EPA Source Apportionment Fall Final Report

STAT 5010W

Deeya Datta, Prateek Mishra, and George Whittington

In collaboration with:

Deron Smith, M.S.
U.S. Environmental Protection Agency

December 10, 2025

1 Introduction

Air pollution is driven by a diverse number of sources such as motor vehicle exhaust, fuel combustion, and natural sources such as dust. As these processes release chemicals into the air simultaneously, the concentrations observed at monitoring sites reflect a blend of potential causes. Identifying the underlying sources responsible for the measured chemical concentrations is known as source apportionment and is essential for designing effective environmental policies while verifying emission inventories. Source apportionment techniques allow us to dissect pollutant mixtures into their underlying contributions, enabling researchers to estimate both the chemical fingerprint of each source and its contribution with respect to time.

Our capstone project was conducted in collaboration with Deron Smith, a researcher and developer at the Environmental Protection Agency (EPA). Here he works extensively on open source analytical tools for air quality measurement including the Environmental Source Apportionment Toolkit (esat). The objective of our partnership is to help strengthen the foundations of these tools and to run our own source apportionment analysis that supports both current and future EPA research.

To address the challenge of separating mixed pollution sources, environmental researchers rely on two classes of source apportionment methods. Bottom-up emission inventory models, which rely on prior knowledge about source activity, and top-down receptor models, which infer source contributions directly from measured concentrations. Receptor-based statistical techniques, such as Chemical Mass Balance (CMB) and Positive Matrix Factorization (PMF), are widely used because they require only pollutant measurements and uncertainty estimates rather than detailed emissions data. PMF decomposes the concentration matrix $\mathbf{X}$ into a matrix of source profiles $\mathbf{F}$, source contributions $\mathbf{G}$, and residual errors $\mathbf{E}$ where

$$\mathbf{X} = \mathbf{GF} + \mathbf{E} \tag{1}$$

and both $\mathbf{G}$ and $\mathbf{F}$ are non-negative.

In practice, applying PMF requires concentration data that capture a wide range of chemical species, as well as reliable estimates of measurement uncertainty. The datasets analyzed in this project were provided by the EPA and consist of pollutant measurements collected from monitoring sites in three U.S. cities: Baton Rouge, Louisiana; Baltimore, Maryland; and St. Louis, Missouri. Although each city contributes information about local air quality conditions, the datasets differ substantially in the pollutants measured, the number of chemical species included, and the temporal structure of sampling. Baltimore and St. Louis contain fine particulate matter ($PM_{2.5}$) data, which include components such as sulfate, nitrate, organic carbon, and various trace metals. In contrast, the Baton Rouge dataset focuses on volatile organic compounds (VOCs), which represent a different chemical class and serve as important precursors to ozone formation.

Each dataset also includes paired uncertainty values for all measurements, which allows PMF to weight observations according to their reliability. This feature is critical because environmental measurements often contain low concentration values and high noise observations that can influence model stability. By incorporating uncertainty into the factorization process, PMF puts less weight on less reliable data and emphasizes observations that contain the strong chemical signals.

The diversity of chemical species and sampling designs across cities allows us to investigate questions about model sensitivity, factor interpretability, and the impact of uncertainty on statistical inference. In this project, we explore these issues and characterize the latent emission sources present in each metropolitan area.

2 Project Goals

The first objective of this project is to figure out how to effectively analyze and pre-process the input data before running any models. Because we have access to the paired uncertainty values for every measurement, we can use the combination to identify which chemical species are actually providing strong signals and which are dominated by noise. Our goal is to calculate signal-to-noise ratios for the variables in each city, allowing us to characterize the reliability of each chemical species. Rather than manually excluding noisy variables, this analysis helps us verify that the modes are correctly utilizing the uncertainty values to automatically down-weight less reliable data points while preserving the information contained in the high-quality measurements.

For the modeling phase, the core goal is to successfully implement source apportionment using multiple algorithms from the family of Non-Negative Matrix Factorization (NMF) to decompose the concentration data into interpretable source profiles and source contributions. While standard Positive Matrix Factorization is the common approach, using these specific NMF algorithms allows us to handle the factorization with more flexibility, particularly in how we account for the error term. The aim is to generate source profiles that are mathematically sound and comparable to standard benchmark results.

Finally, we need to evaluate the results to ensure they are stable and physically meaningful. We will not just run the model once; we plan to use error estimation methods like Bootstrap (resampling the data) and Displacement (shifting the profile values) to quantify the variability in our factors. This will help us confirm that our solutions are robust and not just a project of random chance. Ultimately, the success of the project depends on our ability to translate these statistical outputs into real-world environmental conclusions, correctly identifying distinct pollution sources like vehicle exhaust or industrial emission with the specific context of each city.

3 Approach

The approach is organized into three major components: data preparation, exploratory analysis, and factor modeling. We begin by standardizing the datasets, ensuring consistent naming, units, and formatting across cities. Because the three locations differ significantly in sampling schedules, we treat each observation as an independent snapshot rather than a continuous time series. Although the independence assumption is common in most receptor modeling, including PMF and NMF, air quality measurements might naturally exhibit short term autocorrelation. The irregular spacing of observations makes standard autoregressive modeling infeasible, but mild dependence may still be present. For this exploratory phase, treating observations as independent allows us to apply matrix factorization methods consistently across cities, while future work could incorporate time dependency or clustering to better capture recurring pollution patterns. We also remove redundant or derived chemical species, such as Baltimore's Organic Matter variable, to avoid artificial inflation of specific chemical signals. Each dataset undergoes distributional checks, log transformations, and initial diagnostics to ensure assumptions and stability for factor analysis.

The exploratory analysis focused on understanding the chemical relationships within each dataset, examining correlations among species, and identifying chemically meaningful markers by accounting for uncertainty. These diagnostics provide insight into which chemical species are likely to drive factor separation and which may be dominated by noise or detection limits.

Our modeling strategy is Positive Matrix Factorization (PMF), a receptor modeling technique widely used in environmental applications because it enforces non negativity and incorporates uncertainty directly into the weighting scheme. PMF will be used to estimate latent source profiles and their contributions for each city, and we plan to examine several candidate factor numbers to determine the most interpretable and stable solutions. As part of a comparative analysis, we also consider alternative methods such as Chemical Mass Balance (CMB), a model that relies on known source profiles, to contextualize how PMF derived factors align with established emission signatures. Additionally, we intend to explore Least-Squares NMF (LS-NMF) and Weighted Semi-NMF (WS-NMF) as related matrix factorization approaches that may offer other insights which might have been missed. This allows us to evaluate how chemical composition, uncertainty structure, and sampling design shape the identification and interpretability of emission sources across the three metropolitan areas.

4 Data Description

We received six data files from the U.S. Environmental Protection Agency (EPA) for three cities: Baton Rouge, St. Louis, and Baltimore. Each city has two files, one containing pollutant concentrations and another containing measurement uncertainties. The three cities measured fundamentally different types of air pollutants at different times, reflecting what researchers thought would be

most informative about air quality in each location.

Baton Rouge monitored 42 different volatile organic compounds (VOCs), essentially gaseous chemicals containing carbon, during June 2005. The 307 measurements came from the EPA's monitoring network designed to track compounds that contribute to ozone formation. These VOCs include familiar compounds like benzene (found in gasoline) and propane (used for heating and cooking), as well as dozens of related chemicals emitted by vehicles, industry, and natural gas systems. Concentrations are reported in ppbC (parts per billion carbon), a unit that counts carbon atoms rather than whole molecules, making it easier to compare different organic compounds.

St. Louis and Baltimore took a different approach by measuring tiny particles suspended in the air, called $PM_{2.5}$ (particulate matter smaller than 2.5 micrometers). St. Louis collected 418 hourly samples in June 2001, analyzing 13 chemical components including metals like lead and zinc, ions like sulfate, and carbon based material. Baltimore provided the longest record with 630 daily samples from 2000 to 2007, measuring 27 different particle components. Both cities report concentrations in $\mu g/m^3$ (micrograms per cubic meter), showing how much of each substance exists in a given volume of air.

5 About the Data

The first set of measurements come from Baltimore and it represents the longest timeline out of the three, spanning seven years from December 2000 to July 2007. However, rather than having consistent measurements over time, data seems to have been collected on a strict one-in-three sampling schedule where a sample was taken approximately every three days, but the gap fluctuates from one day between samples, sometimes up to six days between samples. Because of this discontinuous structure of collection, we cannot treat the data as a traditional time series. We instead treat each day as an independent snapshot of the city's air quality. For what was actually measured here, this site focuses on $PM_{2.5}$ speciation, capturing twenty-six distinct chemical species. The variables include major ions such as sulfate and nitrate, as well as carbon species including elemental carbon and organic carbon. It also features a wide array of seventeen trace elements like aluminum, lead, and zinc that act as markers for specific sources like soil dust or brake wear.

The next city, Baton Rouge, was measured a bit differently than Baltimore. Rather than being periodically over a long period of time, the measurements were gathered in the summer months of June through September across 2005 and 2006. The sampling schedule was also highly specific, with measurements taken almost exclusively at 03:00 and 06:00 local time. This timing might suggest the monitoring in this site was designed to capture background conditions or pollutant accumulation overnight before sunlight triggers photochemical reactions and before the morning traffic rush disperses the concentration levels. Chemically, this dataset is also rather different because it measures Volatile Organic Compounds rather than particulate matter. There are forty-one variables covering distinct hydrocarbon groups such as alkanes, alkenes, and aromatics. This

includes the important BTEX group (benzene, toluene, ethylbenzene, and xylene). A key feature of this dataset is also the inclusion of aggregate variables like TNMOC (Total Non-Methane Organic Carbon) and a specific "Unidentified" variable, which seems to account for the mass of carbon that could not be chemically resolved.

The final dataset, St. Louis, acts as a hybrid of the previous two, offering high-frequency data, but over short and discontinuous durations. The data is separated into three specific month-long campaigns occurring in June 2001, November 2001, and March 2002. While the data is nominally hourly, it is not continuous, and we observed significant intra-day gaps. Some days contain a full twenty-four hour record, but others have large blocks of missing data, reinforcing the decision to avoid autoregression time series models. Chemically, the site measures $PM_{2.5}$ like Baltimore, but with a more limited scope. The thirteen variables represent a subset of the previously discussed chemical species list, focusing only on the most abundant markers such as sulfate and specific transition metals including copper, iron, and zinc. This reduced dimensionality suggests the modeling will likely resolve fewer distinct source factors for this location.

6 Exploratory Data Analysis

Our initial examination of the three datasets revealed excellent data quality with over 99% complete records and minimal missing values. After standardizing date formats and column names across the six files, we calculated summary statistics for each chemical species.

Baton Rouge VOC concentrations ranged from trace levels to over 5 ppbC for abundant compounds like ethane, propane, and TNMOC. St. Louis showed $PM_{2.5}$ mass at 3.2 $\mu g/m^3$ as the dominant component, followed by sulfate and organic carbon. Baltimore exhibited $PM_{2.5}$ averaging 11.8 $\mu g/m^3$ with substantial contributions from organic matter, sulfate, and metals.

The distribution of mean concentrations differs substantially across cities. Baltimore sits low overall with a median near zero and limited variability. The top 5 species are $PM_{2.5}$ (2.7 $\mu g/m^3$), organic matter (1.9 $\mu g/m^3$), organic carbon (1.7 $\mu g/m^3$), sulfate (1.6 $\mu g/m^3$), and ammonium ion (1.0 $\mu g/m^3$). St. Louis shows a similar low center but higher variability. The top 5 are $PM_{2.5}$ (3.1 $\mu g/m^3$), organic carbon (1.5 $\mu g/m^3$), sulfate (1.4 $\mu g/m^3$), total nitrate (0.9 $\mu g/m^3$), and elemental carbon (0.5 $\mu g/m^3$).

Baton Rouge shows higher concentrations and far more variability. The top 5 are all VOCs: TNMOC (5.1 ppbC), unidentified compounds (3.8 ppbC), propane (2.6 ppbC), ethane (2.6 ppbC), and isopentane (2.5 ppbC). These numbers appear higher than the other cities' measurements, but ppbC and $\mu g/m^3$ measure different things. Parts per billion carbon counts carbon atoms per billion air molecules, while micrograms per cubic meter measures mass. The key difference is that Baton Rouge measures gases while the other cities measure particles.

Dominant Chemical Species by City

Comparison of Physical Concentrations vs. Log-Transformed Model Inputs

City	species	Arithmetic Mean (µg/m³ or ppbC)	Log Mean (Model Unit)	Max Value
Baltimore	om	6.60	1.94	58.10
Baltimore	sulfate	4.83	1.61	30.20
Baltimore	pm2.5	15.57	2.67	76.30
Baton Rouge	ethane	13.89	2.58	49.97
Baton Rouge	propane	15.78	2.64	70.96
Baton Rouge	tnmoc	199.90	5.15	708.47
St. Louis	sulfate	3.84	1.41	16.90
St. Louis	pm2.5	22.95	3.06	107.00
St. Louis	organic_carbon	4.77	1.53	71.20

Figure 1: *Summary statistics and concentration distributions for all three cities*

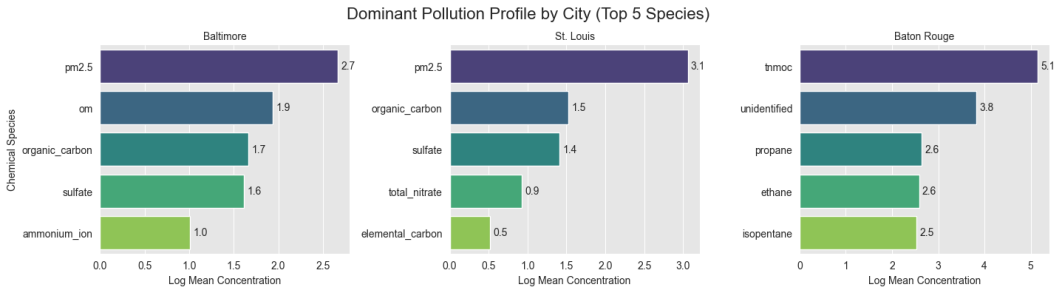

Figure 2: *Concentration distributions by city showing boxplots and top species*

Correlation analysis revealed which species vary together, suggesting common pollution sources. Baton Rouge showed extensive correlation patterns with 171 species pairs above $r = 0.7$. Several groups emerged: xylene isomers (m/p-xylene and o-xylene, $r = 0.935$), ethylbenzene with both xylene isomers ($r > 0.91$), and toluene correlating strongly with all of these ($r > 0.82$). These species likely share emission sources such as traffic exhaust and industrial solvents. Light hydrocarbons also clustered together with isopentane and n-butane at $r = 0.912$.

St. Louis showed minimal correlation structure with only two pairs exceeding $r = 0.7$. Organic carbon correlated moderately with $PM_{2.5}$ mass ($r = 0.715$) and elemental carbon ($r = 0.700$). The OC-EC correlation makes sense because combustion sources release both at the same time. The sparse overall patterns likely reflect both the limited species diversity (13 components) and potentially more independent sources at this industrial monitoring site. Most metals showed weak correlations with each other and with carbon species, suggesting different emission sources for

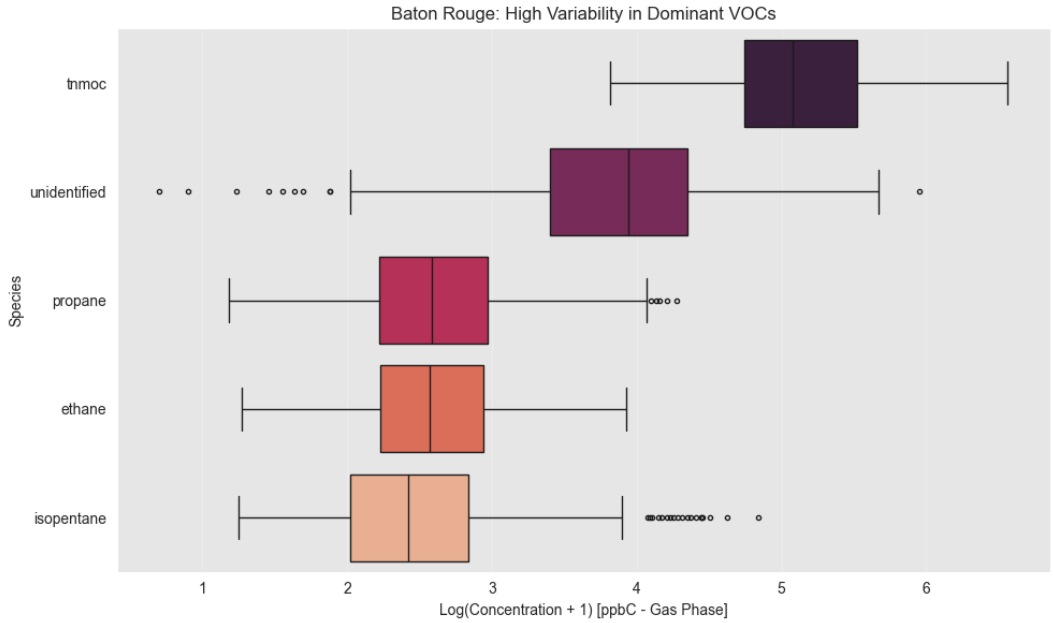

Figure 3: *Baton Rouge: Distribution of concentrations for the top 5 VOC species*

different pollutants.

Baltimore demonstrated moderate correlation structure focused on three main groups. Additionally, we identified one critical data quality issue. Baltimore's Organic Matter correlated perfectly ($r = 1.0$) with Organic Carbon because it is calculated as Organic Carbon multiplied by 1.4. Including both would double count organic contributions, so we flagged Organic Matter for removal. The other sets had no one-to-one correlations among their features. Metals showed specific clustering with chromium and nickel very strongly correlated ($r = 0.952$). Ammonium ion and sulfate were strongly coupled ($r = 0.895$), and $PM_{2.5}$ mass correlated moderately with both sulfate ($r = 0.740$) and organic carbon ($r = 0.725$), identifying these as major contributors to total particle mass.

Tables 1, 2, and 3 present the highly correlated species pairs for each city.

Species 1	**Species 2**	r
cis-2-pentene	trans-2-pentene	0.946
2,2,4-trimethylpentane	2,3,4-trimethylpentane	0.946
m/p-xylene	o-xylene	0.935
ethylbenzene	o-xylene	0.914
isopentane	n-butane	0.912
2,3-dimethylpentane	n-heptane	0.909
ethylbenzene	m/p-xylene	0.909
2,3-dimethylbutane	3-methylpentane	0.901

Table 1: *Baton Rouge: Top 8 Highly Correlated VOC Species Pairs ($r > 0.90$)*

Species 1	**Species 2**	r
Organic Carbon	$PM_{2.5}$	0.715
Elemental Carbon	Organic Carbon	0.700

Table 2: *St. Louis: Highly Correlated Species Pairs ($r > 0.70$)*

Species 1	**Species 2**	r
Organic Matter	Organic Carbon	1.000
Chromium	Nickel	0.952
Ammonium Ion	Sulfate	0.895
Copper	Potassium Ion	0.809
Ammonium Ion	$PM_{2.5}$	0.784
$PM_{2.5}$	Sulfate	0.740
Organic Matter	$PM_{2.5}$	0.726
Organic Carbon	$PM_{2.5}$	0.725

Table 3: *Baltimore: Highly Correlated Species Pairs ($r > 0.70$)*

The max-to-mean ratio reveals highly episodic pollution across all cities. Baltimore $PM_{2.5}$ averaged 15.6 $\mu g/m^3$ but peaked at 76.3 $\mu g/m^3$ (4.9x). Sulfate ranged from 4.8 to 30.2 $\mu g/m^3$ (6.25x) and organic carbon showed 8.8-fold variation. St. Louis exhibited larger swings: organic carbon went from 4.8 $\mu g/m^3$ to 71.2 $\mu g/m^3$ (14.9x), elemental carbon varied 8.3-fold, and mass ranged from 23 to 107 $\mu g/m^3$. Baton Rouge VOCs showed n-hexane varying 10.8-fold, isopentane 7.7-fold, and unidentified compounds spiking to 385 ppbC from a 63 ppbC baseline (6.1x). Pollution comes in massive, short-lived plumes rather than steady streams.

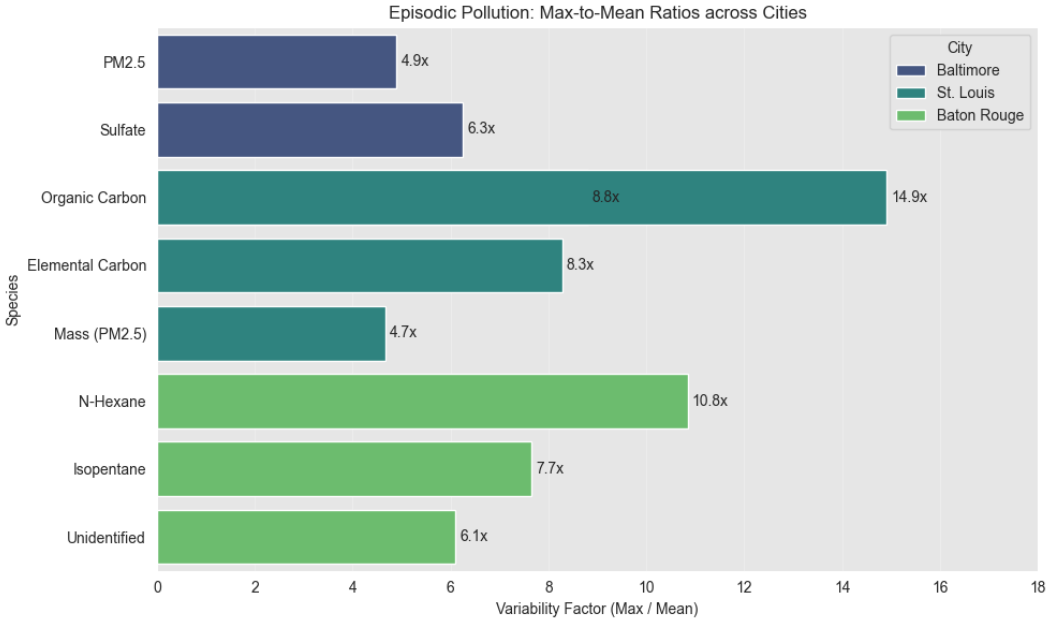

Figure 4: *Extreme value analysis showing max-to-mean ratios for each city*

These spikes create a mathematical problem. If we feed raw data into the model, the algorithm will be biased by magnitude, obsessing over fitting these few massive numbers and ignoring subtle correlations in the rest of the data. Additionally, none of these datasets represent continuous timelines. Baton Rouge and St. Louis captured one-month intensive campaigns, while Baltimore's seven-year record contains gaps between sampling days. Our chosen method, PMF, relies on concentration variability to identify sources, so the spikes and fluctuations we observe provide the necessary information even though we're not capturing every pollution event continuously.

7 Next Steps

Our next steps focus on implementing the source apportionment models and evaluating their performance across the three datasets. We will begin with PMF using the uncertainty weighted model to estimate latent source profiles and contributions. Insights from our exploratory analysis such as chemical species correlations, signal-to-noise ratios, and variability patterns, will help guide our initial selection of candidate factor numbers. We will run PMF across a range of factor counts and identify the most interpretable solutions based on residual structure, stability across initializations, chemical plausibility, and model scores.

In addition to PMF, we will implement Least-Squares NMF (LS-NMF) and Weighted Semi-NMF (WS-NMF) as complementary matrix factorization approaches. LS-NMF provides a baseline non-negative decomposition without uncertainty weighting, while WS-NMF allows weighted contributions and may offer improved performance for chemical species with low or variable concentrations. Comparing these approaches will help identify whether certain datasets benefit from alternative formulations or whether PMF provides the most stable and interpretable factors. Following model estimation, we will also apply clustering to the resulting factor profiles and factor contributions to assess whether the sources have distinct chemical patterns across observations.

Although the datasets do not support formal time series modeling due to irregular sampling intervals, we may still examine temporal patterns such as pollution episodes or recurring VOC peaks to assist interpretation of the extracted factors. Air quality measurements often exhibit short term autocorrelation, and future work could incorporate dependency structures or temporal regularization to better capture dynamic behavior. Together, these steps will allow us to implement and compare multiple factorization methods, integrate clustering for exploratory and validation purposes, and translate the resulting outputs into meaningful environmental insights across the three metropolitan areas.

References

[1] Paatero, P., & Tapper, U. (1994). Positive Matrix Factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, **5**(2), 111–126.

[2] Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, **401**, 788–791.

[3] Smith, D. Environmental Source Apportionment Toolkit (ESAT). U.S. Environmental Protection Agency, Office of Research and Development. Available at: https://github.com/quanted/esat

[4] Guttikunda, S. (2011). Primer on Pollution Source Apportionment. UrbanEmissions.info. Available at: https://urbanemissions.info/publications/primer-on-pollution-source-apportionment/

[5] U.S. Environmental Protection Agency. (2025). Source Apportionment 101 [Lecture slides]. University of Georgia Statistics Seminar.

Code Appendix

The R code for this analysis is available at: https://github.com/g-whittington/EPA-Source-Apportionment