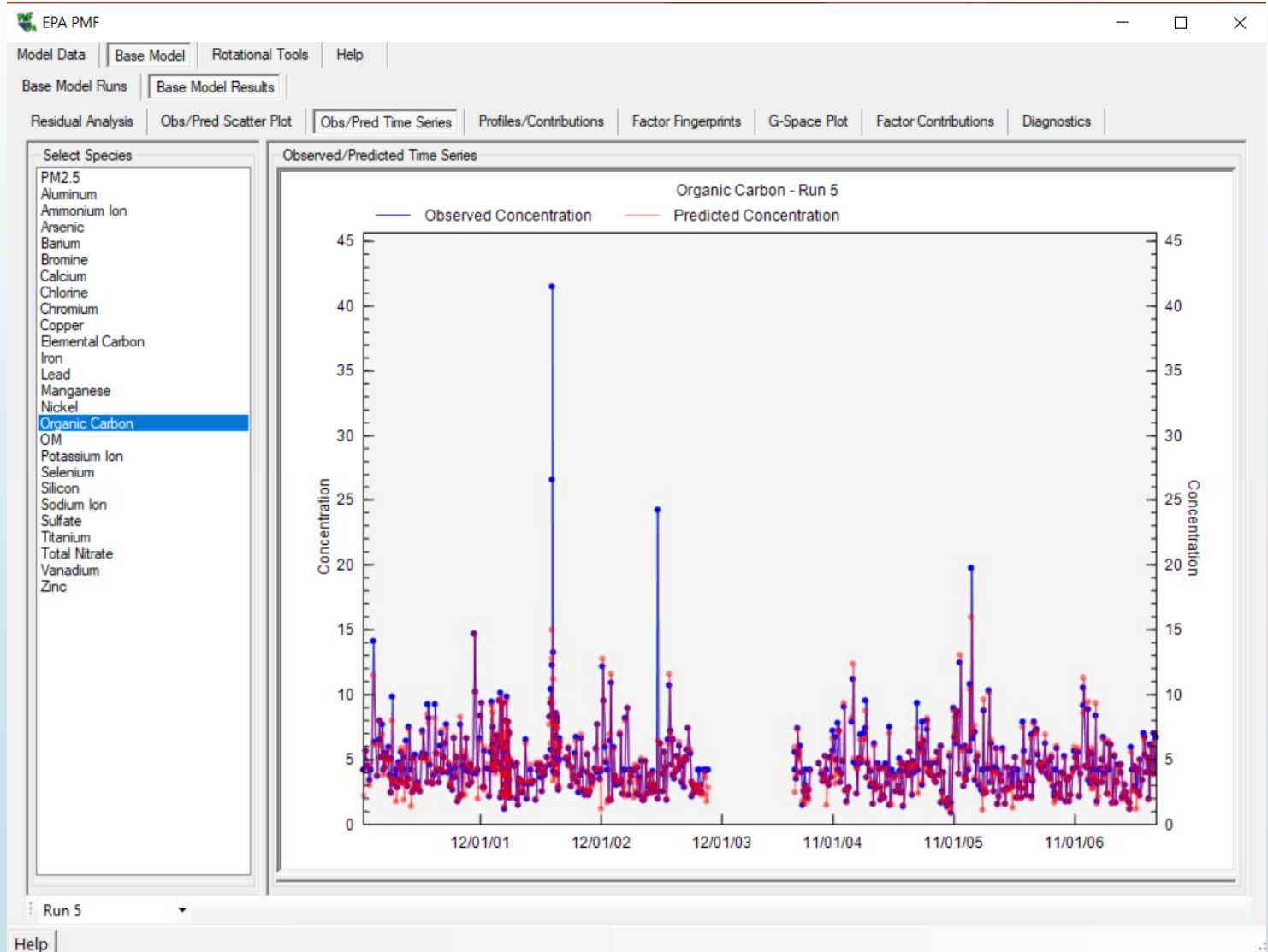
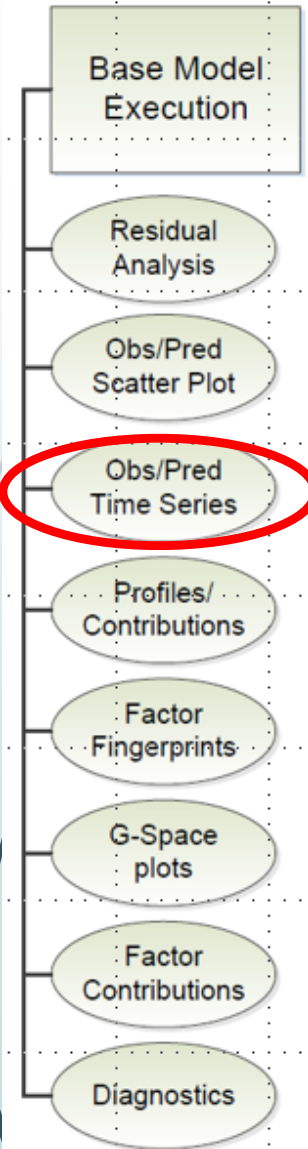
Esecuzione del modello

2



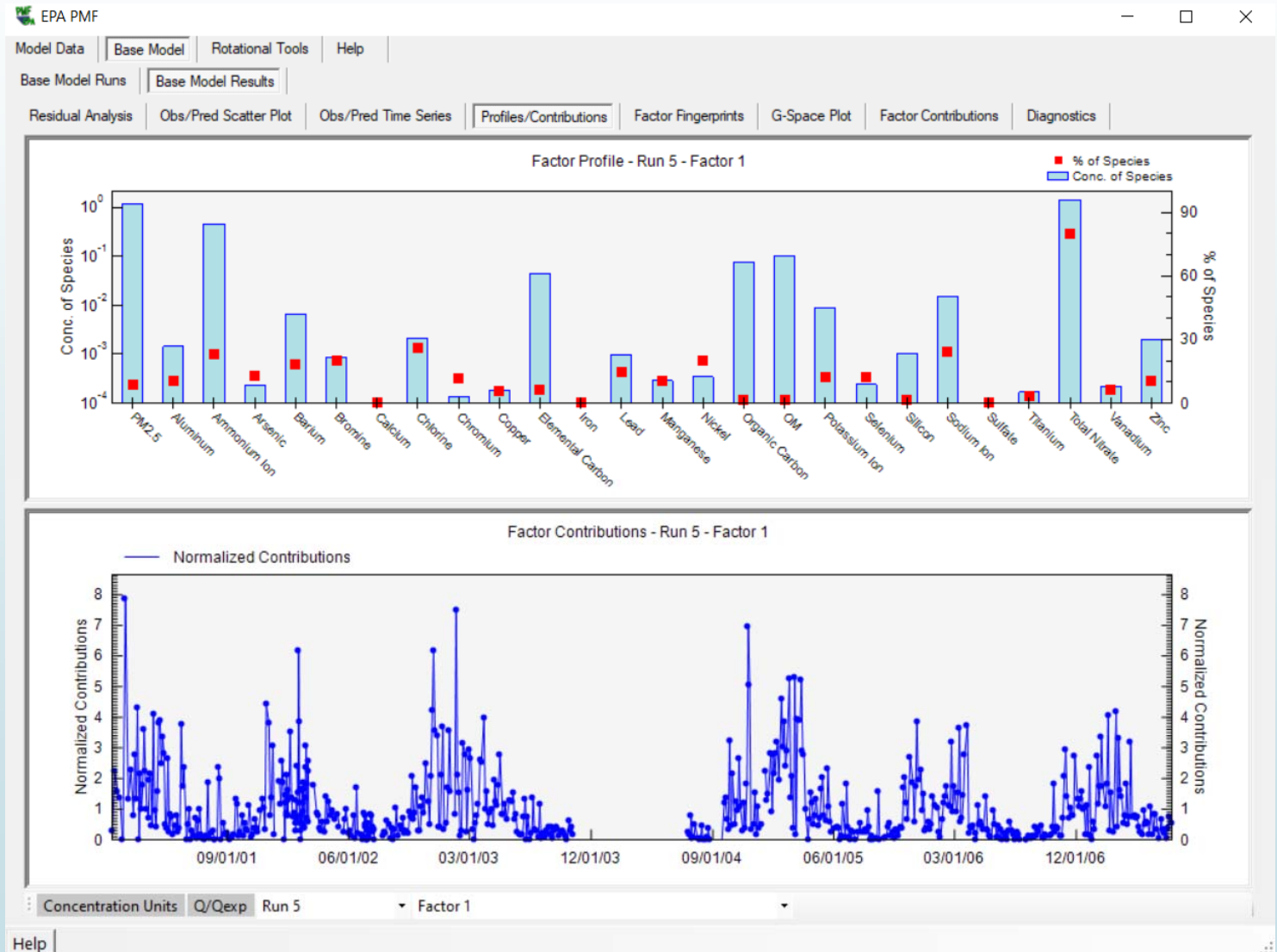
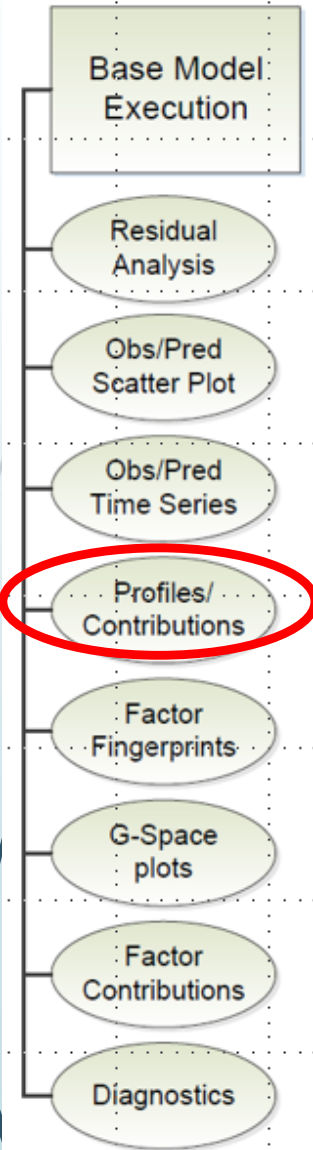
Esecuzione del modello

2



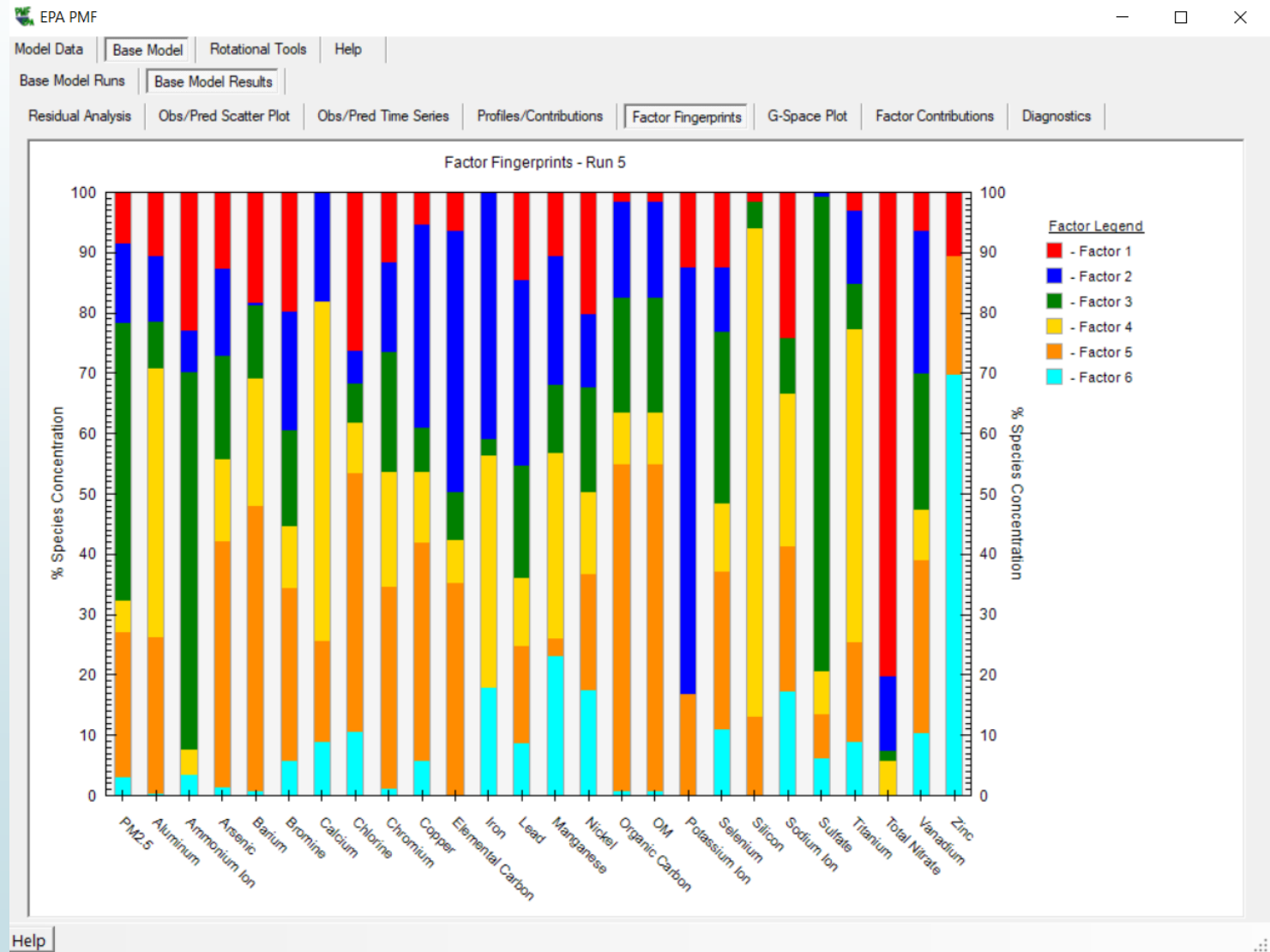
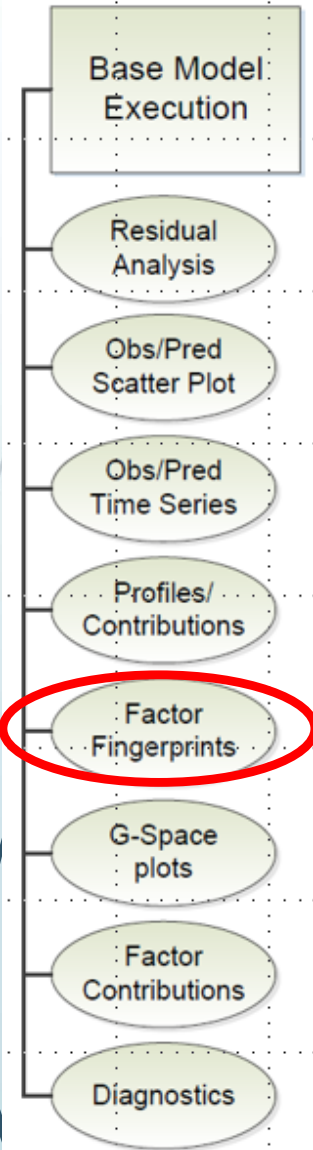
Esecuzione del modello

2



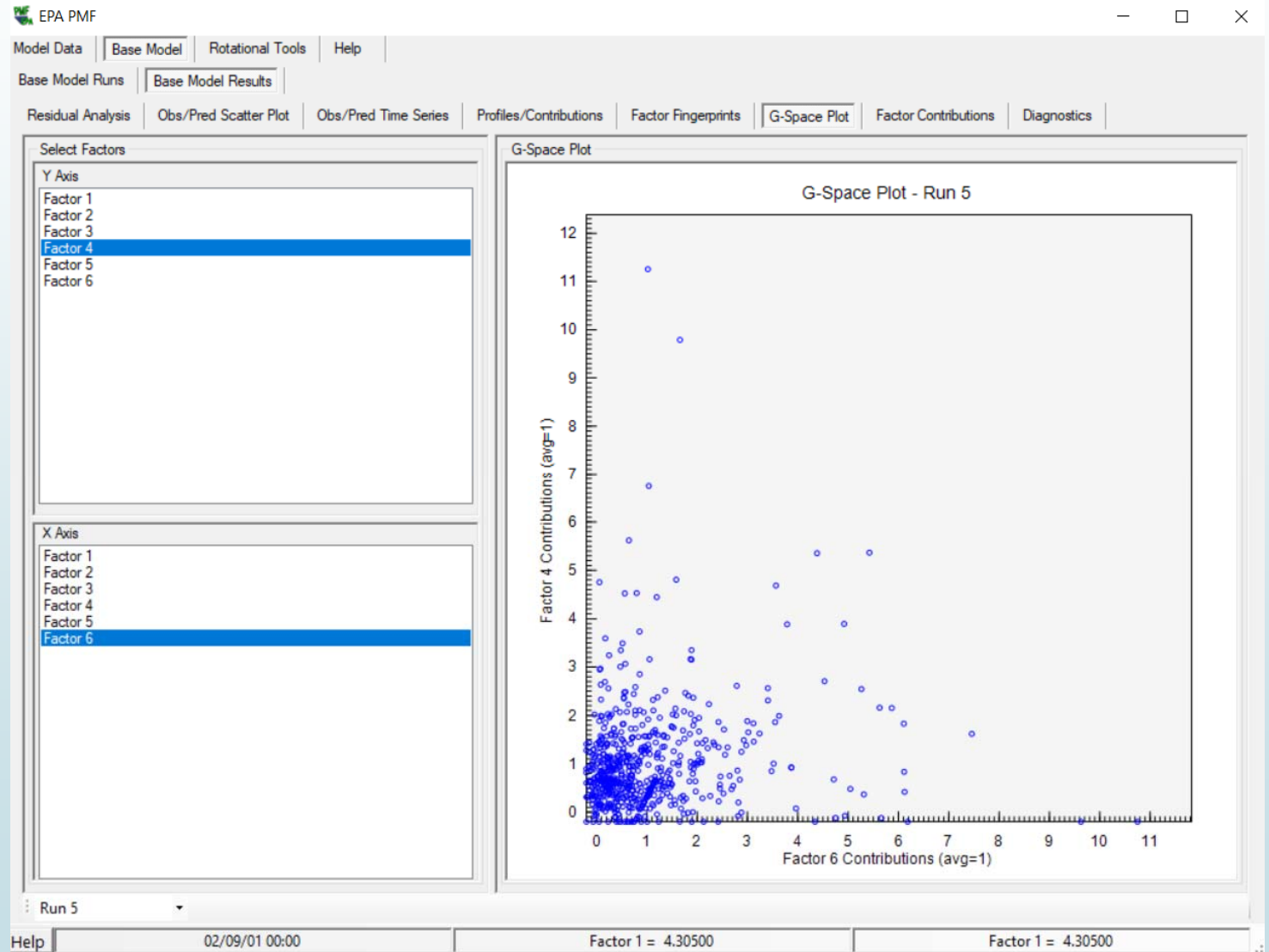
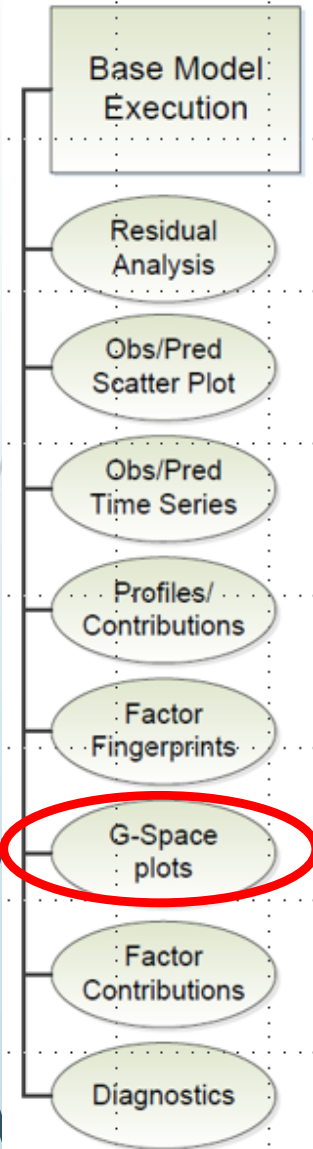
Esecuzione del modello

2



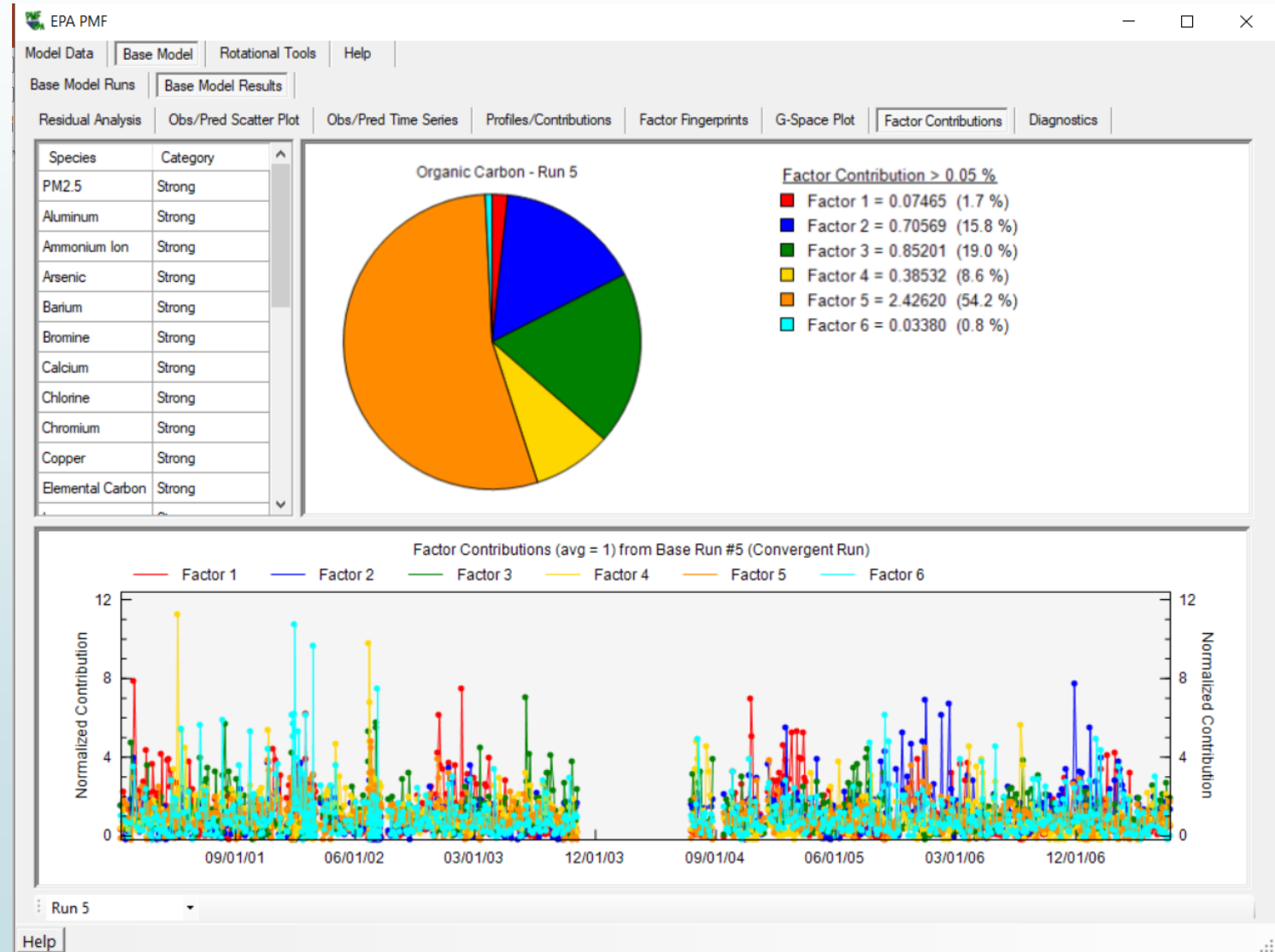
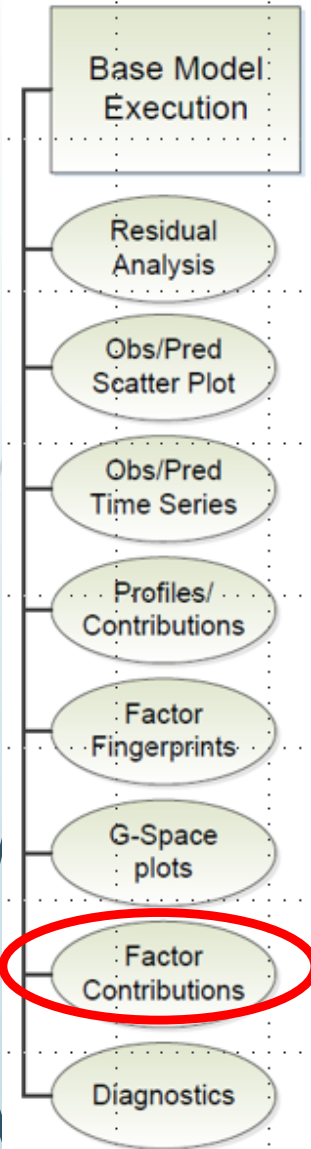
Esecuzione del modello

2



Esecuzione del modello

2





Risorse disponibili on-line

https://www.epa.gov/sites/default/files/2015-02/documents/pmf_5.0_user_guide.pdf

https://www.lcsqa.org/system/files/media/documents/eu_guide_source_apportionment_jrc_2013.pdf

https://fairmode.jrc.ec.europa.eu/document/fairmode/WG3/European%20guide%20SA_3.1_online.pdf