EPA Source Apportionment

Fall 2025 Capstone

Deeya Datta, Prateek Mishra, & George Whittington
Client: Deron Smith, EPA

Table of contents

01

Introduction

Our Project Topic and Client

Source Apportionment

Source apportionment is a "collection of techniques to provide information regarding how much a source (usually a generalized category) contributes to the overall pollutant concentration at receptor (usually a monitoring site) (Rizzo, 2010).

The overall goal is to extract latent factors from the data. We can then interpret these factors to identify specific source profiles.

Deron Smith

Mr. Smith is a UGA alumni with a B.S. in Physics and a M.S. in Computer Science from GaTech. He is now a researcher and software developer for the Environmental Protection Agency

Recent projects include an open source tool called the Environmental Source Apportionment Toolkit (esat), and is currently working on a graphical user interface for the toolkit.

02

Data

Provided Datasets and Research Questions

The Datasets

Baltimore

focuses on fine particulate matter (PM2.5) chemical composition (n=630)

Baton Rouge

focuses on volatile organic compounds (VOCs) rather than particulate matter (n=307)

St. Louis

focuses on fine particulate matter (PM2.5) chemical composition (n=418)

Variables Across Cities

City	Category	Count	Specific Species Included
Baltimore	Trace Elements	17	Al, As, Ba, Br, Ca, Cl, Cr, Cu, Fe, Mn, Ni, Pb, Se, Si, Ti, V, Zn
	Ions	5	NH_4^+, K^+, Na^+, SO_4^{2-}, NO_3^-
	Carbon & Mass	6	EC, OC, OM, PM2.5
Baton Rouge	VOCs	41	BTEX, Alkanes, Alkenes & Isoprene
St. Louis	Trace Elements	8	Cd, Cu, Fe, Mn, Ni, Pb, Se, Zn
	Ions	2	SO_4^{2-}, NO_3^-
	Carbon & Mass	3	EC, OC, PM2.5

Considerations & Challenges

Temporal Resolution & Irregularity

Baltimore:

- Irregular daily measurements with frequent multi-day gaps.

Baton Rouge:

- Sparse hourly sampling (mostly 03:00 or 06:00) rather than continuous monitoring.

St. Louis:

- "Chunked" hourly data (e.g., 00:00–23:00) but with several intraday gaps (missing afternoon blocks).

Considerations & Challenges

Completeness & Detection Limits

Data Integrity:

- There were no missing values found across any of the three datasets.

Signal-to-Noise:

- However, we identified observations where the instrument uncertainty exceeded the measured concentration.

Handling Low Values:

- These "high-noise" points are treated as Below Detection Limit (BDL) but preserved using log-transformation.

Research Questions

Input Data Analysis

How can the input data be effectively analyzed and pre-processed to identify key features?

Source Apportionment Modeling

Can we calculate latent Source Profiles and quantify their Contributions using a standard apportionment model?

Uncertainty Integration

How can we utilize the paired measurement uncertainty into the model to refine our estimates and assess potential error?

Model Evaluation

How can we evaluate the results of multiple models (metrics, plots) to ensure stability?

Environmental Interpretation

What physical conclusions can we make about the real-world pollution sources based on the model outputs?

03

EDA

Exploratory Data Analysis

Gaps Between Extreme Values

Baltimore

Species	mean	max	max_mean_ratio
PM2.5	15.57	76.30	4.90
OM	6.60	58.10	8.80
Organic Carbon	4.71	41.50	8.80
Sulfate	4.83	30.20	6.25
Total Nitrate	1.83	12.60	6.88

Baton Rouge

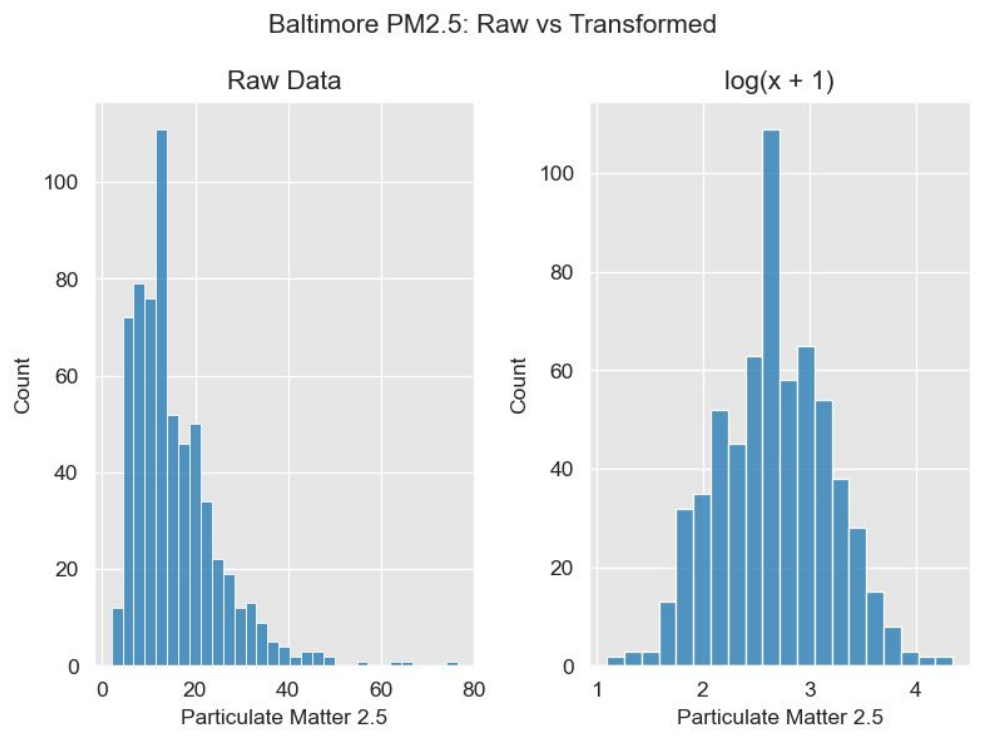
Species	mean	max	max_mean_ratio
TNMOC	199.90	708.47	3.54
Unidentified	63.20	385.18	6.09
Isopentane	16.40	125.50	7.65
N-Hexane	8.54	92.56	10.84
Propane	15.78	70.96	4.50

St. Louis

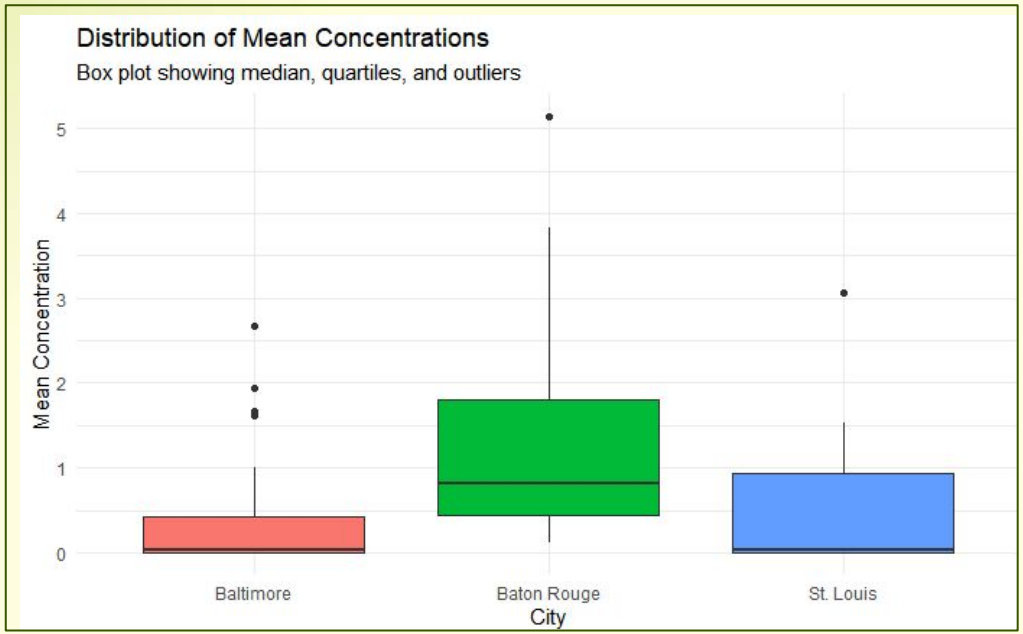
Species	mean	max	max_mean_ratio
Mass	22.95	107.00	4.66
OC	4.77	71.20	14.91
SO4	3.84	16.90	4.40
NO3	1.82	10.30	5.65
EC	0.77	6.38	8.28

log(x + 1) Transformation

A log(x + 1) is applied to the data to try and bring the data to be more symmetric and to also preserve the non-negative status



Concentration Distributions



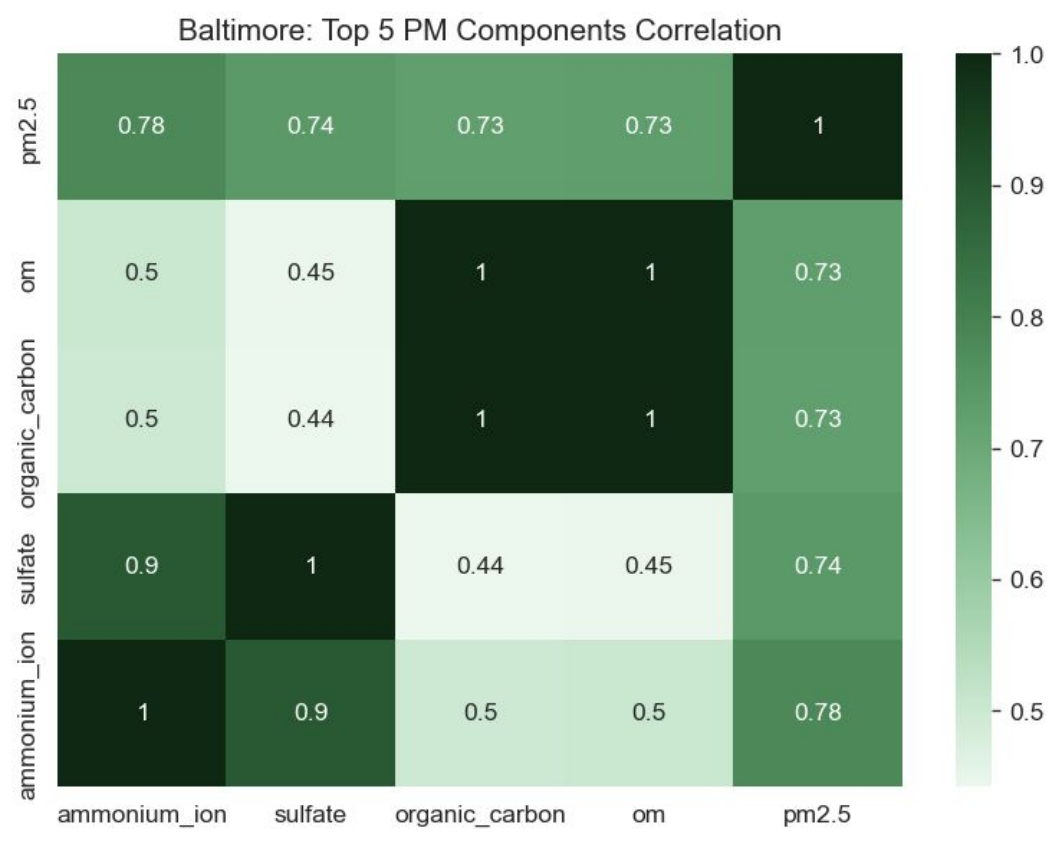

```
==========================================
TOP 5 SPECIES BY MEAN CONCENTRATION
==========================================

BATON ROUGE (VOCs in ppbC):
  1. tnmoc                          5.1488
  2. unidentified                   3.8296
  3. propane                        2.6409
  4. ethane                         2.5839
  5. isopentane                     2.5333

ST. LOUIS (µg/m³):
  1. pm2.5                          3.0585
  2. organic_carbon                 1.5261
  3. sulfate                        1.4111
  4. total_nitrate                  0.9303
  5. elemental_carbon               0.5190

BALTIMORE (µg/m³):
  1. pm2.5                          2.6672
  2. om                             1.9389
  3. organic_carbon                 1.6625
  4. sulfate                        1.6125
  5. ammonium_ion                   1.0076
```

Top Correlations



Organic Carbon and Organic Matter have a perfect correlation of 1.0. This is because OM isn't independently measured, it's simply OC multiplied by a conversion factor to account for the hydrogen and oxygen attached to the carbon.

04

Future Work

Goals For Next Semester

How do we plan to implement a model?

What is Positive Matrix Factorization?

PMF separates pollutants into sources and their contributions.

Source Contributions (G): how much each source contributed on each day

Source Profiles (F): the chemical fingerprint of each source

X=GF+E where X is observed and E are residual errors

Why is it a good fit?

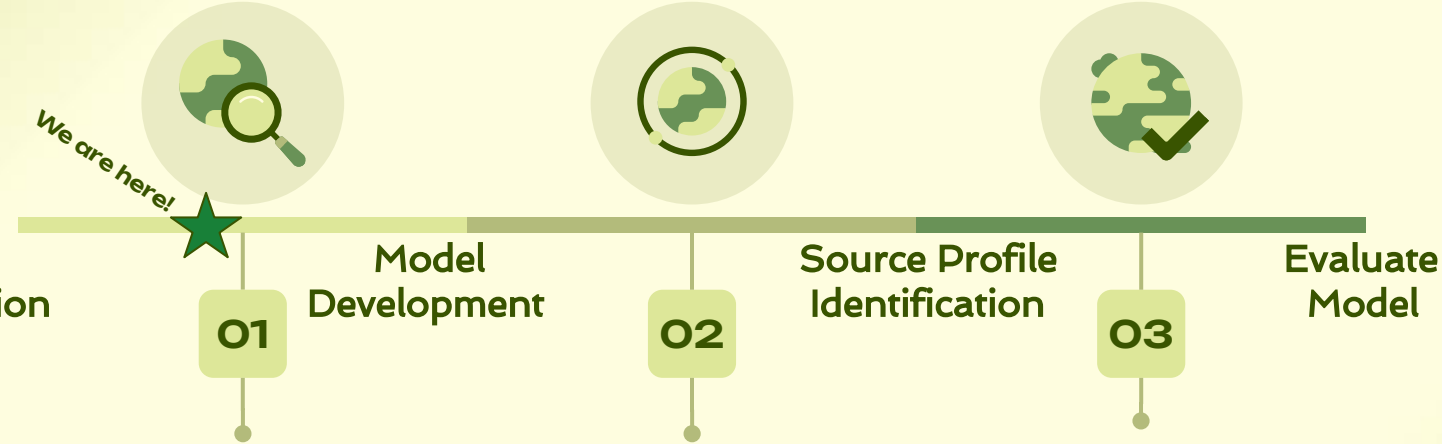
-Our data is a mixture of many unknown sources

-PMF handles uncertainty values (low uncertainty are weighted higher)

-It requires no predefined source profiles

-Air quality researchers widely use it, ensuring reliability

Our Next Steps



Analysis

Conducted data cleaning and exploratory analysis to confirm all three datasets are suitable for PMF source apportionment.

Create Model

Use PMF to create a source apportionment model and calculate source profiles and source contributions.

Assess Model

Evaluate model's potential error and utilize provided input uncertainty to improve its performance or create new iterations.

Thanks!

Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**