

Supporting Information for “Source apportionment of fine particulate matter (PM_{2.5}) personal exposure samples: findings from the Household Air Pollution Intervention Network (HAPIN) study in rural Guatemala”

Erick Mollinedo¹, Katherine A. Kearns^{1, 2}, Armistead G. Russell³, Michael Johnson², Christian L’Orange⁴, Ricardo Piedrahita², Ajay Pillarisetti⁵, Jeremy A. Sarnat⁶, John P. McCracken^{7, 8}, Lisa M. Thompson⁹, Anaité Diaz-Artiga⁸, Maggie L. Clark¹⁰, Kyle Steenland⁶, Lance A. Waller¹¹, Thomas F. Clasen⁶, Jennifer Peel¹⁰, William Checkley¹², Luke P. Naeher^{1, *} and HAPIN Investigators

¹College of Public Health, University of Georgia, Athens, GA, United States

²Berkeley Air Monitoring Group, Berkeley, CA, United States

³School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, United States

⁴Department of Mechanical Engineering, Colorado State University, Fort Collins, CO, United States

⁵Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, CA, United States

⁶Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, GA, United States

⁷Global Health Institute, College of Public Health, University of Georgia, Athens, GA, United States

⁸Center for Health Studies, Universidad del Valle de Guatemala, Guatemala City, Guatemala

⁹Nell Hodgson Woodruff School of Nursing, Emory University, Atlanta, GA, United States

¹⁰Department of Environmental and Radiological Health Sciences, Colorado State University, Fort Collins, CO, United States

¹¹Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA, United States

¹²Division of Pulmonary and Critical Care, School of Medicine, Johns Hopkins University, Baltimore, MD, United States

Corresponding Author: Luke P. Naeher (lnaeher@uga.edu), 150 East Green Street Rm 132, Athens, GA, United States

The following Supporting information contains 3 supplementary tables, 6 detailed equations, 3 supplementary figures and additional references.

Tables	3
Table S1: Summary of personal filter samples by timepoint and study arm	3
Table S2: Summary of the concentration of species used for PMF modeling, categorized by study group	4
Table S3: Summary of error estimates for PMF models with 3 to 6 sources (factors)	6
Equations	8
Equation S1	8
Equation S2	8
Equation S3	8
Equation S4	9
Equation S5	9
Equation S6	10
Figures	11
Figure S1: Q(true)/Q(exp) ratios by different number of factors	11
Figure S2: Error estimates concentrations summary for the chosen four-factor (sources) PMF model	12
Figure S3: G-Space plots of the four-factor (sources) PMF model	13
References	14

Tables

Table S1: Summary of personal filter samples by timepoint and study arm. (Baseline= <20 weeks of gestation; BL-P1: Assessment intermediate between BL and P1; P1: Assessment between 24-28 weeks of gestation; P1-P2: assessment between P1 and P2; P2: 32-36 weeks of gestation, P2-B1: approximately one month after P2)

Study Arm	Timepoint						Total
	Baseline	BL-P1	P1	P1-P2	P2	P2-B1	
Control	60	32	89	29	86	27	323
Intervention	52	33	98	29	74	20	306

Table S2: Summary of the concentration of species used for PMF modeling, categorized by study group

Statistic	Concentration ($\mu\text{g}/\text{m}^3$)		
	Baseline (N= 112)	Control (N= 263)	Intervention (N= 254)
Magnesium (Mg)			
Mean (SD)	0.16 (0.099)	0.14 (0.09)	0.1 (0.07)
Median (IQR)	0.13 (0.092 – 0.19)	0.11 (0.08 – 0.18)	0.081 (0.057 – 0.13)
Min - Max	0.013 – 0.47	0.0066 – 0.53	0.007 – 0.44
Aluminum (Al)			
Mean (SD)	0.78 (0.55)	0.63 (0.41)	0.65 (0.4)
Median (IQR)	0.64 (0.45 – 1.0)	0.56 (0.38 – 0.79)	0.53 (0.41 – 0.83)
Min - Max	0.0077 – 2.6	0.0077 – 2.2	0.008 – 2.0
Silicon (Si)			
Mean (SD)	1.1 (1.0)	0.8 (0.86)	0.76 (0.75)
Median (IQR)	0.87 (0.37 – 1.7)	0.47 (0.17 – 1.2)	0.57 (0.15 – 1.2)
Min - Max	0.0088 – 4.4	0.0088 – 4.4	0.009 – 3.8
Sulfur (S)			
Mean (SD)	0.52 (0.37)	0.41 (0.24)	0.32 (0.25)
Median (IQR)	0.4 (0.26 – 0.64)	0.36 (0.25 – 0.49)	0.24 (0.16 – 0.37)
Min - Max	0.057 – 2.0	0.058 – 1.4	0.003 – 1.4
Potassium (K)			
Mean (SD)	1.7 (1.5)	1.4 (1.2)	0.41 (0.28)
Median (IQR)	1.3 (0.8 – 2.1)	1.0 (0.64 – 1.8)	0.34 (0.23 – 0.53)
Min - Max	0.2 – 9.7	0.18 – 8.9	0.002 – 2.0
Calcium (Ca)			
Mean (SD)	0.48 (0.35)	0.37 (0.32)	0.22 (0.18)
Median (IQR)	0.39 (0.21 – 0.65)	0.3 (0.17 – 0.48)	0.18 (0.098 – 0.27)
Min - Max	0.045 – 1.5	0.026 – 2.9	0.005 – 1.2
Titanium (Ti)			
Mean (SD)	0.05 (0.042)	0.039 (0.029)	0.037 (0.032)
Median (IQR)	0.037 (0.023 – 0.069)	0.03 (0.016 – 0.058)	0.028 (0.011 – 0.05)
Min - Max	0.011 – 0.28	0.011 – 0.17	0.011 – 0.19
Manganese (Mn)			
Mean (SD)	0.025 (0.025)	0.02 (0.017)	0.016 (0.016)
Median (IQR)	0.016 (0.008 – 0.035)	0.016 (0.0068 – 0.028)	0.011 (0.0038 – 0.02)
Min - Max	0.0038 – 0.15	0.0038 – 0.11	0.004 – 0.1
Iron (Fe)			
Mean (SD)	0.4 (0.34)	0.29 (0.23)	0.29 (0.26)
Median (IQR)	0.27 (0.17 – 0.5)	0.23 (0.11 – 0.42)	0.22 (0.11 – 0.39)
Min - Max	0.034 – 2.0	0.015 – 1.1	0.006 – 1.5
Zinc (Zn)			
Mean (SD)	0.013 (0.02)	0.013 (0.013)	0.011 (0.012)
Median (IQR)	0.0046 (0.0031 – 0.017)	0.0064 (0.0031 – 0.018)	0.0046 (0.0031 – 0.015)
Min - Max	0.0031 – 0.18	0.0031 – 0.067	0.003 – 0.06
Chlorine (Cl)			
Mean (SD)	2.6 (2.9)	2.0 (2.7)	0.3 (0.38)

Median (IQR)	1.5 (0.71 – 3.4)	1.3 (0.45 – 2.3)	0.17 (0.07 – 0.35)
Min - Max	0.034 – 16.0	0.016 – 17.0	0.004 – 2.5
Sodium (Na)			
Mean (SD)	0.15 (0.094)	0.13 (0.089)	0.15 (0.098)
Median (IQR)	0.12 (0.089 – 0.2)	0.1 (0.052 – 0.17)	0.12 (0.052 – 0.19)
Min - Max	0.052 – 0.44	0.052 – 0.51	0.052 (0.52)
Black carbon (BC)			
Mean (SD)	13 (9.2)	12 (5.8)	4.8 (4.4)
Median (IQR)	11 (8.9 – 15)	11 (8.1 – 15)	2.9 (2.6 – 5.7)
Min - Max	2.6 – 96	1.7 - 53	0.95 – 40
Fine particulate matter (PM_{2.5})			
Mean (SD)	155 (135)	127 (103)	30 (22)
Median (IQR)	113 (63 – 198)	94 (58 – 176)	24 (16 – 37)
Min - Max	21 - 780	11 - 632	9.7 – 195

Table S3: Summary of error estimates for PMF models with 3 to 6 sources (factors)

Factors	Q(true)/ Q(exp)	DISP		BS		BS-DISP		
		Decrease in Q	Factor swaps at lowest dQ	Factor mapping	Unmapped samples	Decrease in Q	Accepted cases (%)	Factor swaps at lowest dQ
3	11.85	0	No swaps	Factor 1: 100% Factor 2: 100% Factor 3: 100%	None	1.08	100	No swaps
4	8.98	0.08	No swaps	Factor 1: 100% Factor 2: 100% Factor 3: 100% Factor 4: 52%	None	1.2	96	4 swaps in factor 2 4 swaps in factor 3
5	6.45	0.05	No swaps	Factor 1: 99% Factor 2: 100% Factor 3: 100% Factor 4: 100% Factor 5: 92%	None	51.67	99	No swaps
6	5.69	11.17	No swaps	Factor 1: 100% Factor 2: 82% Factor 3: 100% Factor 4: 100% Factor 5: 100% Factor 6: 99%	None	29.59	84	1 swap in factor 1 1 swap in factor 2 6 swaps in factor 3 7 swaps in factor 4

Explanation from Table S3^{1,2}:

- DISP: Displacement error. Conducts repeated fits in the model by perturbing (dispersing) each variable. Measures the effect of random errors and rotational ambiguity.
- BS: Bootstrap error. Conducts data resampling, this measures the effects from random errors and partially, the effects of rotational ambiguity.
- BS-DISP: A method that combines Bootstrap data errors and Displacement error rotational ambiguity.
- Q(true)/Q(expected) ratio: The ratio of the goodness-of-fit parameter calculated using all the observations 'Q(true)' and the observations without analytical difficulties or anomalous values 'Q(expected)'. The lowest value for this parameter means the model has the highest resolution (in this case, the 6-factor model).
- Decrease in Q (DISP and BS-DISP): Nominal decrease of the parameter 'Q' after the DISP analysis. If a high value is reported, it means that the base solution was probably not the best (in this case, the 6-factor model), and another solution could be the most appropriate.

- Factor swaps at lowest dQ (DISP and BS-DISP): Describe if the solution by the model is considered as the universal solution. If swaps are present, it means that the factor can be interpreted as another one from the model solution (in this case, the 4-factor model shows 4 swaps between factor 2 and 3, which means that this model may not easily distinguish factor 2 from 3).
- Factor mapping (BS): A parameter to determine the percentage of resamples by each factor that were mapped against the original factors determined by the model. The closest to 100% shows a more stable solution (less uncertainty).

Equations

The average 24-hr PM_{2.5} concentrations were estimated using the following equation:

$$[PM_{2.5}] = \frac{M}{V} \quad (\text{Equation S1})$$

where $[PM_{2.5}]$ is the PM_{2.5} concentration (µg/m³). M is the mass deposition, which is the difference between the average post- and pre- weights, corrected by the median of field blank filters. The limit of detection (LoD) of M was calculated as 3 times the standard deviation of the mass deposition, and values below LoD were replaced with LoD/2^(0.5)³. V is the average volume of the sampler, estimated as the sampling duration multiplied by the average air flow rate.

Black carbon concentrations were calculated using the following equation:

$$[BC] = \frac{\ln(I_f/I_0) * A/V * 10^6}{\sigma} \quad (\text{Equation S2})$$

where $[BC]$ is the black carbon concentration (µg/m³). I_f and I_0 are the average post- and pre-sample collection attenuations from the transmissometer. A is the effective sampling area of the filter, V is the volume of the sampler (as described above) and σ is the attenuation cross section value (13.7 µg/m³). The σ value comes from Garland et al.⁴, which describes PTFE filters with samples of similar sources found in the present study.

The uncertainties for each inorganic element were calculated using the following equation:

$$S_{XRF} = \sqrt{\left(\frac{A}{V}\right)^2 * S_d^2 + \left(\frac{D}{V}\right)^2 * S_a^2 + \left(\frac{-D * A}{V^2}\right)^2 * (S_v * V)^2} \quad (\text{Equation S3})$$

Where:

S_{XRF} = Estimated uncertainty of the inorganic species samples

A = Area (cm²)

V = Volume (m³)

D = Chemical species Deposition (µg/cm²)

S_d = Adjusted XRF Uncertainty (µg/cm²)

S_a = Area uncertainty = (0.05 cm²)

S_v = Volume uncertainty = (0.05 m³)

The uncertainties of BC samples were determined using the following equation:

$$S_{BC} = \frac{\sigma}{100} * \sqrt{\left(\ln\left(\frac{pre}{post}\right) * 100\right)^2 * SA^2 + A^2 * Satn^2} \quad (\text{Equation S4})$$

Where:

S_{BC} = Estimated uncertainty for BC samples

$\sigma/100$ = Attenuation cross-section value from Garland et al (2017) ($\sigma= 1370$), divided by 100

$\ln(pre/post) * 100$ = Attenuation, estimated as the Natural logarithm of the intensity of light of a post-sampled filter divided by a pre-sampled filter by 100

A = Area in m^2

SA = Area uncertainty in m^2

$Satn$ = Uncertainty of the attenuation

The uncertainties of the $PM_{2.5}$ samples were estimated using the following equation:

$$S_{PM2.5} = \sqrt{\left(\frac{1}{V}\right)^2 * s_m^2 + \left(\frac{-M}{V^2}\right)^2 * \left(\frac{S_v}{V}\right)^2} \quad (\text{Equation S5})$$

Where:

$S_{PM2.5}$ = Estimated uncertainty of the $PM_{2.5}$ samples

V = Volume (m^3)

M = Mass Deposition (μg)

S_m = Mass uncertainty = (2.0)

S_v = Volume uncertainty = (0.05 m^3)

The objective function Q from PMF is defined as:

$$Q = \sum_{i=1}^n \sum_{j=1}^m \left[\frac{X_{ij} - \sum_{k=1}^p g_{ik} f_{kj}}{u_{ij}} \right]^2 \quad (\text{Equation S6})$$

Where, Q is the critical parameter for PMF, X_{ij} is the concentration of species j on sample i , p is the number of factors (sources), f_{kj} is the concentration of species j in the source profile k , g_{ik} is the relative mass concentration of source k contributing to the sample i , e_{ij} is the residual for each species and sample and u_{ij} is the uncertainty estimate in the species j measured in the sample i .

Figures

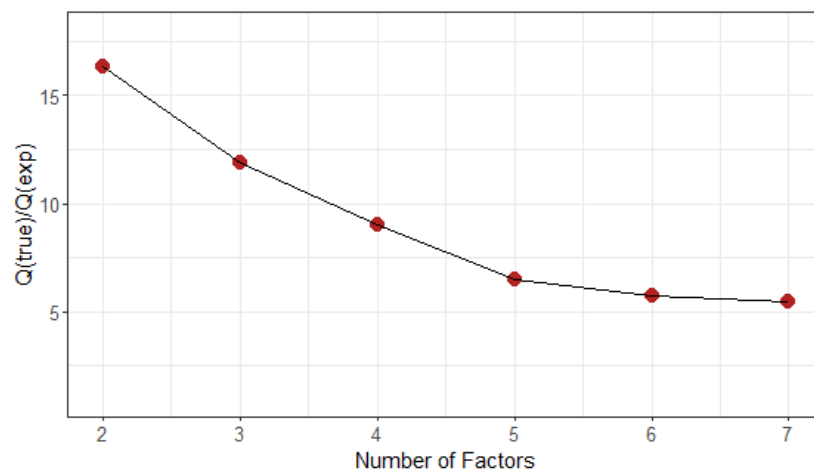


Figure S1: $Q(\text{true})/Q(\text{exp})$ ratios by different number of factors.

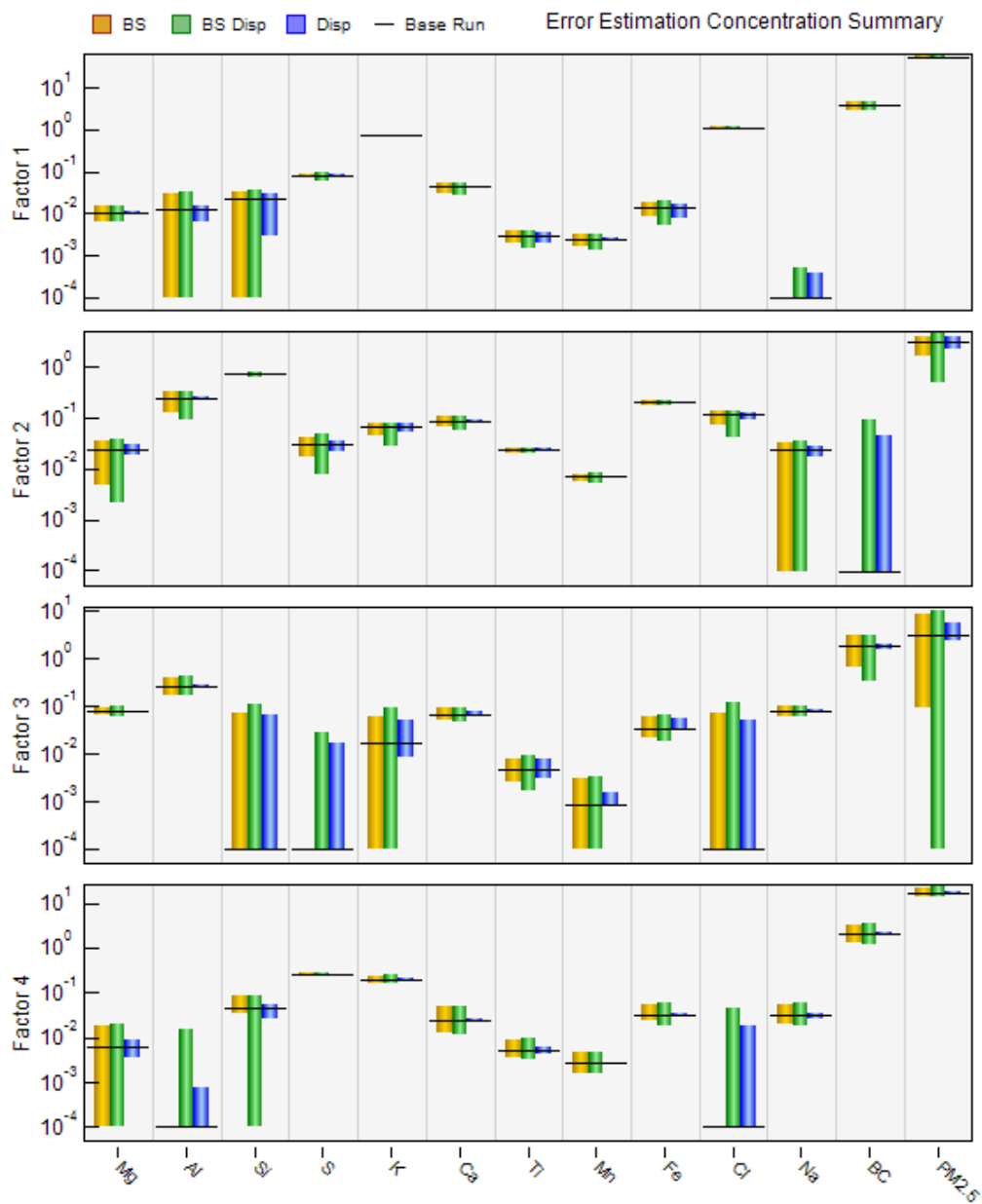


Figure S2: Error estimates concentrations summary for the chosen four-factor (sources) PMF model. (Output directly from the EPA PMF 5.0 software)

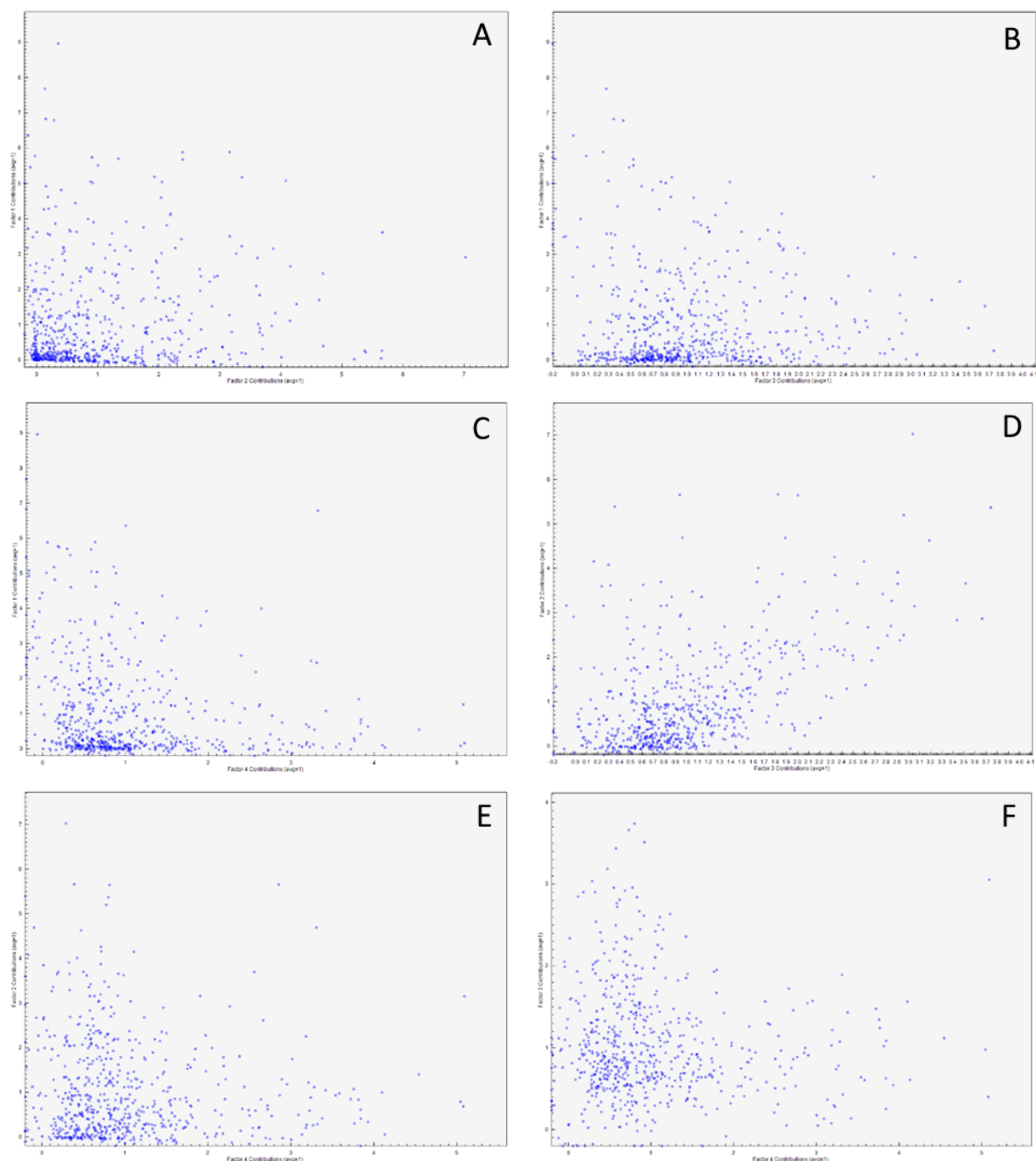


Figure S3: G-Space plots of the four-factor (sources) PMF model. (A= Source 1 vs Source 2; B= Source 1 vs Source 3; C= Source 1 vs Source 4; D= Source 2 vs Source 3; E= Source 2 vs Source 4; F= Source 5 vs Source 6) (Output directly from the EPA PMF 5.0 software).

The G-Space plots provide a graphical representation when plotting one factor against another⁵. A more stable solution (less rotational ambiguity) is observed when the observations are distributed close to zero in both X and Y axes (Figures S3: A, B, C, E and F)

References:

1. Paatero, P., Eberly, S., Brown, S. G. & Norris, G. A. Methods for estimating uncertainty in factor analytic solutions. *Atmospheric Meas. Tech.* **7**, 781–797 (2014).
2. 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. *Sci. Total Environ.* **518–519**, 626–635 (2015).
3. Johnson, M. *et al.* Exposure Contrasts of Pregnant Women during the Household Air Pollution Intervention Network Randomized Controlled Trial. *Environ. Health Perspect.* **130**, 097005 (2022).
4. Garland, C. *et al.* Black carbon cookstove emissions: A field assessment of 19 stove/fuel combinations. *Atmos. Environ.* **169**, 140–149 (2017).
5. Norris, G., Duvall, R., Brown, S. & Bai, S. Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide.