Deron Smith
Computer Scientist

US Environmental Protection Agency

Research Focus

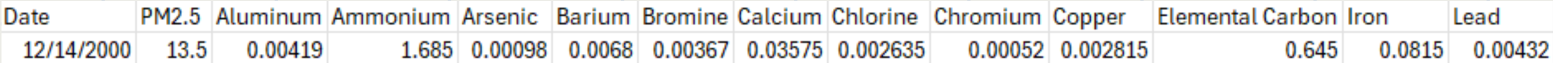
Data Analytics and Machine Learning

Numerical Modeling and Optimization Problems

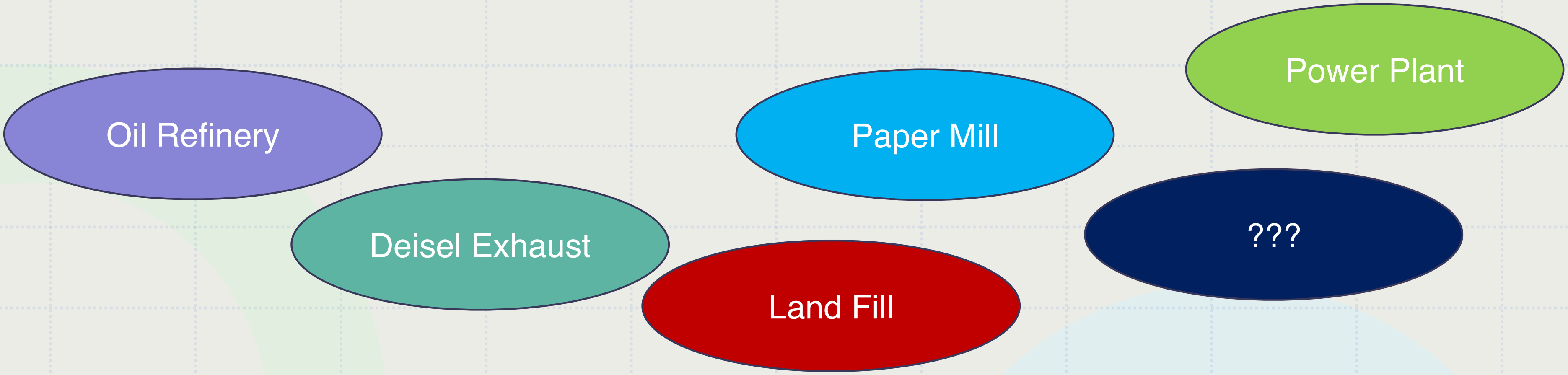
Full-Stack Software Development

Project: Source Apportionment

Environmental sampling and sensor data measurements are often the product of many different unknown sources. Source apportionment is the process of statistically estimating the relative contributions from a set of sources to the observed sample data.



Date	PM2.5	Aluminum	Ammonium	Arsenic	Barium	Bromine	Calcium	Chlorine	Chromium	Copper	Elemental Carbon	Iron	Lead
12/14/2000	13.5	0.00419	1.685	0.00098	0.0068	0.00367	0.03575	0.002635	0.00052	0.002815	0.645	0.0815	0.00432



Source Apportionment Data

SOURCE APPORTIONMENT IS A NUMERICAL PROCEDURE THAT IS TYPICALLY INDEPENDENT OF UNITS.

DATA COULD REPRESENT CHEMICALS, CONTAMINANTS, HEAVY METALS, OR ANY FEATURE WHICH CAN BE REPRESENTED NUMERICALLY.

THE DATA COULD BE AIR QUALITY SENSOR READINGS, WATER QUALITY MEASUREMENTS, OR OTHER SENSOR DATA.

ENVIRONMENTAL SOURCE APPORTIONMENT DATA IS GENERALLY TIME-SERIES DATA, THOUGH IS CONSIDERED TEMPORALLY INDEPENDENT.

THE CONTEXT OF THE DATA DIRECTLY DETERMINES HOW YOU INTERPRET SOURCE APPORTIONMENT OUTPUT.

ENVIRONMENTAL TIME-SERIES DATA CAN ALSO HAVE CORRESPONDING UNCERTAINTY VALUES, DUE TO SENSOR UNCERTAINTY/LIMITATIONS.

Source Apportionment Output

The objective of source apportionment is to produce source profiles and source contribution time-series.

- Source Profiles – Feature profiles that represent identifying signature of the source, often considered static.
- Source Contributions – How much the source contributions to each sample.

Data

Three sample datasets are available for this project. Each dataset represents air quality data measurements taken from a single location at irregular time intervals.

Baton Rouge – The data was downloaded from the EPA Air Quality System. The data contains 307 samples and 41 features, collected from a Photochemical Air Monitoring Station (PAMS) between the dates June-August 2005 and June-September 2007. The input and uncertainty data files are comma separate list (csv).

St Louis – The data is taken from the East St. Louis Supersite (Superfund site) during June 2001, November 2001, and March 2002. The data contains 418 samples and 13 features. The input and uncertainty data files are csv files.

Baltimore – The data contains 630 samples and 26 features. The data represents particulate matter (PM) measurements taken from December 2000 to July 2007. The input and uncertainty data are .txt files, which are tab separated.

Research Objectives

How can the input data be analyzed?

Calculate source profiles and source contributions, a source apportionment model.

Evaluate source apportionment model against input data, metrics and plots.

Utilize the input uncertainty into the model.

Can we evaluate the model’s potential error?

How can we evaluate the results of multiple models, created from the same input data?

What other conclusions can we make about what the input data and model represent?

Contact Information: smith.deron@epa.gov