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

Project: Part 1

Erick E. Mollinedo

2024-02-01

# 1. Introduction

## 1.1 General Background Information

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 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.2,3 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.4,5

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.6 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.7-9

## 1.2 Description of data and data source

The data comes from filter samples of PM2.5 from personal exposure measurements conducted in pregnant women from the Household Air Pollution Intervention Network [(HAPIN)](https://www.hapintrial.org/) trial, in Guatemala. In brief, this is 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). The filter samples represent the exposures to PM2.5 in a 24-hour period. In total, six hundred and forty eight sample filters are available, where eighteen correspond to field blanks. For these filters, the elemental composition was determined using X-Ray Fluorescence (XRF) to determine the concentrations of 24 elements.

1. Concentrations:

As seen in the following glimpse from the one of the spreadsheets that contains the concentrations data, each filter sample has its unique ID in Filter. There are other 24 variables that list the concentrations of each element. The concentrations are listed in micrograms per square centimeter (ug/cm2). Also, each filter is categorized in either two types (Personal sample or Blank) under Type.

concentration <- read\_excel(here("data", "raw-data", "conc\_unc\_hapingt\_emollinedo.xlsx"), sheet = "Concentration")  
  
#Have a 'glimpse' of the data  
glimpse(concentration)

Rows: 648  
Columns: 26  
$ Filter <chr> "3M53155", "3M53146", "3M53145", "3M53144", "3M53129", "3M53137…  
$ Type <chr> "Personal sample", "Personal sample", "Personal sample", "Perso…  
$ Mg <chr> "5.0500000000000003E-2", "7.6399999999999996E-2", "0.2369999999…  
$ Al <dbl> 0.2664, 0.3422, 0.7431, 0.9458, 0.6668, 0.8040, 0.8879, 1.4259,…  
$ Si <chr> "0.38869999999999999", "0.66539999999999999", "1.96459999999999…  
$ S <dbl> 0.14694, 0.28000, 0.23270, 0.21852, 0.32850, 0.27250, 0.24223, …  
$ K <chr> "0.1278", "1.03", "0.66180000000000005", "0.29330000000000001",…  
$ Ca <chr> "7.8299999999999995E-2", "0.29089999999999999", "0.3604", "0.31…  
$ Ti <chr> "2.9000000000000001E-2", "1.89E-2", "5.04E-2", "7.4399999999999…  
$ Cr <chr> "0", "[0.0014]", "[0.0022]", "[0.0007]", "[0.0007]", "[0.0014]"…  
$ Mn <chr> "9.4999999999999998E-3", "1.03E-2", "1.9800000000000002E-2", "2…  
$ Fe <chr> "0.15490000000000001", "0.21199999999999999", "0.45700000000000…  
$ Ni <chr> "[0.0007]", "0", "[0.0015]", "3.3999999999999998E-3", "[0.0030]…  
$ Cu <chr> "1.2999999999999999E-2", "0", "0", "1.09E-2", "1.73999999999999…  
$ Zn <chr> "[0.0070]", "[0.0100]", "[0.0100]", "8.9999999999999993E-3", "2…  
$ Ga <chr> "[0.0033]", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "…  
$ As <chr> "0", "0", "0", "[0.0040]", "0", "0", "[0.0024]", "0", "[0.0016]…  
$ Se <chr> "0", "0", "[0.0008]", "[0.0008]", "0", "0", "7.7000000000000002…  
$ Cd <chr> "0", "0", "[0.03]", "0", "[0.03]", "0", "0", "[0.03]", "[0.12]"…  
$ In <chr> "[0.05]", "[0.13]", "0", "0", "[0.17]", "[0.13]", "0", "0", "0"…  
$ Sn <chr> "[0.05]", "[0.11]", "[0.06]", "[0.01]", "[0.05]", "[0.05]", "0"…  
$ Te <chr> "[0.23]", "0", "[0.23]", "0", "0", "[0.87]", "0", "0", "[0.39]"…  
$ I <chr> "0", "0", "0", "0", "[0.84]", "0", "[0.22]", "[0.28]", "0", "0"…  
$ Pb <chr> "[0.0064]", "0", "0", "0", "[0.0032]", "0", "0", "0", "0", "0",…  
$ Cl <chr> "4.4499999999999998E-2", "0.84099999999999997", "0.726999999999…  
$ Na <chr> "7.2499999999999995E-2", "0.1002", "0.22720000000000001", "0.23…

#Check which values are under the `Type` variable  
unique(concentration$Type)

[1] "Personal sample" "Blank"

1. Uncertainties:

The uncertainties are stored on a different tab from the same excel file, but they match to the concentrations for each filter sample (rows) and each element (columns). These values are also stored in ug/cm2.

uncertainty <- read\_excel(here("data", "raw-data", "conc\_unc\_hapingt\_emollinedo.xlsx"), sheet = "Uncertainty")  
  
#Produce a summary of uncertainties by each element  
summary(uncertainty)

Filter Type Mg Al   
 Length:648 Length:648 Min. :0.000000 Min. :0.000000   
 Class :character Class :character 1st Qu.:0.002200 1st Qu.:0.002800   
 Mode :character Mode :character Median :0.002600 Median :0.003100   
 Mean :0.002622 Mean :0.003095   
 3rd Qu.:0.003000 3rd Qu.:0.003500   
 Max. :0.004500 Max. :0.005500   
 Si S K Ca   
 Min. :0.000000 Min. :0.0000450 Min. :0.000000 Min. :0.000000   
 1st Qu.:0.002000 1st Qu.:0.0006600 1st Qu.:0.005000 1st Qu.:0.004100   
 Median :0.002500 Median :0.0008200 Median :0.006800 Median :0.005000   
 Mean :0.002455 Mean :0.0008541 Mean :0.007731 Mean :0.005226   
 3rd Qu.:0.003100 3rd Qu.:0.0009900 3rd Qu.:0.009700 3rd Qu.:0.006200   
 Max. :0.005200 Max. :0.0020000 Max. :0.026000 Max. :0.013000   
 Ti Cr Mn Fe   
 Min. :0.000000 Min. :0.000000 Min. :0.000000 Min. :0.001300   
 1st Qu.:0.004575 1st Qu.:0.000000 1st Qu.:0.002600 1st Qu.:0.006700   
 Median :0.007100 Median :0.000000 Median :0.004300 Median :0.008450   
 Mean :0.006463 Mean :0.001528 Mean :0.003869 Mean :0.009043   
 3rd Qu.:0.008600 3rd Qu.:0.003700 3rd Qu.:0.005400 3rd Qu.:0.011000   
 Max. :0.012000 Max. :0.007900 Max. :0.008700 Max. :0.023000   
 Ni Cu Zn Ga   
 Min. :0.000000 Min. :0.000000 Min. :0.00000 Min. :0.000000   
 1st Qu.:0.000000 1st Qu.:0.004600 1st Qu.:0.00440 1st Qu.:0.000000   
 Median :0.001300 Median :0.006000 Median :0.00580 Median :0.002400   
 Mean :0.001609 Mean :0.006719 Mean :0.00651 Mean :0.003214   
 3rd Qu.:0.003600 3rd Qu.:0.011000 3rd Qu.:0.01000 3rd Qu.:0.007600   
 Max. :0.006700 Max. :0.015000 Max. :0.01300 Max. :0.011000   
 As Se Cd In   
 Min. :0.000000 Min. :0.000000 Min. :0.00000 Min. :0.0000   
 1st Qu.:0.000000 1st Qu.:0.000000 1st Qu.:0.00000 1st Qu.:0.0000   
 Median :0.000000 Median :0.000000 Median :0.00000 Median :0.0000   
 Mean :0.002297 Mean :0.002467 Mean :0.06757 Mean :0.1104   
 3rd Qu.:0.002900 3rd Qu.:0.003300 3rd Qu.:0.14000 3rd Qu.:0.2425   
 Max. :0.021000 Max. :0.013000 Max. :0.24000 Max. :0.3400   
 Sn Te I Pb   
 Min. :0.0000 Min. :0.0000 Min. :0.0000 Min. :0.000000   
 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.0000 1st Qu.:0.000000   
 Median :0.0000 Median :0.3900 Median :0.8100 Median :0.000000   
 Mean :0.1436 Mean :0.3998 Mean :0.8602 Mean :0.004691   
 3rd Qu.:0.2325 3rd Qu.:0.9400 3rd Qu.:2.0000 3rd Qu.:0.006200   
 Max. :0.4800 Max. :1.2000 Max. :2.4000 Max. :0.036000   
 Cl Na   
 Min. :0.0000 Min. :0.000000   
 1st Qu.:0.0085 1st Qu.:0.003700   
 Median :0.0140 Median :0.004300   
 Mean :0.0176 Mean :0.004285   
 3rd Qu.:0.0230 3rd Qu.:0.004800   
 Max. :0.0740 Max. :0.006800

1. Other variables from the HAPIN study:

Another source of data is a .csv file that contains over 1000 variables collected from the study, associated to a single filter sample or a single participant (depending on the perspective). I am only interested in exploring some of the variables such as “Treatment” and “Timepoint” to categorize the filters according to study arm or type of fuel. “PM2.5”, “Black Carbon” and “Carbon monoxide” to have a reference of the concentrations of the three air pollutants measured. And other variables that record exposures such as trash burning, vehicle emissions, tobacco smoke among others.

## 1.3 Questions/Hypotheses to be addressed

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, there are certain questions that I plan to answer:

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?

Answering these questions will also be helpful before performing the PMF analysis using the EPA software, to check patterns and associations between elemental species.

# 2. Methods

## 2.1 Data processing

First, the concentration values will be cleaned by removing unnecessary character such as brackets, and by standardizing in the same numeric notation. Then, all elements where more than 50% of the samples are below the detection limit (DL) will be filtered out. The detection limit for each elemental species is listed in the same excel spreadsheet that contains the raw concentrations and uncertainties. For the remaining elemental species, the concentrations will be adjusted based on the blank filters, and the values below DL will be replaced as their corresponding DL divided by the square root of 2. Finally, the concentrations will be transformed to micrograms per cubic meter (ug/m3) operating the concentration by the filter area, divided by the sample volume. Also, the uncertainties will be transformed to ug/m3 to match the units of the concentrations. In addition, the black carbon concentrations will be added to add another species to the analysis.

The uncertainties will be processed by first, removing all the elemental species that had more than 50% of their samples below the detection limit, as seen above. Then, the uncertainties will be 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).

## 2.2 Statistical analysis

The correlation (r2) between elemental 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. 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. 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).
3. 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).
4. 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).
5. 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.
6. 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).
7. 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).
8. Sharma, S. K. et al. Source Apportionment of PM2.5 in Delhi, India Using PMF Model. Bull Environ Contam Toxicol 97, 286–293 (2016).
9. 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).