Elemental characterization of personal filter samples from the HAPIN trial for Source Apportionment

Project: Part 2

Erick E. Mollinedo

2024-02-23

# 1. Introduction

## 1.1 Background

Household air pollution (HAP), due to the combustion of solid fuels for cooking and heating, is one of the most prominent environmental health issues in low- and middle-income countries (LMICs).1 HAP has been listed as the second environmental health risk factor worldwide, just after outdoor air pollution.2 Exposure to HAP has been linked to more than 2.31 million deaths in 2019 and it is associated with adverse cardiovascular and respiratory health outcomes.3,4 Particulate matter with an aerodynamic diameter less than 2.5 micrometers (PM2.5) is one of the major air pollutants chemically and physically composed of crustal material, black carbon (BC), organic compounds, metals and others. Toxicity from PM2.5 depends, among many factors, on the chemical composition of the particles which heavily depend on the sources of emissions.5,6

The identification of point sources of pollution is important for the development of strategies to improve air quality. Source apportionment (SA) is a methodology to reconstruct the impact of emissions from different sources, a method applied to different pollutants such as particulate matter.7 The most used SA methods includes receptor models, an example of this is the positive matrix factorization (PMF). One advantage of PMF is that it does not require the source profiles prior to analysis and there is no limit on the number of sources, however, it is required the knowledge of potential source profiles. Another important piece before conducting source apportionment is to correctly process the data (which includes concentrations and uncertainties), and to explore associations to try to identify the potential sources. Just a few studies about SA in LMICs have been published, therefore it is important to characterize chemical composition and potential sources of air pollutants in these settings.8-10

The Household Air Pollution Intervention Network (HAPIN) trial was a randomized controlled trial where a liquefied petroleum gas (LPG) stove was delivered as an intervention to half of the pregnant women participants, meanwhile the other half remained cooking using their biomass stove. The intervention was evaluated in terms of reduction of exposures and specific health outcomes by comparing the two study groups (Control and Intervention). Baseline measurements were taken to the participants, and then followed-up for about 18 months, conducting exposure measurements three times during pregnancy and three times after the birth of the child. A description of the HAPIN trial and the methods for exposure sampling can be found elsewhere.11,12 A source apportionment pilot study from HAPIN was conducted in 2022. This study was conducted in 64 and 59 personal exposure filter samples from the pilot phase in Guatemala and Rwanda, respectively. Twenty-two chemical species were analyzed in all samples using X-ray fluorescence (XRF), a non-destructive spectrometry analytical method, but only 12 species were detected at both study sites in comparable levels. Based on the chemical species detected, four potential sources of PM2.5 were identified in both sites. One study caveat is the small sample size and limited assessment of chemical composition and sources between study arms, so further studies to complement the findings are needed.13

## 1.2 Aims and Hypothesis questions

One of the objectives of this project is to create a clean output that consists in a file with adjusted concentrations for all samples and each elemental species, and another file with the adjusted uncertainties for the same samples and elemental species. These outputs will be used as inputs for the EPA PMF 5.0 software to try to determine the potential sources that contributed to these samples. However, the questions I plan to answer with the data analysis of this project is:

1. What elements are correlated among each other from these filter samples?
2. Is there a difference in concentrations if we categorize the samples by type of fuel and by study arm?
3. What exposures could explain the variations in the concentrations of the most significant elemental species?

# 2. Methods

## 2.1 Data collection

The main exposure data comes from filter samples of personal exposure measurements in pregnant women participants from the HAPIN trial, in Guatemala. In total, six hundred and forty eight sample filters are available, where eighteen correspond to field blanks. From the HAPIN trial study design, in total, six exposure measurements (visits) were conducted during pregnancy in around 800 participants, but for this scenario, only 648 filter samples were randomly selected for chemical species characterization. Other exposures where collected as categorical variables using standardized questionnaires administered in each visit. These exposures capture some of the sources of exposure such as tobacco smoke, trash burning, use on incense or coil, kerosene and others.

## 2.2 Concentration estimates

The average 24-hour PM2.5 concentrations were calculated using the following equation:

Where, M is the mass deposition of particles in the filter and V is the volume of air the sampler used to collect the particles.

The black carbon 24-hour concentrations were calculated using the following equation:

Where, If is the post-attenuation and I0 is the pre-attenuation, A is the area of the filter, V is the volume of air the sampler used to collect the particles and is the mass absorption coefficient (a constant).

The concentration of the chemical species (elements) on the filters was determined using X-Ray Fluorescence (XRF), and the uncertainties were estimated also based on the XRF instrument.

## 2.3 Concentration and Uncertainties data processing

First, all chemical species where more than 50% of the samples are below the detection limit (LOD) were filtered out, leaving a total of 10 chemical species. For these remaining species, the concentrations were adjusted based on the blank filters, and the values below LOD were replaced as their corresponding LOD divided by the square root of 2. Finally, the concentrations were transformed to micrograms per cubic meter (ug/m3). Also, the uncertainties were transformed to ug/m3 to match the units of the concentrations. In addition, the black carbon concentrations were added to add another chemical species to the analysis.

The uncertainties were processed by, first removing all the chemical species that had more than 50% of their samples below the detection limit, as seen above. Then, the uncertainties were adjusted based on equations derived from the law of propagation of uncertainty. In brief, the equations consider the individual uncertainties from each one of the variables, in this case the XRF instrument uncertainty, the area of the filter, volume, and the attenuation (in case of the Black carbon measurements).

## 2.4 Statistical analysis

The correlation (r2) between chemical species will be determined using either individual paired correlations or by a correlation plot matrix. Additional plots, such as boxplots, to visualize and compare how the concentrations differ by type of fuel and by study arm. These differences will be assessed for statistical significance. After exploring the distribution that best fits the dataste, the most appropriate model will be implemented to explore what exposure variables are mostly associated with the concentrations of the most significant elemental species.

# 3. Preliminary Results

## 3.1 Description of data

1. Concentrations:

As seen in the following glimpse from the cleaned concentration dataframe, it contains the ID of the 630 sample filters, along with the concentrations of 11 chemical species (Mg, Al, Si, S, K, Ca, Ti, Mn, Fe, Zn, and BC), and to which Arm and Fueltype they belong.

concentration <- read\_rds(here("data", "processed-data", "concentration.rds"))  
  
glimpse(concentration)

Rows: 630  
Columns: 14  
$ filter\_id <chr> "3M50038", "3M50039", "3M50040", "3M50046", "3M50080", "3M50…  
$ Mg <dbl> 0.0692, 0.1172, 0.1216, 0.1992, 0.1087, 0.0300, 0.1708, 0.18…  
$ Al <dbl> 0.1938, 0.4610, 0.0077, 1.6462, 0.1857, 0.0077, 0.0077, 0.79…  
$ Si <dbl> 0.0088, 0.7402, 0.2425, 1.8990, 0.8719, 0.0088, 0.0731, 2.13…  
$ S <dbl> 0.1313, 0.2661, 0.3313, 0.2360, 0.4874, 0.1635, 0.2891, 0.28…  
$ K <dbl> 0.3547, 1.0150, 1.9132, 1.8146, 0.7701, 0.8615, 1.9753, 1.71…  
$ Ca <dbl> 0.1180, 0.6348, 0.6317, 0.5638, 1.0556, 0.0376, 0.3737, 0.69…  
$ Ti <dbl> 0.0111, 0.0323, 0.0276, 0.1105, 0.0276, 0.0111, 0.0161, 0.10…  
$ Mn <dbl> 0.0038, 0.0161, 0.0133, 0.0241, 0.0038, 0.0068, 0.0161, 0.03…  
$ Fe <dbl> 0.0827, 0.2321, 0.1395, 0.8825, 0.1456, 0.0539, 0.1554, 0.63…  
$ Zn <dbl> 0.0031, 0.0046, 0.0064, 0.0031, 0.0031, 0.0031, 0.0082, 0.01…  
$ BC <dbl> 2.584503, 13.579366, 95.612828, 10.034089, 10.659048, 10.848…  
$ arm <chr> "Control", "Intervention", "Control", "Control", "Interventi…  
$ fueltype <chr> "Biomass", "Biomass", "Biomass", "Biomass", "Biomass", "Biom…

1. Uncertainties:

The cleaned uncertainty dataframe contains the uncertainty of the 11 chemical species. It contains the same name of variables as the concentration dataframe.

uncertainty <- read\_rds(here("data", "processed-data", "uncertainty.rds"))  
  
glimpse(uncertainty)

Rows: 630  
Columns: 14  
$ filter\_id <chr> "3M50038", "3M50039", "3M50040", "3M50046", "3M50080", "3M50…  
$ Mg <dbl> 0.0068, 0.0106, 0.0112, 0.0172, 0.0101, 0.0047, 0.0147, 0.01…  
$ Al <dbl> 0.0164, 0.0377, 0.0090, 0.1335, 0.0161, 0.0090, 0.0090, 0.06…  
$ Si <dbl> 0.0104, 0.0601, 0.0202, 0.1539, 0.0708, 0.0104, 0.0074, 0.17…  
$ S <dbl> 0.0107, 0.0216, 0.0269, 0.0192, 0.0395, 0.0133, 0.0234, 0.02…  
$ K <dbl> 0.0302, 0.0837, 0.1564, 0.1483, 0.0640, 0.0712, 0.1614, 0.13…  
$ Ca <dbl> 0.0118, 0.0529, 0.0529, 0.0473, 0.0868, 0.0076, 0.0322, 0.05…  
$ Ti <dbl> 0.0132, 0.0097, 0.0085, 0.0165, 0.0078, 0.0132, 0.0056, 0.01…  
$ Mn <dbl> 0.0046, 0.0056, 0.0043, 0.0077, 0.0046, 0.0032, 0.0049, 0.00…  
$ Fe <dbl> 0.0121, 0.0239, 0.0164, 0.0765, 0.0169, 0.0087, 0.0176, 0.05…  
$ Zn <dbl> 0.0037, 0.0088, 0.0075, 0.0037, 0.0037, 0.0037, 0.0073, 0.01…  
$ BC <dbl> 3.1826, 1.0593, 3.8767, 1.6588, 2.0040, 1.0923, 3.4024, 0.17…  
$ arm <chr> "Control", "Intervention", "Control", "Control", "Interventi…  
$ fueltype <chr> "Biomass", "Biomass", "Biomass", "Biomass", "Biomass", "Biom…

1. Other variables from the HAPIN study:

The cleaned hapin\_samples dataframe contains information of the PM2.5, BC and CO (carbon monoxide) concentrations, besides other important recorded exposures, as seen from the glimpse below. It also contains the filter IDs, the arm and fueltype.

hapin\_samples <- read\_rds(here("data", "processed-data", "hapin\_samples.rds"))  
  
glimpse(hapin\_samples)

Rows: 630  
Columns: 16  
$ arm <chr> "Control", "Intervention", "Control", "Control", "Control"…  
$ filter\_id <chr> "3M50148", "3M51092", "3M50107", "3M50888", "3M50821", "3M…  
$ stove <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Y…  
$ smoke <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"…  
$ coil <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"…  
$ trash <chr> "Yes", "No", "No", "No", "No", "No", "No", "No", "No", "No…  
$ kerosene <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"…  
$ incense <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"…  
$ generator <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"…  
$ smoky <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"…  
$ crop <chr> "No", "No", "No", "No", "No", "No", "No", "No", "No", "No"…  
$ stove\_other <chr> "Yes", "No", "Yes", "No", "No", "No", "No", "No", "No", "Y…  
$ pm25 <dbl> 112.513867, 40.814201, 105.409475, 85.149929, 27.093514, 1…  
$ bc <dbl> 15.976181, 8.164598, 10.848641, 9.697128, 5.112407, 2.5547…  
$ co <dbl> 0.52393802, 0.01826375, 0.04087828, 0.73995536, 0.40697640…  
$ fueltype <chr> "Biomass", "LPG", "Biomass", "Biomass", "Biomass", "LPG", …

## 3.2 Exploratory data analysis

### 3.2.1 Chemical species concentrations

The concentration of the chemical species are distributed in different ranges depending on their type. For example, the highest concentrations were determined to be for Black carbon (BC), with values from 1 to 100 ug/m3. The remaining chemical species have lower concentrations, ranging from less than 0.1 ug/m3 to 10 ug/m3. Other noticeable patterns is that some of the species were found in higher concentrations for personal samples from participants that used biomass stoves, compared to LPG stoves ([Figure 1](#fig-concentrations)), such as BC, K, and Ca, and similar patterns were observed when categorized by intervention arm (Supplementary material). The following step will be to determine which of these chemical species have statistically significant differences among both comparison groups.

In terms of correlations, Spearman correlation was estimated between all the chemical species and [Figure 2](#fig-correlations), shows the associations highlighting the R2 value. It is observed that the highest correlation is observed between Si and Fe, Ti-Fe, and Al-Si. These values could indicate that these species come from the same source, so it will probably be easier to identify using the PMF software after the conclusion of this project.

It is important to remember that the concentrations of the chemical species were recalculated as referenced in the data\_cleaning\_code.qmd file, alongside with the uncertainties. A summary statistics table of the uncertainties can be found in the data\_cleaning\_code.qmd code.

|  |
| --- |
| Figure 1: Chemical species concentrations by type of fuel |

|  |
| --- |
| Figure 2: Correlation plot for chemical species |

### 3.2.2 Other Exposure variables

From the hapin\_samples dataframe, it was observed that there are some missing values for most of the exposure variables according to the Missing variables plot ([Figure 3](#fig-missing)). The variable with more missing values (>30) was the CO concentrations. Given that this variable had the highest number of missing concentrations, it may not be ideal to include it as part of the final analysis. Other missing values are from the recorded exposures such as kerosene, trash burning, stove use, external stoves (‘stove\_other’), and others. It was also observed that there is one observation without BC concentration. This observation could be dropped out since the PMF analysis after the conclusion of this project requires all the concentrations and uncertainties to be available (No NA/missing values).

The bar plot ([Figure 4](#fig-categorical)), shows that most of the participants recorded little exposures, with the exception of stove use at home (‘stove’), stoves from other homes (‘stove\_other’) and trash burning (‘trash’). Having not so many recorded exposures other than the previously mentioned, could make the analysis difficult because of sample size or any other statistical adjustments will have to occur before applying a regression model analysis approach.

[Figure 5](#fig-pm) shows the distribution of PM2.5 concentrations by type of fuel. It is observed that the concentrations do not follow a normal distribution and it seems like, as expected, the LPG stove users samples have lower concentrations than the biomass stove users. The similar trends are observed for the CO and BC concentrations (Supplementary material).

|  |
| --- |
| Figure 3: Plot for missing values for the hapin\_samples dataframe |

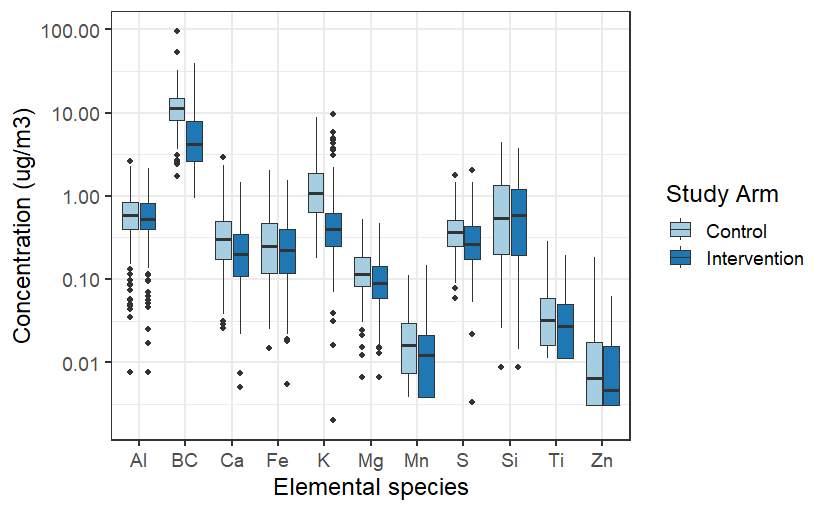
|  |
| --- |
| Figure 4: Distribution of responses by the Categorical variables |

|  |
| --- |
| Figure 5: PM2.5 Concentrations |

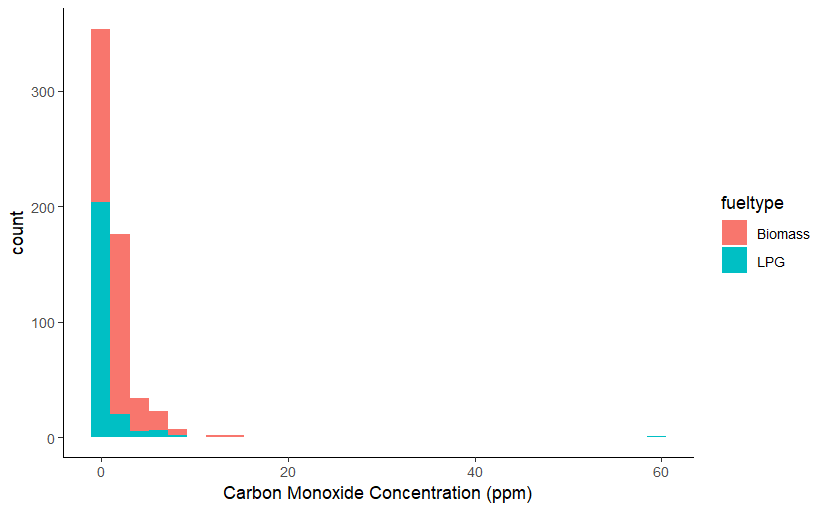
# 4. References

1. Gordon, S. B. et al. Respiratory risks from household air pollution in low and middle income countries. The Lancet Respiratory Medicine 2, 823–860 (2014).
2. GBD. Global Burden of Disease Study 2019 (GBD 2019) Results. http://ghdx.healthdata.org/gbd-results-tool (2020).
3. Murray, C. J. L. et al. Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. The Lancet 396, 1223–1249 (2020).
4. Pope, C. A. & Dockery, D. W. Health Effects of Fine Particulate Air Pollution: Lines that Connect. Journal of the Air & Waste Management Association 56, 709–742 (2006).
5. Clasen, T. et al. Design and Rationale of the HAPIN Study: A Multicountry Randomized Controlled Trial to Assess the Effect of Liquefied Petroleum Gas Stove and Continuous Fuel Distribution. Environ Health Perspect 128, 047008 (2020).
6. Mohammed, M. O. A. et al. Potential Toxicological and Cardiopulmonary Effects of PM2.5 Exposure and Related Mortality: Findings of Recent Studies Published during 2003-2013. Biomed Environ Sci.
7. Viana, M. et al. Source apportionment of particulate matter in Europe: A review of methods and results. Journal of Aerosol Science 39, 827–849 (2008).
8. Zhou, Z. et al. Chemical Characterization and Source Apportionment of Household Fine Particulate Matter in Rural, Peri-urban, and Urban West Africa. Environ. Sci. Technol. 48, 1343–1351 (2014).
9. Sharma, S. K. et al. Source Apportionment of PM2.5 in Delhi, India Using PMF Model. Bull Environ Contam Toxicol 97, 286–293 (2016).
10. Martínez-Cinco, M., Santos-Guzmán, J. & Mejía-Velázquez, G. Source apportionment of PM 2.5 for supporting control strategies in the Monterrey Metropolitan Area, Mexico. Journal of the Air & Waste Management Association 66, 631–642 (2016).
11. Clasen, T., Checkley, W., Peel, J. L., Balakrishnan, K., McCracken, J. P., Rosa, G., Thompson, L. M., Barr, D. B., Clark, M. L., Johnson, M. A., Waller, L. A., Jaacks, L. M., Steenland, K., Miranda, J. J., Chang, H. H., Kim, D.-Y., McCollum, E. D., Davila-Roman, V. G., Papageorghiou, A., … HAPIN Investigators. Design and Rationale of the HAPIN Study: A Multicountry Randomized Controlled Trial to Assess the Effect of Liquefied Petroleum Gas Stove and Continuous Fuel Distribution. Environmental Health Perspectives, 128(4), 047008. (2020).
12. Johnson, M. A., Steenland, K., Piedrahita, R., Clark, M. L., Pillarisetti, A., Balakrishnan, K., Peel, J. L., Naeher, L. P., Liao, J., Wilson, D., Sarnat, J., Underhill, L. J., Burrowes, V., McCracken, J. P., Rosa, G., Rosenthal, J., Sambandam, S., de Leon, O., Kirby, M. A., … HAPIN Investigators. Air Pollutant Exposure and Stove Use Assessment Methods for the Household Air Pollution Intervention Network (HAPIN) Trial. Environmental Health Perspectives, 128(4), 047009. (2020).
13. Kearns, K. A. Exposures to household air pollution in women across four diverse countries as part of a multi-country cookstove intervention trial. (2022).

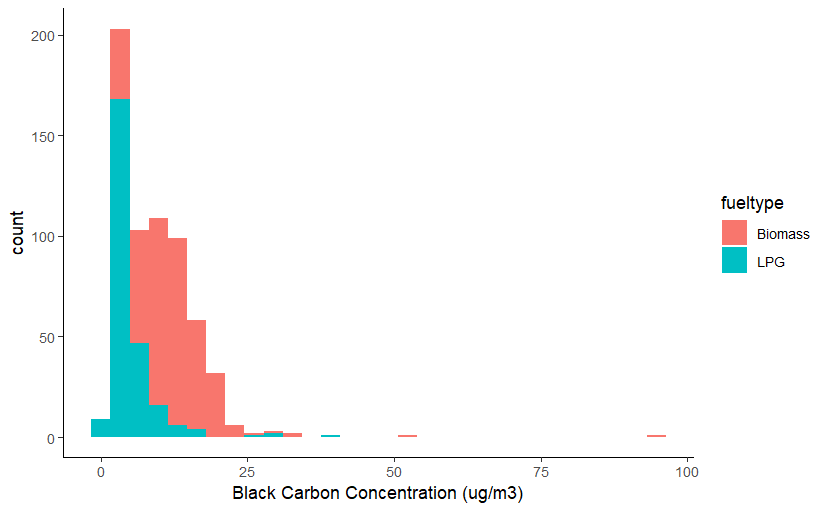
# 5. Supplementary material



Chemical species concentrations by study arm



Carbon monoxide (CO) Concentrations



Black carbon (BC) Concentrations