Source Apportionment of Combustion Generated Particulate Matter Air Pollution using Excitation Emission Matrix Fluorescence Spectroscopy and Machine Learning

Jay W Rutherford

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2019

Reading Committee:

Jonathan D. Posner, Chair

Igor Novosselov, Chair

David A. C. Beck

Program Authorized to Offer Degree:

Chemical Engineering

University of Washington

Abstract

Source Apportionment of Combustion Generated Particulate Matter Air Pollution using Excitation Emission Matrix Fluorescence Spectroscopy and Machine Learning

Jay W. Rutherford

Chair of the Supervisory Committee:
Professor Jonathan D. Posner
Department of Mechanical Engineering

Exposure to particulate matter (PM) air pollution is the world's largest environmental health risk accounting for millions of premature deaths and disability-adjusted life years annually. PM originates from natural and anthropogenic sources such as dust from soil, combustion engines, and forest fires, among many others. PM exposure is quantified by measuring its mass concentration in air. This measurement alone does not identify the sources of PM exposure, which can inform effective mitigation strategies and allow for studying source-specific health effects. There are several options for source apportionment (e.g. GC-MS and X-ray fluorescence), but they are costly and time consuming to conduct. Alternative methods for source apportionment using low-cost techniques would be beneficial to the study of air pollution and its health effects. In this dissertation, I develop a method for source apportionment of

combustion generated PM using fluorescent Excitation Emission Matrix (EEM) fluorescent spectroscopy and machine learning.

First, I collected PM samples from combustion sources of concern to human health in the laboratory. I analyzed cyclohexane extracts of cigarette smoke, diesel exhaust and wood smoke by EEM fluorescent spectroscopy and using the World Health Organization's guideline for annual mean PM exposure of 10 $\mu g/m^3$ as a basis of comparison I show EEM is sensitive enough to detect combustion generated PM at levels well below those of concern to human health.

Next, mixtures of the same laboratory sources are analyzed using EEM. Combining measurements of the individual sources with those of mixtures, I apply several machine learning techniques and a simple linear model to perform source apportionment and identification from the mixtures and compare the results. A convolutional neural network (CNN) is found to have the best performance of all methods investigated. I describe in detail the architecture and data augmentation approach used for the CNN.

Finally, the EEM-Machine Learning approach is used for source apportionment of environmental samples. Results and filter samples from an exposure assessment panel study are used for this analysis. The samples were analyzed using X-ray fluorescence and source apportionment was conducted using Positive Matrix Factorization. Filters, archived in a freezer, were extracted with cyclohexane and analyzed by EEM. The resulting EEM spectra and source contribution estimates from PMF were used as training data for the application of machine learning. A CNN with the same architecture as applied to the laboratory samples and Principal Component Regression showed similar results in predicting contributions from combustion generated PM. These methods were able to reproduce the XRF-PMF results with R^2 values as high as 0.84 for vegetative burning and 0.52 for traffic emissions.

Table of Contents

List of Figures

List of Tables

Citation to previously published work

Portions of this dissertation, Chapter 2 and 3, have appeared in the following peer-reviewed article:

Rutherford, J. W.; Dawson-Elli, N.; Manicone, A. M.; Korshin, G. V.; Novosselov, I. V.; Seto, E.; Posner, J. D. Excitation Emission Matrix Fluorescence Spectroscopy for Combustion Generated Particulate Matter Source Identification. *Atmospheric Environment* **2019**, 117065. https://doi.org/10.1016/j.atmosenv.2019.117065.

Acknowledgments

This work was funded by the National Institute of Biomedical Imaging and Bioengineering (NIBIB) grant U01 EB021923. Data analysis was supported by the Data Intensive Research Enabling Clean Technology (DIRECT) NSF National Research Traineeship (DGE-1633216) and the UW eScience Institute Data Science Incubator.

I thank Ben Sullivan, Garrett Allawatt, and Devin Udesen of the UW Clean Cookstoves Lab for their assistance collecting wood smoke samples, Jim Stewart for his help collecting diesel samples, and Prof. Anne Manicone and Keqin Gong for allowing me to collect cigarette smoke samples.

Without the help of Prof. Tim Larson and Tim Gould, much of my work would not have been possible. I'd like to thank Tim Gould for his guidance on sampling, lending me sampling equipment and helping me locate archived environmental samples. I'd like to thank Prof. Tim Larson for giving me a freezer full of samples, the data for those samples, and many valuable discussions.

Thank you to Prof. Edmund Seto, Prof. Gregory Korshin, Byron Ockerman, Jiayang "Joe" He, Gaurav Mahamuni, and Ravi Vaddi for your valuable discussions at our regular project meetings. Thank you to Prof. David Beck, Neal Dawson-Elli and Ben Ponto for your help getting the data analysis aspect this research to work for the first time and thank you to Bernease Herman and others at the eScience Institute for your help in refining my data analysis.

I also owe thanks to many who were not directly involved in the work presented in this dissertation. Thank you to Russel Dills, Shar Sammy, and Jacqui Ahmad for teaching me how a certified laboratory operates and allowing me to work in your lab. Thank you to Profs. Eric Stuve,

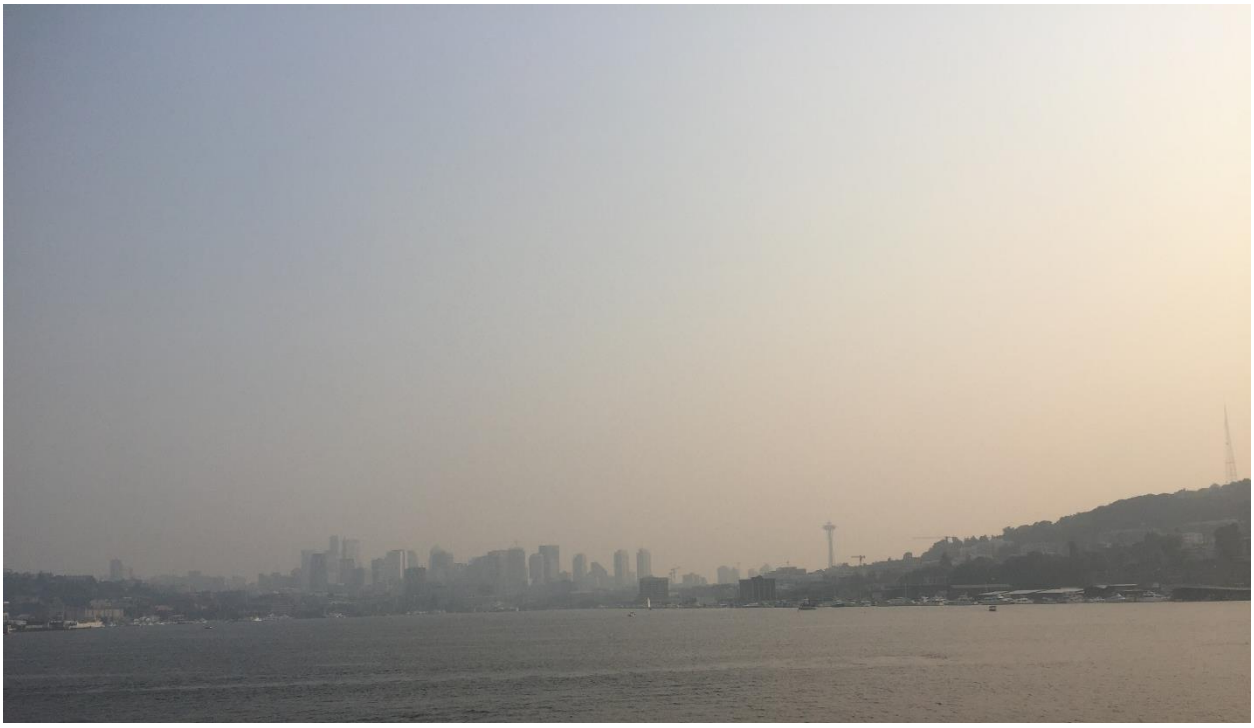
Stuart Adler, Qiuming Yu, Elizabeth Nance, and Andy Kim for the opportunity to teach alongside you and all the lessons learned from those experiences.

Finally, I'd like to thank my advisors Prof. Johnathan Posner and Prof. Igor Novosselov. Dr. Posner, thank you for your guidance on my research and mentorship during times when I was faced with difficult decisions. Igor, thank you for never giving up hope that I could make the EEM analysis method work.

Dedication

To my family and friends for your support.

Chapter 1: Introduction



1.1 Air Pollution

According to the 2017 Global Burden of Disease study, air pollution is the world's largest environmental health risk accounting for 4.9 million deaths and 147 million disability-adjusted life-years annually.[1] About 90% of the world's population lives in areas that exceed World Health Organization guidelines for air quality based on recent estimates.[2,3] Efforts to combat air pollution in North America and Europe have resulted in reduced levels of air pollution and this has contributed to an increase in life expectancy and quality of life.[4,5]

In the United States, air quality regulation began at the state level in the 1950s led by California and was followed by federal regulation starting in 1955. The first regulations were largely ineffectual and updates to these regulations have improved effectiveness and adjusted standards based on current understandings of the health effects of air pollution. The Clean Air Act of 1963 followed by the Clean Air Act Amendments of 1970 led to the creation of the national ambient air quality standards (NAAQSs) that are promulgated by the United States Environmental Protection Agency (USEPA). The pollutants and levels regulated by NAAQSs have changed over time based on the regular review of epidemiological and toxicological studies of the health effects of air pollution.[6] Currently, the NAAQSs regulate six types of air pollution: carbon monoxide (CO), lead, nitrogen dioxide (NO_2), ozone, particulate matter of two different sizes ($PM_{2.5}$, PM_{10}), and sulfur dioxide(SO_2).[7]

As part of the Clean Air Act, local agencies report the Air Quality Index (AQI) to keep the public informed about air pollution levels. The AQI is on a scale from 0 to 500 with various ranges corresponding to levels of health concern. For example, an AQI between 0 and 50 is "good" and does not indicate any increased health risks and an AQI of 151 to 200 is "unhealthy" indicating that individuals should avoid exposure by staying indoors and limiting exertion. The

various levels and colors associated with the AQI are shown in Table 1-1. The AQI can be elevated due to levels of any of the regulated air pollutants. These pollutants are measured using different units, $\mu g/m^3$ for PM and lead, ppm for ozone and CO, and ppb for SO_2 and NO_2 and each pollutant has varying health effects per unit of concentration. The purpose of the AQI is to simplify interpreting various levels of these pollutants. The EPA has established the process for converting pollution levels to AQI.[8] In summary, the levels of each pollutant are measured and the AQI is calculated from each of these levels by interpolating the corresponding pollution concentration to a value of the AQI (see Table 1-1 for corresponding $PM_{2.5}$ and ozone concentrations). After calculating the AQI in this way for all pollutants, the highest AQI is reported. For example, if $PM_{2.5}$ was at a concentration of 55.5 $\mu g/m^3$ and ozone was at 0.0055 ppm the AQI would be reported as: "151 – Unhealthy, due to $PM_{2.5}$."

Table 1-1: Air Quality Index (AQI) Levels. The AQI ranges from 0 to 500 and is calculated differently for each of six pollutants (ozone, $PM_{2.5}$, PM_{10}, CO, SO_2 or NO_2). The table shows the corresponding concentration ranges for $PM_{2.5}$ and ozone for illustration.

AQI	**Corresponding Concentration Range**		**Description**	**Color**
	$PM_{2.5}$ ($\mu g/m^3$, 24 hr. avg.)	**Ozone** (ppm, 8 hr. avg)		
0 to 50	0.0 - 12.0	0.000 - 0.054	Good	Green
51 to 100	12.1 - 35.4	0.0055 - 0.070	Moderate	Yellow
101 to 150	35.5 - 55.4	0.071 - 0.085	Unhealthy for Sensitive Groups	Orange
151 to 200	55.5 - 150.4	0.086 - 0.105	Unhealthy	Red
201 to 300	150.5 - 250.4	0.106 - 0.200	Very Unhealthy	Purple
301 to 500	250.5 - 500.4	--*	Hazardous	Maroon

*Calculated using 1-hour average levels

1.2 Particulate Matter Air Pollution

Particulate matter (PM) is a major component of air pollution consisting of microscopic particles that remain suspended in the atmosphere for minutes to weeks. These microscopic particles come from natural and anthropogenic sources such as diesel engines, agricultural burning, cooking with biomass, electrical power generation, pollen, bacterial spores, dust from soil, evaporation of sea spray, and forest fires. In addition to the examples above of *primary particles*, there are *secondary particles* that form in the atmosphere as a result condensation of volatile organic compounds, interactions between particles in clouds, and chemical reactions between primary particles, other pollutants, and solar radiation.[6]

The human respiratory system is effective at removing many of the particles from the air we breathe before they enter deep into the lungs. Larger particles are removed by impaction and sedimentation in the ciliated airways of the lungs and smaller particles can penetrate deeper into the lungs.[9] Breathing rate, mouth vs. nose breathing, and the health of an individual determine the deposition of particles in the respiratory tract. To more easily estimate how much PM is deposited in the respiratory tract regulators and epidemiologists have adopted the practice of classifying PM into several size categories that are likely to be deposited in the lungs. PM sizes refer to the aerodynamic equivalent diameter (AED); the diameter of a sphere of unit density with the same terminal settling velocity as the particle. Generally speaking, nearly all particles larger than 10 μm AED are removed from the air we breathe before leaving the head, particles ranging in size from 2.5 μm to 10 μm are able to penetrate into the ciliated conducting airways of the respiratory tract and particles smaller than 2.5 μm ($PM_{2.5}$) can bypass the body's natural defenses against PM and reach the gas exchange regions of the lungs,[10] making $PM_{2.5}$ of the most concern in health studies. These cut off sizes are what is most widely used in the study of PM health effects, but their

appropriateness is not absolute and other metrics may be used.[11] Figure 1.1 is a graphical illustration of the size of $PM_{2.5}$ relative to beach sand and a human hair showing that an aerosol of 2.5 μm in diameter, the upper bound for $PM_{2.5}$, is about 1/20 the diameter of a human hair.

The NAAQSs set by the USEPA regulate levels of $PM_{2.5}$ air pollution. The NAAQSs standards for maximum exposure levels are 15 $\mu g/m^3$ annual average and 35 $\mu g/m^3$ daily average. The World Health Organization (WHO) guidelines are lower at 10 and 25 $\mu g/m^3$, respectively.[12,13]

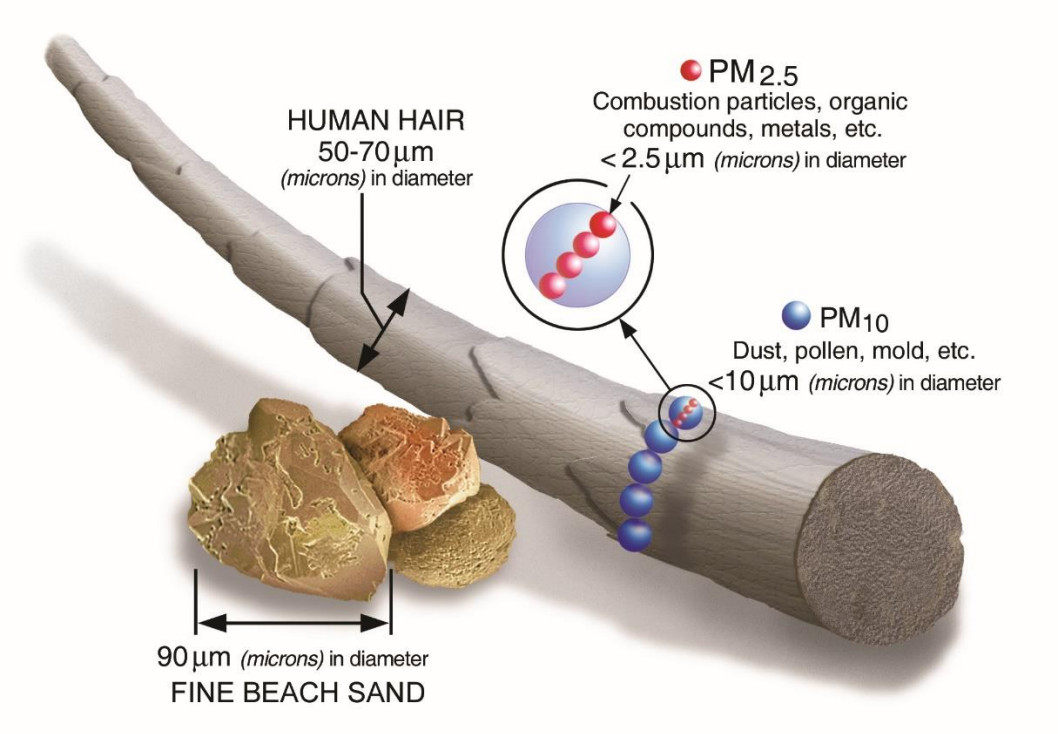


Figure 1.1: Graphical illustration of the relative size of particulate matter (available from EPA.gov).[14]

1.3 Health Impacts of Air Pollution

Many studies have shown links between mortality and exposure to elevated $PM_{2.5}$ mass

concentration.[15] For example, a recent study by Dai et. al. at the Harvard School of Public Health found $PM_{2.5}$ was associated with increases in mortality. An increase of $PM_{2.5}$ concentration of 10 $\mu g/m^3$ (averaged over 2 days) was found to increase respiratory deaths by 1.71% (95% confidence interval-1.06-2.35%).[16] To give perspective on $PM_{2.5}$ mass concentrations: for Seattle in 2015 the yearly average concentration was 7 $\mu g/m^3$ and the maximum 2-day average was 26 $\mu g/m^3$, in Beijing these values were 83 $\mu g/m^3$ and 440 $\mu g/m^3$ respectively.[17,18]

It is certain that increased $PM_{2.5}$ mass concentration has negative health effects and regulation of $PM_{2.5}$ in North America and Western Europe has led to improvements in health. However, the relative impact of specific sources of PM and their chemical composition is not as well understood. Some studies suggest that certain sources of PM are worse for health than others. For example, black carbon, which is associated with traffic, has been shown to be worse for health than $PM_{2.5}$ alone,[19,20] but the body of evidence as a whole does not conclusively show what sources or combinations thereof are the worst for health.[4,21–23] To address this lack of consensus more studies are needed that quantify mass concentration, chemical composition and source of PM simultaneously. Unfortunately, this type of information is difficult and expensive to gather with current monitoring techniques that require chemical analysis using techniques such as gas chromatography-mass spectrometry (GC-MS) and X-ray fluorescence (XRF) to be performed in addition to measurement of PM mass concentration.

Asthma is known to be exacerbated by $PM_{2.5}$ exposure.[24] In practice, clinicians advise asthmatic patients to avoid exposure to pollution that they are sensitive to and to avoid exertion outdoors when air quality is poor.[25] A study using parental questionnaires and proximity to roadways found increased asthma risk in children with exposure to second-hand smoke, but not with roadway proximity.[26] New, more efficient tools for monitoring source-specific PM exposure

will enable continued and more quantitative research in the area of source-specific health impacts for respiratory diseases like Asthma.

1.4 Source Apportionment by Receptor Modeling

If the composition of air pollution is measured and the composition of sources is known, air pollution can be unmixed using a process called receptor modeling. Receptor modeling assumes a mass balance between the sources of air pollution and the receptors (i.e. where samples are collected) expressed as:

$$c_{ij} = \sum_{k=1}^{P} g_{ik} f_{kj}$$

Where c_{ij} represents airborne mass concentration in sample *i* of species *j*, g_{ik} is the total mass concentration for sample *i* for source *k*, and f_{kj} is the mass fraction of species *j* in source *k*. There are various mathematical approaches for solving this mass balance. The chemical mass balance model (CMB) takes source profiles (f_{kj}) and species concentration data (c_{ij}) and solves for source contributions (g_{ik}). The CMB can be applied to one sample at a time and is useful when underlying sources are known. The number of sources and the source profiles may be adjusted to improve the model results.[27]

The most widely used method for receptor modeling is Positive Matrix Factorization (PMF). In PMF, the source profiles are derived from the samples, so the underlying source compositions don't need to be specified. Unlike the CMB model, PMF cannot be applied to individual samples, and many samples with varying source contribution levels are needed for the model to derive accurate source profiles from the samples. In PMF modeling, the expected number of contributing sources are specified and the PMF model finds source profiles (f_{kj}) and

source contributions (g_{ik}) to best fit the species concentration data (c_{ij}). The derived profiles are identified by comparison to known source profiles. The profiles do not need to match exactly to be identified because source profiles are expected to vary from location to location. The accuracy of the derived profiles, compared to existing profiles, is an indication of model validity so models with varying numbers of sources are fit to evaluate the correct number of sources to be used in the modeling process.[27]

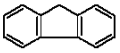
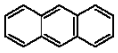
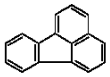
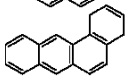
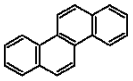
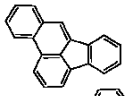
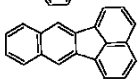
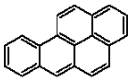
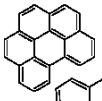
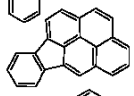
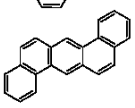
Source apportionment is used to estimate the heath-effects of different sources of particulate matter as discussed in section 1.3. Another application of source apportionment is the evaluation of the effectiveness of air pollution mitigation efforts. For example, in a wood smoke pollution impacted community about 1,200 old woodstoves were replaced with cleaner-burning models. Source apportionment for samples collected in the winter months the year before and the year after the changeout found the wood smoke component of $PM_{2.5}$ air pollution was reduced by 28% showing the effectiveness of the changeout.[28] Regardless of the mathematical method of receptor modeling used for source apportionment, some type of chemical composition measurement is required.

1.5 Chemical Analysis of Particulate Matter

Particulate matter comes from many sources, as illustrated by the list of examples in section 1.1. Even if each source produced a single chemical, one would expect the chemical composition of atmospheric PM to be complex, but it is not this simple. Each source of particulate matter produces a complex mixture of chemicals, for example, in a study by Samburova et al., 113 PAH compounds were measured using GC-MS in PM samples from burring peat, ponderosa pine needles, and cheatgrass. The investigators used the measured PAH concentrations to account for the UV-Vis absorbance of the sample extracts and found the 113 chemicals they measured

accounted for less than 0.2% of absorbance suggesting there are thousands of chemicals present.[29] In another similar study of biomass burning emissions over 2,000 unique chemical formulas were identified by high-resolution mass spectroscopy.[30] Each formula could represent several unique compounds, for example, Benzo[b]fluoranthene, Benzo[k]fluoranthene, and Benzo[a]pyrene (shown in Table 1-2) have the same molecular formula, $C_{20}H_{12}$. Determining the exact chemical makeup of atmospheric PM would be difficult, if not impossible, but it is not necessary to measure the exact composition to identify sources of PM. Methods to determine a subset of chemicals or elements in PM are used for identifying sources of air pollution using receptor modeling.

Table 1-2: EPA Priority PAH compounds. IARC Groups: 1- Carcinogenic to human beings; 2A– probably carcinogenic to human beings; 2B– Possibly carcinogenic to human beings.

PAH	IARC	Structure
Naphthalene (NA)		
Acenaphthene (AC)		
Acenapthylene (ACN)		
Fluorene (FL)		
Phenanthrene (PHE)		
Anthracene (AN)		
Fluoranthene (FA)		
Pyrene (PY)		
Benz[a]anthracene (BaA)	2B	
Chrysene (CHR)	2B	
Benzo[b] fluoranthene (BbF)	2B	
Benzo[k] fluoranthene (BkF)	2B	
Benzo[a]pyrene (BaP)	1	
Benzo[ghi]-perylene (BgP)		
Indeno[1,2,3-cd]-pyrene (IP)	2B	
Dibenz[a,h]-anthracene (DaA)	2A	

1.5.1 Polycyclic Aromatic Hydrocarbons

An important chemical component of PM generated by combustion processes is polycyclic aromatic hydrocarbons. PAHs consist of a series of fused benzene rings and many of these compounds are considered carcinogenic by the International Agency for Research on Cancer (IARC). PAH content of particulate matter is normally measured by extraction into a solvent followed by GC-MS (e.g. EPA Method TO-13A[31], European Standard EN 15549[32]). The US Environmental

Protection Agency designated 16 priority PAH pollutants, these compounds are shown in Table 1-2 along with their IARC classification as carcinogens.[33] Measuring the PAH content of a PM sample allows one to distinguish if the PM came from a combustion source or not. Additionally, the ratios of individual PAH species produced vary from source to source (e.g., traffic vs. non-traffic), so PAH content can be used to identify sources of PM.[32]

1.5.2 Elemental Analysis

Elemental analysis of aerosols is also a useful tool for identifying sources of PM. X-ray fluorescence (XRF) is one technique employed to measure the relative abundance of elements present in a PM sample. XRF is performed on a thin layer of atmospheric PM collected on a Teflon membrane filter (e.g. EPA Method IO-3.3[34]). The filter samples are bombarded with X-rays and the resulting emission spectra can be attributed to the known characteristic emission spectra of individual elements. The lower an element's atomic number, the lower its X-ray fluorescent yield. For this reason, lighter elements are more difficult to measure with XRF.[35] For example, the lightest element measured in a source apportionment study by Larson et al. was aluminum.[36] Carbon is a key component in PM air pollution from combustion sources so other techniques in addition to XRF are used to measure carbon content and may be included with elements measured by XRF for the purpose of source apportionment by receptor modeling.

1.5.3 Measurement of PM Carbon Content

An example of measurement of carbon content is the analysis of organic and elemental carbon ratios (OC/EC). Particulate matter emitted by combustion contains carbonaceous material in the form of a wide range of chemical compounds as well as graphitic material. The numerous chemical compounds are referred to as OC and the graphitic material is referred to as EC. Measurements of OC and EC are accomplished using thermal desorption of PM from quartz filters

(e.g. NIOSH method 5040[37]). OC is measured by thermal desorption in an inert atmosphere followed by thermal oxidation of EC accomplished by introducing oxygen. The utility of this non-specific measurement is that various sources have varying ratios of OC/EC and varying levels of total carbon per mass of PM. This non-specific chemical measurement method has been shown to be useful in source apportionment. In particular, measurement of OC/EC has been used to differentiate between gasoline and diesel vehicle emissions[38,39]

In addition to the thermal measurement of OC/EC content described above, there are optical measurements used to estimate PM carbon content. One such method is the integrating plate method (IPM). This method measures the amount of light transmitted through a filter before and after PM is collected to determine absorption by the PM. This measurement can be correlated to other types of carbon measurement.[40] For example, Larson et al. correlate IPM measurements of absorbance to EC carbon content measured by thermal desorption.[36]

1.5.4 Tracer Compounds

Some chemical compounds found in atmospheric PM come only from specific emission sources. These compounds may be used as tracers for certain sources of PM. Levoglucosan, a sugar anhydride, may be used as a tracer compound for wood smoke. Measurement of levoglucosan requires a multi-step sample preparation followed by GC-MS.[41] The disadvantage of using a tracer compound is that it only provides information about one source, but it is a powerful method when studying a single source like wood smoke.

1.6 Fluorescent EEM Spectroscopy

Another technique that provides chemical information is fluorescence spectroscopy. Although fluorescence does not quantify specific compounds, the signal obtained is underpinned by the

chemical composition of the sample and provides information that may be used for source apportionment. Fluorescence spectroscopy is a sensitive analytical technique with the ability to detect fluorescence from a single molecule using sophisticated instrumentation,[42] and with widely available benchtop fluorimeters, limits of detection are roughly 1 ng/mL for polycyclic aromatic hydrocarbons, a common chemical component of PM air pollution.[43,44] For this reason, fluorescence spectroscopy is an attractive analytical technique for PM analysis due to the typical sample sizes of PM being small. Although fluorescence is a very sensitive technique, it is not highly specific due to many analytes having overlapping signals. One method to increase the specificity of fluorescence spectroscopy is to collect fluorescent emission spectra at many excitation wavelengths, giving a 2D dataset or matrix of fluorescence intensities referred to as an Excitation – Emission Matrix (EEM).[45] Figure 1.2 is an illustration of how an EEM is assembled from individual spectra and represented graphically.

EEM spectroscopy has been widely applied to the analysis of complex environmental water samples[46] as well as analysis of atmospheric PM.[47–53] Mladenov et. al. suggested EEM could be useful as a source identification tool for atmospheric PM but did not evaluate the ability of EEM alone to identify sources.[51] Other work applying EEM to atmospheric aerosols discusses the chemical composition of various regions of fluorescence but does not discuss using EEM as a source apportionment or identification tool.[47–50,52,53] In this dissertation, I apply EEM spectroscopy to identify sources of air pollution.

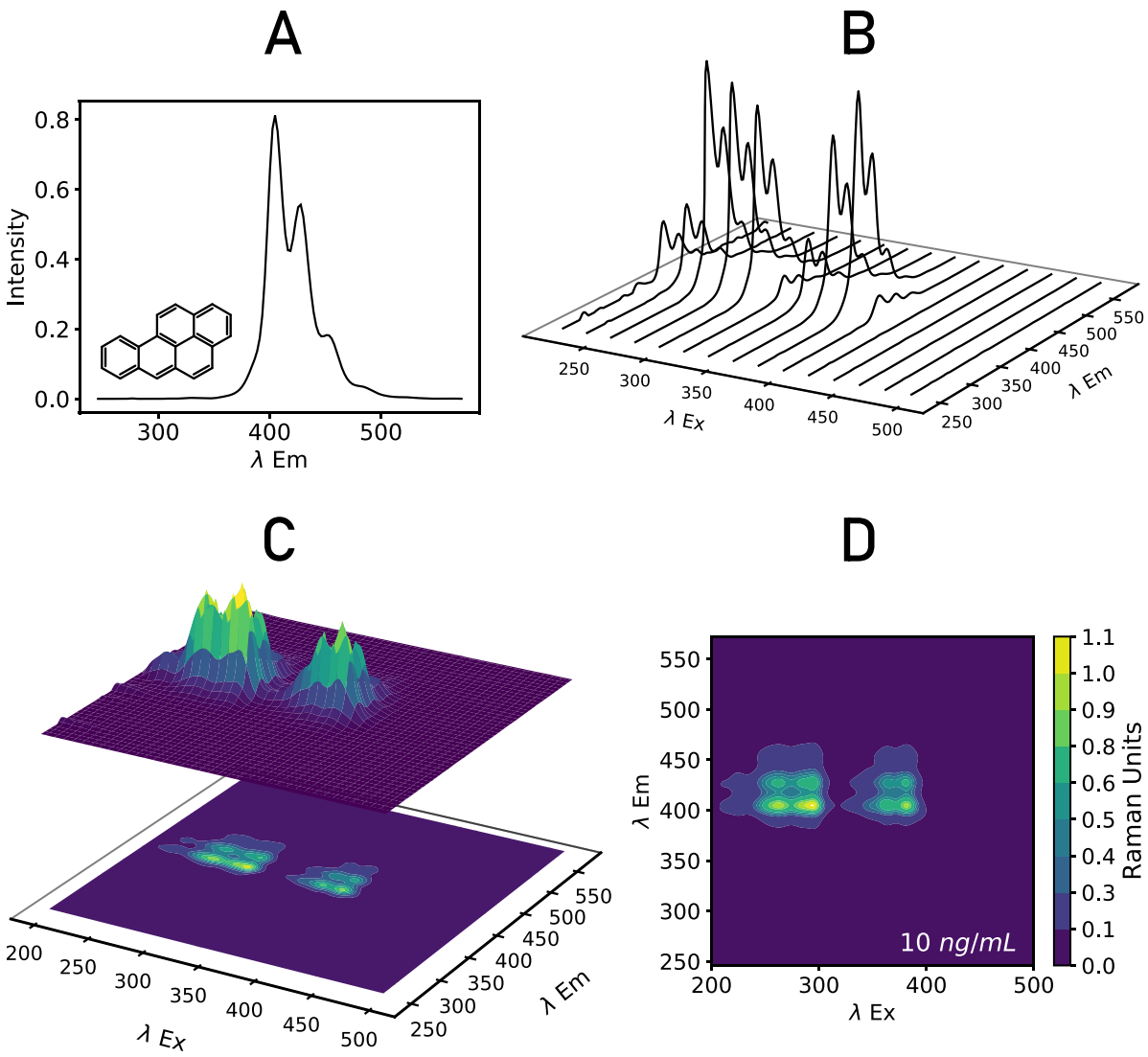


Figure 1.2: EEM spectroscopy. Individual fluorescence emission spectra are collected (a), the excitation wavelength is incremented, and additional spectra are collected (b). An EEM may be represented as a 3D surface (c) or as using a filled contour plot (d). The example EEM spectrum shown here is from Benzo[a]pyrene at 10 ng/mL in cyclohexane.

1.7 Fluorescent EEM Data Analysis

EEMs provide complex spectral information consisting of thousands of wavelength-dependent fluorescent intensities (~20,000 data points for the EEMs in this work), as such, a variety of approaches have been used to interpret EEM spectra. Fluorescent regional integration

is one example. This method considers specific regions of an EEM spectrum based on compounds of interest that display fluorescence in various regions.[48] This approach has the advantage of simplicity, but it is unable to distinguish overlapping spectra, which was a problem in this work. Other techniques that are useful for interpreting EEM spectra, can handle overlapping spectra, and have been applied to the analysis of atmospheric PM include partial least squares regression (PLS), parallel factor analysis (PARAFAC), principal component regression (PCR), and multivariate curve resolution (MCR).[43,47,50,54,55] Given that EEMs are 2-dimensional data similar to images, an alternative approach for interpretation of EEMs is using a convolutional neural network (CNN). CNNs are a machine-learning technique widely applied to image classification and many other fields including weather forecasting, natural language processing (i.e. written language) and speech recognition.[56–59] CNNs are used in many applications due to their flexibility to learn from the data provided.

1.8 Convolutional Neural Networks

A convolutional neural network is a series of mathematical operations used to convert an input, such as an image or an EEM spectrum, into a classification or regression output. There are many separate operations performed in a CNN referred to as layers. CNNs consist of varying combinations of layers depending on their application that include convolutional layers, pooling layers, and fully connected layers. CNNs are widely used in image classification and have achieved the best results in the annual competition ImageNet Large Scale Visual Recognition Challenge.[60] A well-known CNN that achieved winning results in the 2012 challenge is AlexNet, named after one of its creators.[61] The CNN architecture used in this work is similar to the architecture used in AlexNet.

The first layers in a CNN are convolutional layers. These layers consist of filters of

user-defined size which are iteratively scanned, or convolved, across the data. The filter is a matrix of values that is multiplied element-wise with the input data and summed giving a value in the output, called the feature map, at the corresponding location. Higher values in the output feature map correspond to the feature being present and negative values correspond to the inverse of the feature. The filter may be moved across the data one element at a time or in larger steps. The size of these steps is called the stride. Filters may be applied with or without adding zero padding around the edges of the input. If zero padding is added the input and output data shapes are the same, otherwise, the data size is reduced depending on the size of the filter.[62] An example of convolution with a 3-by-3 filter is shown in Figure 1.3.

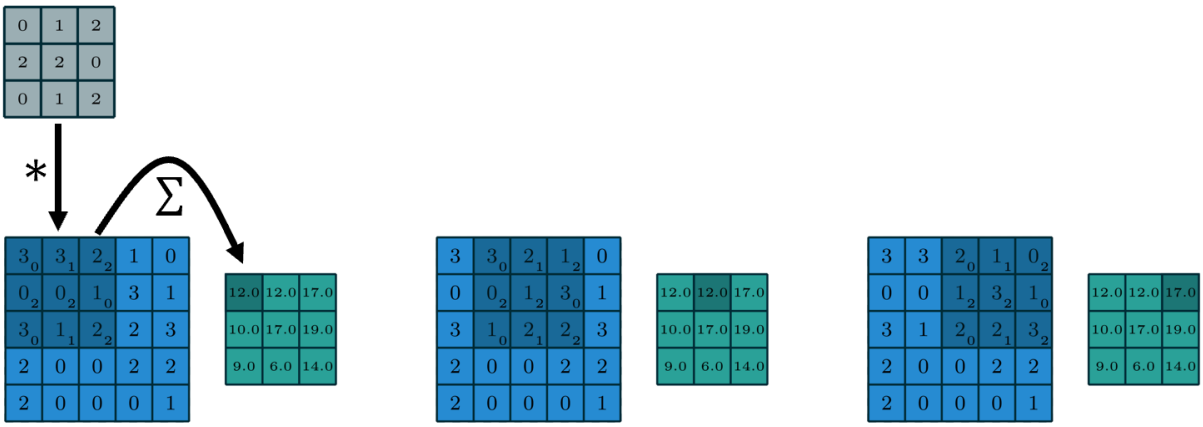


Figure 1.3: Example convolution with a 3-by-3 filter with a stride of 1 and no zero padding. Figure adapted from Dumoulin and Visin.[62]

The first convolutional layer in a CNN is used to recognize the lowest level features in the data. In the case of image recognition, this would be things like edges. Additional convolutional layers allow for the complexity of the represented features to build by looking for patterns of previous filters. For example, if an image of a bicycle was being identified by a CNN trained to identify bicycles the first convolutional layer would recognize edges, a second may recognize patterns of edges that resemble individual parts of the bike like a wheel, frame, and handlebar and

a third layer may recognize two wheels, a frame and a handlebar as a bicycle. In this way, a very large number of patterns can be represented and identified using CNNs. When applied to EEM data the convolution process can recognize peaks and valleys and patterns thereof.

It is common to have pooling layers between each convolutional layer. The purpose of pooling layers is to down-sample the data in order to reduce the size of data that must be processed in each layer as well as decrease the sensitivity to the exact location of a feature. A common method of this is called max pooling which is used in this work. Max pooling takes the maximum value in a region of data and uses that value in a new representation of the data that is reduced in dimension as illustrated in Figure 1.4.[62]

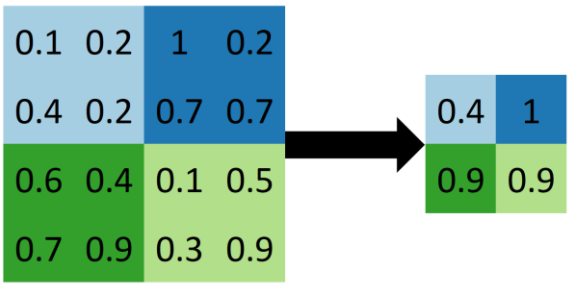


Figure 1.4: Example of max pooling with a 2-by-2 pool size.

The fully connected layers serve the purpose of translating the multidimensional feature map into a one-dimensional output such as classification or regression values. The fully connected layers could be an artificial neural network (ANN) or a deep neural network (DNN), this depends on how many layers the neural network has. The first and the last layer of a neural network are the input and output layers; all the layers in between are referred to as hidden layers. If a network has more than one hidden layer it is generally referred to as a deep neural network. The concept of the neural network was first proposed by Frank Rosenblatt in 1958 in his paper on the *perceptron*.[63]

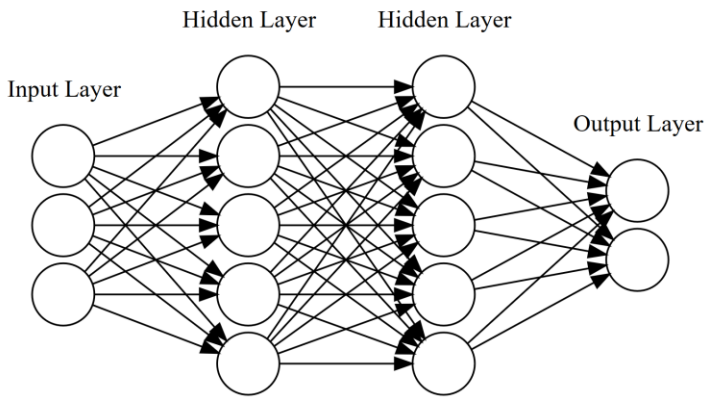


Figure 1.5: Simple neural network diagram with 2 hidden layers.

Each input node of a neural network, represented by a circle in Figure 1.5, takes in a value and passes it forward to every node of the hidden layer. As an input is passed forward to the next layer it is multiplied by a weight. All the values passed forward to each hidden node are summed together along with a bias term and this value is passed through an activation function to give the output of the hidden layer neuron. The outputs of each hidden layer neuron are passed forward in the same manner to the next hidden layer or to the output layer. The details of this process are described in The Elements of Statistical Learning by Hatsie and many other places in literature.[64]

If all the weights and biases are set correctly, the output of the neural network will give the correct value, but it is not possible to know what the correct values for these parameters are ahead of time, the parameters must be adjusted to get the desired output. The process of adjusting the weights and biases in the neural network is called *training*. The training process also allows for adjustment of the values used in the filters. To begin the training process filters, weights and biases are set using randomly generated values. A training dataset, consisting of inputs and desired outputs, is passed through the algorithm. Using the example of identifying a bicycle: an image of a bicycle would be the input and an output encoding bicycle would be a single training example. It is expected that on the first pass the network produces poor results. For example, it may assign

an image of a bicycle with a high probability of being a dog. The error in the output is captured using a cost function such that the cost is higher the more incorrect the result is. The incorrect output is corrected by minimizing the cost function through a process called backpropagation and gradient descent.

Backpropagation is the process of looking backward through the network and computing the gradient of how the cost changes in terms of all the parameters in the network. This is accomplished by repeated application of the chain rule, the specifics of which are widely explained in the literature.[64,65] Once the gradient of the cost function is calculated the parameters of the network are adjusted in the direction to reduce the cost function, referred to as gradient descent. Using all training data to update the network parameters one time is called a training epoch. It is not practical to calculate the gradient of the cost function with respect to all training examples simultaneously, so the training examples are broken up into batches and the gradient is calculated according to each batch and network parameters are updated after each batch. Training by this process is called stochastic gradient descent.

As with any model, overfitting is possible. To evaluate overfitting, a fraction of the training data are held back from the training process, this is called the test set. After training is complete the accuracy of the model predictions on the test set are evaluated. If the accuracy on the test set is poor compared to the training set this is an indication of overfitting. In some cases, not enough data are available to have a separate training and testing set. In this case, the dataset may be split into a given number of approximately equally sized parts. Each part, in turn, is held out as the test set, while the remaining parts are used for training. This approach, called cross-validation, is depicted graphically in Figure 1.6.

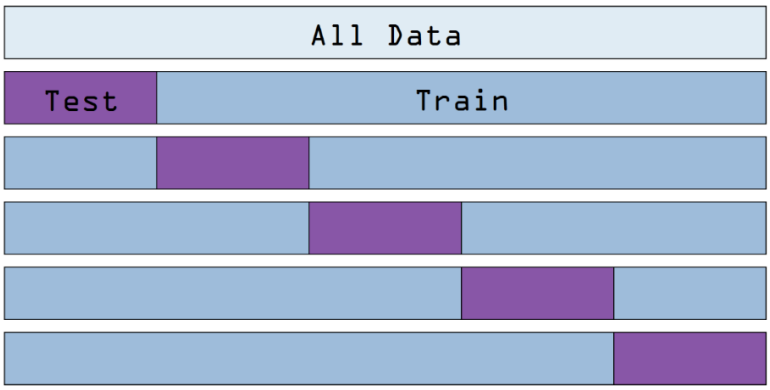


Figure 1.6: Graphical representation of 5-fold cross-validation. The dataset is split into five parts and 1/5th of the data are held back as the test set. The 1/5th of the data held back is rotated in each "fold" of the process.

Using a test-train split or cross-validation does not prevent model overfitting, so a strategy must be employed during the training process. One approach to prevent overfitting is called *early stopping*. This approach stops model training after a limited number of epochs when the accuracy on the test set begins to decrease while accuracy on the training set continues to increase. The difficulty of this approach is deciding when to stop. Additional details and difficulties of this approach are discussed in detail by Prechelt.[66] The approach I use in this work is called *dropout*. In this approach, a randomly selected subset of the weights in the network are left out of each training iteration. This helps prevent co-adaptation of neurons, which reduces overfitting. The details of *dropout* are explained by Hilton et al.[67]

Neural networks and CNNs have been applied to a variety of PM exposure related problems. For example, a CNN has been used to predict continuous $PM_{2.5}$ concentrations from discrete measurements.[68] Source apportionment has been conducted using a neural network with elemental composition as the input, which is a similar application to interpreting the EEM data in this work, but the input data was 1D elemental data so convolution layers are not used.[69] Neural networks

have been used to process EEM spectra for analysis of water samples, contaminates in olive oil, and antibiotics in urine.[70–73] In these examples, neural networks are used because of their ability to fit non-linear behavior without the need for assumptions about the underlying data and their ability to interpret information from the entire EEM, both of which are advantages for this work.

1.9 Principal Component Regression

PCR is a two-step process in which predictors are transformed into a lower-dimensional representation by principal component analysis (PCA) and linear regression is performed on a chosen number of the principal components (PCs).[74] In the case of the EEM data in this work, there are ~22,000 fluorescent values (predictors) in each data matrix. Linear regression could be performed using all values in the EEM independently but would result in overfitting. Instead, linear regression is performed on the PC representation of the data. The number of PCs used is smaller than the number of original predictors and is selected to capture nearly all the variance in the data, give the best possible model fit, and avoid overfitting.

1.10 Objectives

Particulate matter air pollution is the world's most significant air pollution problem. Current evidence clearly shows that increased $PM_{2.5}$ mass is associated with adverse health outcomes. There are some studies that suggest certain types of air pollution are worse for health than others, but the evidence for this is not conclusive. A significant challenge in studying the health effects of specific categories of PM pollution is source apportionment. Source apportionment adds cost and difficulty to air pollution studies because it requires measurements of PM composition in addition to mass measurements. New tools that reduce the barriers to performing source apportionment will enable continued improvement of air quality and study of source-specific health effects.

In this dissertation, I develop a method for source apportionment of combustion generated particulate matter using fluorescent excitation-emission matrix spectroscopy and machine learning. The method is intended as a new tool for source apportionment that reduces the difficulty of source apportionment of combustion generated PM.

The objectives of this dissertation include:

1. Show EEM can detect combustion generated PM at concentrations below the levels associated with adverse health effects,
2. Demonstrate EEM and machine learning can be used for source apportionment of controlled combustion sources in the laboratory,
3. Show that EEM and machine learning can be used for source apportionment of environmental PM samples using results obtained by orthogonal measurements.

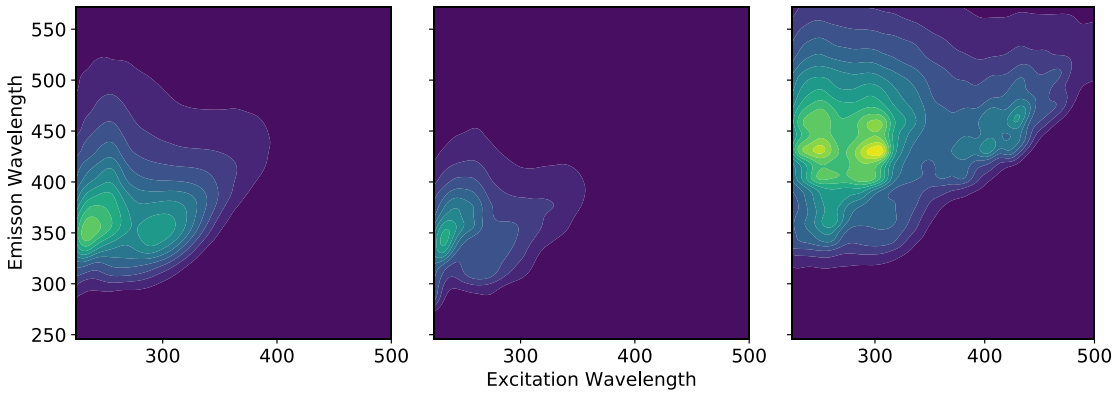
For the first objective, I have leveraged existing resources at the University of Washington (UW) to collect three common sources of particulate matter: diesel exhaust particulate from the UW diesel exposure chamber, wood smoke from the UW clean cookstoves laboratory and cigarette smoke from a smoking machine. $PM_{2.5}$ was collected and mass determined using standard procedures for PM collection. Using these samples, I developed a procedure for PM extraction and analysis by EEM spectroscopy. Using EEM spectra from the laboratory samples I determine the limit of detection of the EEM method for each of the sources. The procedures and results of this work are described in chapter 2.

To address the second objective, EEM spectra collected from the controlled sources in the laboratory were used to train various machine learning models. Mass measurements of the PM are used to quantify the concentration of each source in the extracts analyzed by EEM. Then the

ability of several machine learning models, and a simple linear model, are evaluated for source apportionment of the controlled PM sources using EEM spectra. A convolutional neural network is found to give the best results. The application of this method is discussed in detail and compared to the other modeling approaches in chapter 3.

I begin to address the third objective by applying the CNN trained on the laboratory samples to a small set of environmental samples. The lessons learned from this are described in chapter 3 and motivate the application of the EEM-machine learning method to a larger set of field samples that were collected as part of an exposure assessment panel study in Seattle WA. Source apportionment was conducted for these samples, and the results were previously published. Using these source apportionment results I show the ability of the EEM-machine learning approach to reproduce the source apportionment results for combustion sources in chapter 4.

Chapter 2: Excitation Emission Matrix Fluorescence Spectroscopy of Combustion Generated PM from Controlled Sources



2.1 Particulate Matter Sampling

I sampled $PM_{2.5}$ from cigarettes, diesel exhaust, and wood smoke using 2.0 µm pore PTFE membrane filters (Pall Zefluor®, Pall Cat. # P5PJ037) housed in Harvard School of Public Health Personal Exposure Monitor (BGI, Butler, NJ Cat. # HP2518) sampling cassettes. Filters were operated at a flowrate of 1.8 lpm using either portable or stationary vacuum pumps (AirChek XR5000 pump, SKC Inc., Eighty Four, PA or VP0625-V1014-D2-0511, Medo USA, Roselle, IL with a custom manifold of 9 VFB-65-BV roto-meters, Dwyer Instrument, Michigan City, IN). Flow rates were verified using an airflow calibrator (Gillian Gilibrator PN# 800268, Sensidyne, St. Petersburg, FL).

I collected wood smoke by burning 1 ½ by ¾ inch Douglas fir sticks cut from dimensional lumber in a side feed, natural draft, prototype improved cookstove. Sampling devices were placed in a sealed chamber connected to the exhaust hood duct at the sampling point described for gravimetric sampling by Sullivan et. al.[75] I collected diesel exhaust particulate from the exposure room at the UW's controlled inhalation diesel exhaust exposure facility.[76] The diesel PM is generated by a 435 cc direct-injection single-cylinder diesel engine fueled with ultra-low-sulfur diesel. I collected cigarette smoke either by lighting cigarettes in a fume hood and allowing them to smolder or from the exposure chamber of a cigarette smoking machine (Model TE-10B, Teague Enterprises, Woodland, CA). The TE-10B is a microprocessor-controlled machine that produces mainstream smoke mixed with sidestream smoke from filtered 3R4F research cigarettes (Tobacco Research Institute, University of Kentucky, Lexington, KY). Two cigarettes were puffed simultaneously for 2 seconds for a total of 8 puffs, at a flow rate of 1.05 l/min. The smoke collected represents approximately 10% mainstream and 90% sidestream to more closely resemble secondhand smoke.

2.2 Gravimetric Analysis

Following collection, filters are removed from the samplers and placed in a chamber with 37% (SD = 4%) relative humidity for 24 hours. The filters were then weighed using a micro-balance with 0.5 µg resolution (Mettler-Toledo UMT-2, Greifensee, Switzerland). Initial weights of each filter are recorded in the same manner and I use the difference to calculate the amount of $PM_{2.5}$ collected.

Figure 2.1: Photos of PM collection. From left to right: Laboratory sources of particulate matter (diesel exposure chamber, cigarettes, and clean cookstove), particulate matter collected on a PTFE filter, and PTFE filter in cyclohexane for extraction.

2.3 Filter Extraction

I place the filters into 20 mL glass vials (Cat # 89096-774 VWR, Edison, NJ), submerge them in cyclohexane, (Uvasol® Cyclohexane for Spectroscopy, MilliporeSigma Cat. #1.02822.2500) and sonicate for 30 minutes (40 kHz, 2510DTH Branson, Ultrasonic Corp., Danbury, CT). Filters were generally submerged in ~10 mL of cyclohexane to achieve an initial extract concentration of 5 µg PM/mL cyclohexane or greater. For filters with low PM loading, I cut the filter into fourths

to enable extraction in as little as 3 mL of cyclohexane to maintain extract concentrations at or above 5 µg/mL. Typically, the PM is not dislodged from the filter during extraction allowing for direct analysis of the extract. If enough PM was dislodged and suspended to cause turbidity, the extract was filtered with a 0.2 µm PTFE syringe filter (VWR Cat. #28145-491) before analysis.

2.4 EEM Collection

PM extracts were stored in 4 ml vials (Cat # 66009-876 VWR, Edison, NJ) until analysis. For EEM spectroscopy, ~ 3 ml of PM extract was transferred to a 1 cm x 1 cm quartz cuvette (Item # CV10Q3500FS, Thorlabs Inc., Newton, New Jersey). I collected EEM data using a fluorometer with an extended-UV 150W xenon-arc lamp (Aqualog-880-C, HORIBA Instruments Inc. Edison, New Jersey). I excited samples between 200 to 500 nm at 2 nm increments with an excitation slit width of 5 nm and recorded emission spectra between 246 and 826 nm on a CCD array. The CCD array has 1000 pixels each covering 0.58 nm. Data were collected using 4-pixel binning giving an effective emission slit width of 2.32 nm. I kept emission data between 246 and 572 nm and excitation data between 224 and 500 nm. Emission data above 572 was discarded because minimal fluorescence was observed above this wavelength and excitation wavelengths below 224 were removed due to low excitation lamp intensity between 200 and 224nm. The raw fluorescent signal is corrected for detector response and lamp intensity by the instrument[77] and is normalized to Raman Units.[78] Daily solvent blanks are recorded and used for blank subtraction to minimize the effect of Rayleigh and Raman scattering. To further reduce the effects of Rayleigh scatter values within 10 nm of the first and second-order Rayleigh scattering bands were excised followed by replacement of the values using 2-dimensional interpolation.[79] I did not correct for the inner filter effect because I observed absorbance below 0.2 for my samples.

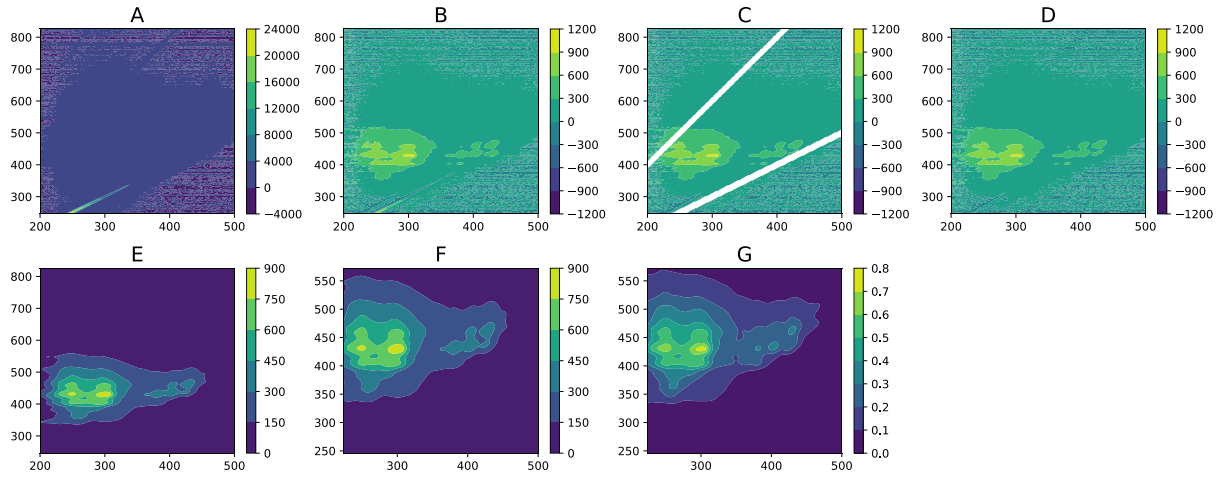


Figure 2.2: EEM of wood smoke extract showing data processing steps: Starting with the raw data (a), the solvent blank is subtracted (b), then values are excised for scatter removal (c), the excised values are interpolated (d), negative values are replace with zero and 2D-gaussian smoothing is applied (e), the EEM is cropped (f) and normalized to Raman units (g)

PM extracts from cigarette, wood smoke, and diesel show unique EEM spectra, as shown in Figure 2.3. Cigarette spectra consist of two peaks at ~350 nm emission wavelength. Diesel spectra consist of a single primary peak, also located at ~350 nm emission with less fluorescence surrounding the peak than cigarette. Wood smoke has spectra consisting of six peaks in the region from 400-475nm emission and 225 – 275 nm excitation. Cigarette and wood smoke have similar maximum fluorescent intensity levels per mass of PM while diesel has a lower intensity. Wood smoke shows fluorescence over the broadest region and generally at higher emission wavelengths than cigarette smoke and diesel exhaust PM samples. The spectra from the different sources have overlapping regions so there can be some difficulty in identifying individual sources from mixed samples.

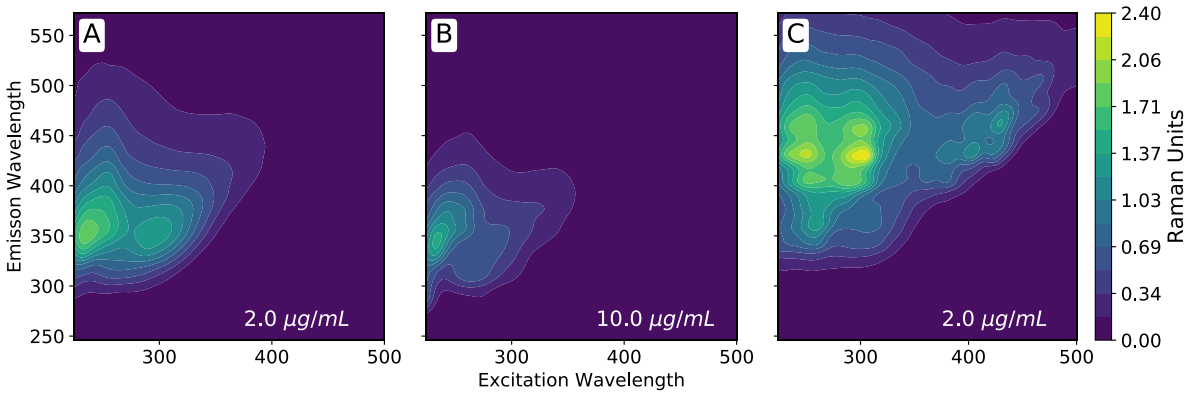


Figure 2.3: Fluorescence EEM spectra of laboratory PM sources. (a) cigarette smoke at an extract concentration of 2 µg/ml, (b) diesel soot at 10 µg/ml, and (c) wood smoke from a clean cookstove at 2 µg/ml. The three sources have been extracted in cyclohexane and exhibit unique spectral fingerprints. Cigarette and wood smoke have similar maximum fluorescent intensity on a per mass basis while diesel has a lower signal intensity.

2.5 EEM Data

I collected a total of 37 filter samples and used the extracts from these filters to create 113 samples for EEM analysis. They consisted of 81 single-source samples diluted to concentrations between 0.2 µg/mL and 10 µg/mL, 21 mixtures of the single-source samples, and five samples from filters with mixed PM from serial sampling of the sources. I also collected six spectra from liquid extracts of filters that were loaded into sampling devices and weighed, but no air was drawn through the filters (method blanks). Data augmentation was applied to 12 of the 113 EEM samples to create the training data leaving 101 samples for testing the algorithm. Table 2-1 summarizes the total number of each type of sample used for training and testing.

Table 2-1: Number of unique filter samples and liquid extract samples generated from extracts, dilutions of extracts and mixtures of extracts for each category of sample. I collected a total of 113 EEM spectra. Twelve of these spectra were used to generate training data leaving 101 spectra in the test set.

Sample Type	Number of Unique Filters	Samples for EEM	Spectra used for training
Cigarette	9	26	4
Diesel	10	29	4
Wood smoke	9	26	4
Extract Mixtures	N/A*	21	0
Multiple-Exposure	5	5	0
Method Blanks	4	6	0
Total	**37**	**113**	**12**

*Mixtures of Cigarette, Diesel, and Wood smoke samples

2.6 Limit of Detection

Using the integrated fluorescent intensity, the limit of detection (LoD) for each source was determined using the Clinical and Laboratory Standards Institute method for determining LoD.[80,81] This procedure accounts for variation present in blank measurements and measurements of low levels of the analyte to assign a LoD that represents 95% confidence in differentiating a low concentration sample from a blank and visa-versa. First, the limit of blank (LoB) was determined as,

$$LoB = \mu_B + 1.645\sigma_B, \qquad \text{(Equation 2-1)}$$

where μ_B is the mean and σ_B is the standard deviation of the integrated fluorescent intensity for the method-blank samples. This method assumes measurements are normally distributed. 1.645

is the z-score for which 95% of values in the standard normal distribution are below. Next, dilutions of extracts at low levels were analyzed. Extract concentrations were 0.5, 1.0, and 2.0 µg/mL for cigarette and wood smoke and 1.0, 2.0 and 3.0 for diesel. From these low-level measurements, the LoD was calculated as,

$$LoD = LoB + 1.645/\left(1 - \frac{1}{4f}\right) \cdot \sigma_S, \qquad \text{(Equation 2-2)}$$

where σ_S is the average standard deviation of integrated fluorescent intensity for the low-level sample measurements and f is the degrees of freedom calculated as the number of low-level samples analyzed minus one. For cigarette and wood smoke, σ_S was calculated from the 0.5, 1.0 and 2.0 µg/mL samples and for diesel from the 1.0, 2.0 and 3.0 µg/mL samples. Figure 2.4 shows error bars calculated as $1.645/\left(1 - \frac{1}{4f}\right) \cdot \sigma_S$ for all the low-level samples measured. Graphically, the LoD is where these error bars intersect with the LoB which is shown as a black horizontal dashed line in Figure 2.4. The LoD in terms of actual concentration (shown on the x-axis) is determined by finding the value at which the line of best fit (shown with a dotted line in Figure 2.4) intersects the value of LoD. The LoD, as reported in Table 2-2, is shown in Figure 2.4 by the red dashed vertical line.

.

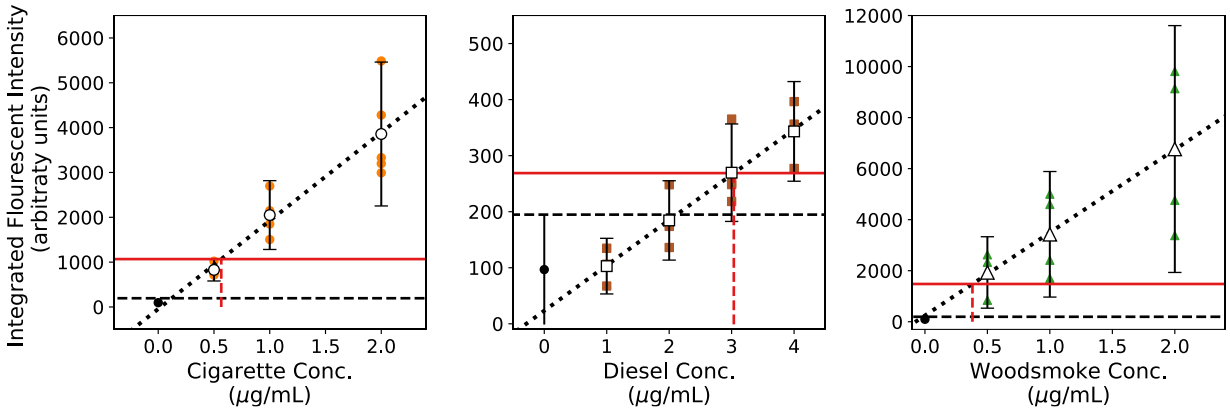


Figure 2.4: Plots showing data used to determine LoD. Samples containing a single source were used to determine the LoD. The LoB (Equation 2-1) is shown by the black horizontal dashed line, the LoD (Equation 2-2) on the integrated fluorescence intensity axis (y-axis) is shown by the red horizontal line. The LoD on the concentration axis is shown by the vertical red dashed line. These values (reported in Table 2-2) are determined using the line of best fit which is shown as a dotted line.

The measured LoD for each source is provided in Table 2-2 in mass of particulate matter per volume of extraction liquid as well as a calculated particulate matter concentration per volume of air sampled. The LoDs in units of volume of air sampled were calculated assuming a sampling time of 24 hours, an air sampling rate of 1.8 liters/min, and an extraction volume of 3 ml. An example conversion is shown below:

$$1\frac{\mu g}{ml\ solvent} * \frac{3\ ml\ solvent}{sample} * \frac{sample}{24\ hours} * \frac{1\ hour}{60\ mins} * \frac{1\ min}{1.8\ liters\ air} * \frac{1000\ liters}{1\ m^3} = 1.16\frac{\mu g}{m^3}$$

Diesel has the weakest fluorescence intensity and thus the highest LoD of 3.0 µg/mL cyclohexane or 3.5 µg/m^3 air. The LOD for each source in a 24-hour sampling period is significantly lower than the WHO and US EPA 24 hour mean exposure guidelines of 25 and 35 µg/m^3 respectively. This shows the LOD of this method is low enough to provide meaningful results at exposure levels

expected to be present in typical indoor and outdoor environments.

Table 2-2 LoD for samples containing pure sources. The column reporting LOD in µg/mL is determined using PM mass measurement of filters dispersed in a volume of cyclohexane. The column reporting µg/m^3 in air is determined by converting the LOD in µg/mL to µg/m^3 assuming a 24-hour sampling time at an air sampling rate of 1.8 liters per minute.

Source	LOD [µg/mL cyclohexane]	LOD [µg/m^3 air]
Cigarette	0.6	0.7
Diesel	3	3.5
Wood smoke	0.4	0.4

2.7 Extraction in Alternative Solvents

Extracts in various solvents were measured by dividing filters into quarters using a scalpel. I compared spectra from extracts in ultrapure water, methanol, and cyclohexane. Results were similar for methanol and cyclohexane while water extracts showed different spectra. Methanol and cyclohexane extracts were analyzed at lower concentrations (5ug/mL) compared to water extracts (25ug/mL and 10 ug/mL for cigarette and wood smoke respectively). Accounting for the concentration difference between the extracts, the cyclohexane extracts had the highest signal intensity. The goal of our work is to identify combustion sources of PM, so cyclohexane is a good solvent choice given the observed signal intensity and the fact that combustion products are expected to contain non-polar fluorophores such as PAHs.

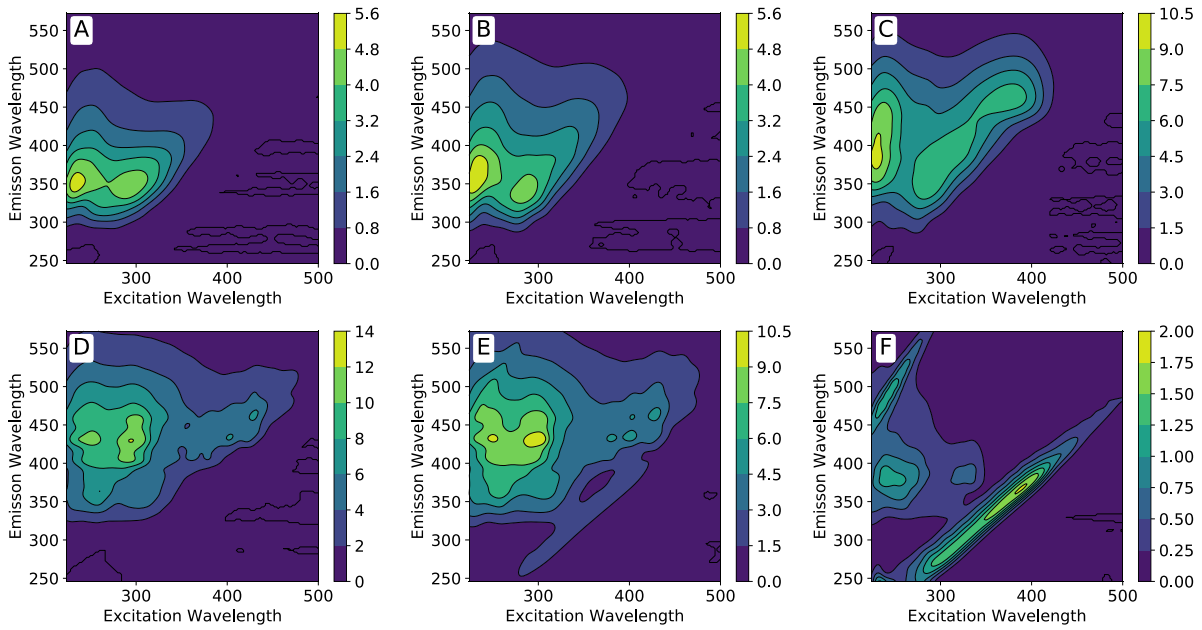
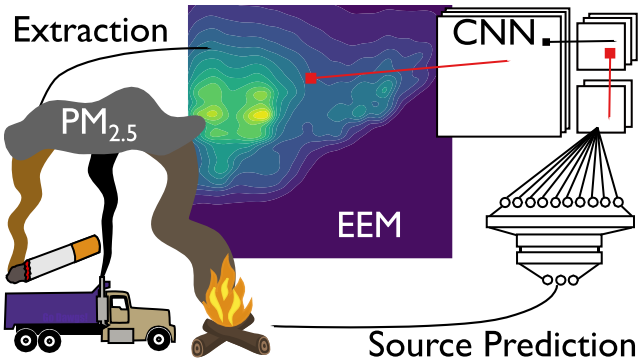


Figure 2.5: EEMs of in various solvents. Cigarette in cyclohexane at 5 µg/mL (A), in methanol at 5 µg/mL (B), and in ultra-pure water at 25 µg/mL (C). Wood smoke in cyclohexane at 5 µg/mL (D), in methanol at 5 µg/mL (E), and in ultra-pure water at 10 µg/mL (F). Scale bars are in Raman units.

Chapter 3: Source Apportionment Using EEM Data from Controlled Sources and Machine Learning



3.1 Training Dataset and Data Augmentation

A training dataset consisting of 6,375 spectra was generated by linearly scaling and mathematically combining spectra from 12 unique samples (4 from each of the 3 sources, cigarette, diesel and wood smoke). The process for generating spectra is shown schematically in Figure 3.1. To simulate noise and variability I average the four spectra from each source together with a randomly weighted average to produce a prototypical spectrum for each source (Figure 3.1A to C). The prototypical spectra for the three sources are scaled to the desired concentration and combined to generate a training spectrum (Figure 3.1D to G). For each spectrum generated the entire process is repeated in a loop to simulate variability associated with PM sampling, extraction and EEM collection.

I employed two sampling strategies to create an augmented training dataset. First, I created 1000 spectra for each of the three sources containing only one source in a linearly spaced concentration range from 0 to 5 µg/mL, resulting in 3000 spectra. Then I created digital mixtures of the three sources in a logarithmically spaced concentration range from 0.01 to 6.3 in fifteen steps (15^3 combinations) giving 3375 training spectra consisting of mixtures.

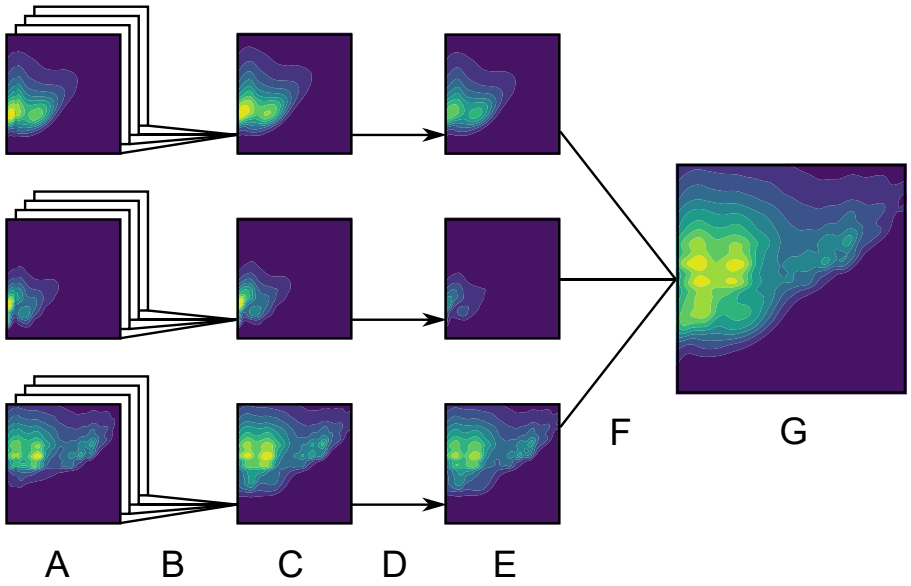


Figure 3.1: Graphical representation of the loop used to generate training spectra. Twelve original spectra, four from each of cigarette, diesel, and wood smoke, (A) are averaged using random weights (B) producing prototypical spectra for each source (C) that are scaled (D) giving single source spectra at target concentration (E) that are combined (F) to generate a training spectra (G). This loop is repeated from the beginning for each training spectra generated in order to simulate variability associated with PM sampling and EEM collection.

The data augmentation process assumes that the Beer–Lambert law applies to absorbance, constant fluorescence quantum yield, and negligible inner filter and matrix effects.[82] I tested these assumptions in 2 ways. First, I analyzed a series of dilutions made using extracts from single filters. These dilution series are plotted in Figure 3.2, showing fluorescence is indeed proportional to concentration as assumed when mathematically scaling the spectra. Next, to test that matrix effects due to mixing multiple sources together are minimal I compared liquid and digital mixtures of the 3 sources to confirm that liquid mixture and the digital mixture gave similar results. The results of this analysis are shown in Figure 3.3.

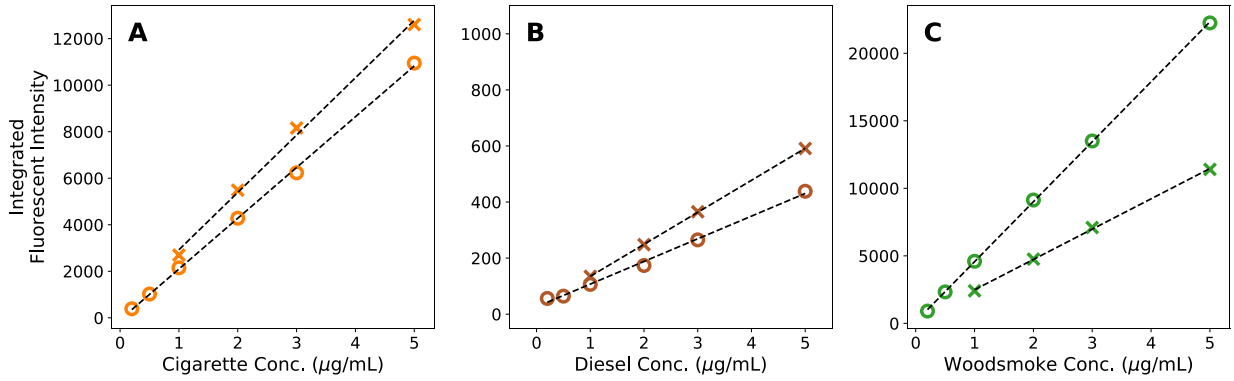


Figure 3.2: Integrated fluorescence intensity for single-source spectra vs. PM extract concentration. Two dilution series are shown for each source (marked with circles and X's) for cigarette (a) diesel (b) and wood smoke (c). The R^2 values for linear fit to these dilution series are all above 0.998. These high R^2 values support the method of linearly scaling fluorescent intensity when generating training data.

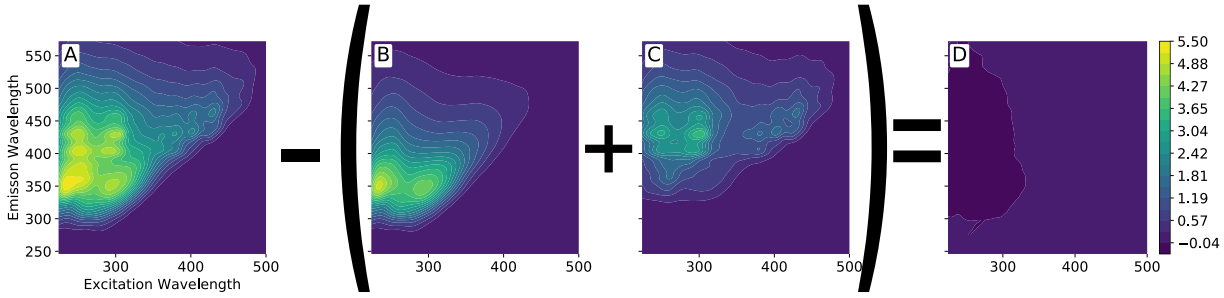


Figure 3.3: Comparison of liquid and digital mixtures. (a) shows a liquid mixture of cigarette and wood smoke extracts each at 5 µg/mL, (b) and (c) show cigarette and wood smokes extracts respectively also at 5 µg/mL, and (d) is the result of subtracting (b) and (c) from (a) showing nearly zero remaining signal which illustrates that matrix effects due to mixing are minimal.

Although fluorescence intensity scales linearly with concentration for a single filter, there is significant variation in fluorescence from filter to filter for a given mass concentration. The coefficient of variation for samples at 5 µg/mL is 0.09, 0.27, and 0.66 for cigarette, diesel, and wood smoke. For this reason, if a single linear line is fit to all single source spectra for each source, R^2 values of 0.97, 0.84, and 0.69 for cigarette, diesel, and wood smoke, respectively are obtained (Figure 3.4). Possible sources of this variation include the variability in source concentration in the air sampled as well as differences in combustion conditions from sample to sample. For example, I have observed that the fluorescent intensity from a filter decreases when HEPA filtered air is passed through the filter after sampling the source.

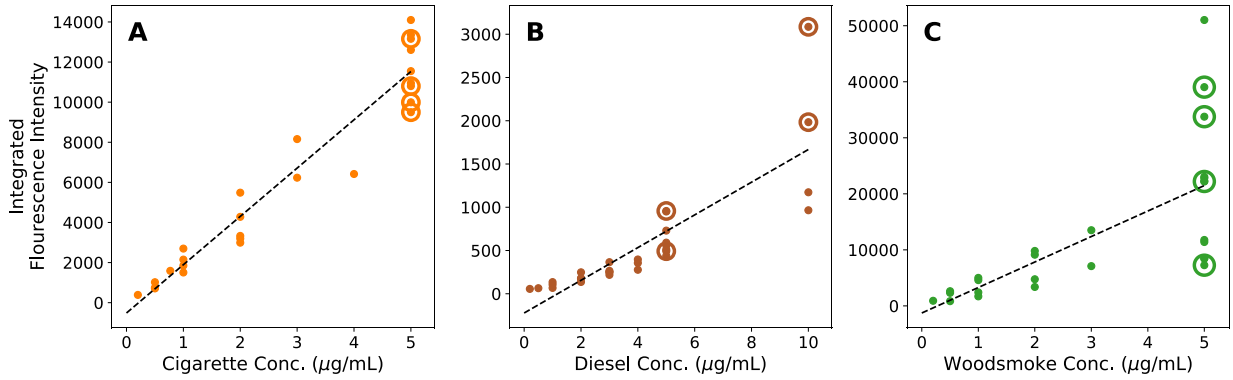


Figure 3.4: Soot mass vs. integrated fluorescence intensity. Cigarette (A) and wood smoke (C) have higher fluorescence per mass than diesel soot (B). Samples from multiple filters show a positive correlation with concentration, r^2=0.97 for cigarette (26 spectra), 0.84 for diesel (29 spectra) and 0.69 for wood smoke (26 spectra). The four points in each plot with a circle around them are the samples used for generating the training data.

Figure 3.4 shows the integrated fluorescent intensity of single-source spectra for cigarette, diesel, and wood smoke. The twelve spectra chosen for the data augmentation process are denoted by the circled points. These spectra were chosen so that the samples used to generate the training

data are not present in any samples in the test data. This means the samples used for data augmentation are not present in any dilutions or mixtures. Table 2-1 summarizes the number and sources of spectra in the testing and training datasets.

The four cigarette spectra used for data augmentation lay roughly on the line of best fit for all cigarette spectra (Figure 3.4A), for diesel three of the calibration spectra lie roughly on the line of best fit and one has above average signal intensity (Figure 3.4B) and for wood smoke, two calibration spectra are above the line of best fit, one below and one near (Figure 3.4C). The relative intensity of the spectra chosen for training influences the model to over or underpredict the test spectra. For example, the wood smoke calibration spectra are above average fluorescent intensity in total, so the model is expected to underpredict the concentrations of wood smoke in the test data.

3.2 Convolutional Neural Network

EEMs are 2D spatial data made up of combinations of peaks and valleys, which correlate to a particular chemical or combination of chemicals that are extracted from the PM samples. These peaks and valleys vary in their intensity across emission and excitation dimensions. The convolution filters learn to fit these varying shapes to better detect peak presence as they are iteratively applied over the EEM. For example, one filter may identify broad peaks, as seen in the EEM of cigarette (Figure 2.3A), while another may identify sets of narrower peaks as are characteristic of a wood smoke spectrum (Figure 2.3C). Subsequent convolutional layers are used to identify patterns of lower-level features, for example, a second convolutional layer may look for a group of narrow peaks identified in the previous convolutional layer. The results of the convolutional layers are feature maps showing the presence or absence of features. The feature map data are fed into fully connected layers that map this information to the desired output: predictions of concentrations for each source.

3.2.1 Network Architecture

The CNN used in this work consists of three convolutional layers, each followed by max-pooling,[62] as shown in Figure 3.5. All convolutions are performed using padding, so the dimensions of input and output data are the same.[62] The first convolutional layer contains twenty 5-by-5 filters equating to 11.6 nm in emission (height) and 10 nm in excitation (width), as shown by the red box (Figure 3.5a). This is followed by 3-by-3 max-pooling, which reduces the data from 143-by-139 to 47-by-46. The second convolutional layer is ten 10-by-10 filters followed by 3-by-3 max-pooling. The final convolutional layer applies ten 15-by-15 filters to the 15-by-15 feature maps. The output of the third convolutional layer is max pooled to a size of 5-by-5 and then flattened and connected to a dense neural network with three hidden layers having 512, 256, and 256 nodes in each layer, respectively. A dropout rate of 20% is used between all convolutional and fully connected layers.[83] The exponential linear unit was used as the activation function for all convolutional and hidden layers and a linear activation function was used for the output layer,[84] the loss function was the mean-squared-error.[64] The results described are from a network that was trained for 80 epochs. Details of how I selected the training duration are discussed in section 3.2.2. The CNN was implemented in Python 3 using Keras[85] and TensorFlow.[86]

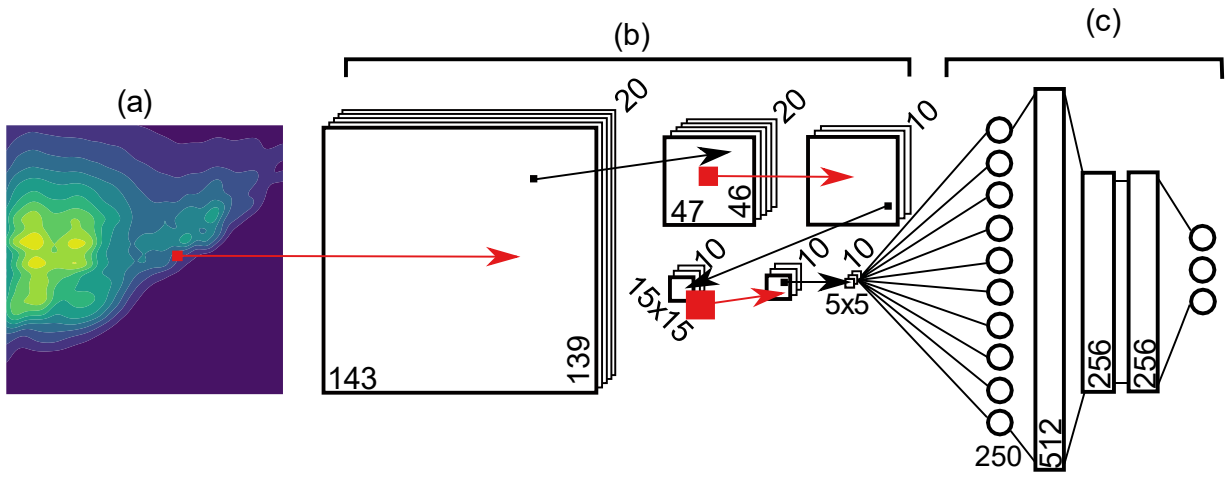


Figure 3.5: CNN Network Diagram. (a) Input spectra are first convolved with twenty 5-by-5 filters. (b) convolutional layers and max pooling layers are shown with associated data shapes. Convolutions (10x10 followed by 15x15) are shown in red and 3-by-3 max pooling is shown in black. Filters, data, and pooling sizes are shown to scale. (c) Output of the convolutional layers is flattened to a shape of 250 by 1 and fed into fully connected layers resulting in 3 output values (not to scale).

3.2.2 CNN Training Details

The CNN was trained for a total of 150 epochs. At each epoch, I saved the model parameters and prediction results. At the end of the 150 epochs of training, I selected the model with the best overall classification accuracy (see section 3.2.6 for discussion of classification accuracy). The results presented in this thesis are for a single model training run. The model selected was trained for 80 epochs to achieve the reported results. I repeated the training process a total of three times and each time the overall accuracy reached a maximum of 89%. The other two model training runs reached the same maximum accuracy at 73 and 85 epochs.

In Figure 3.6 I plot overall accuracy vs. epoch with a solid blue line and show the individual accuracies of cigarette, diesel, and wood smoke with circles, squares, and triangles, respectively.

In just a few epochs the model learns to accurately classify cigarette and wood smoke, but it takes much longer to improve accuracy on diesel. The diesel classification accuracy varies more over each epoch than the others due to the lower signal intensity of diesel and the resulting higher limit of detection. Figure 3.6B plots loss vs. epoch, using the mean-squared-error as the loss function. The training loss is noisy due to the small size of the training data set. This plot shows that the test loss reaches a minimum near 80 epochs suggesting that training beyond this point results in overfitting.

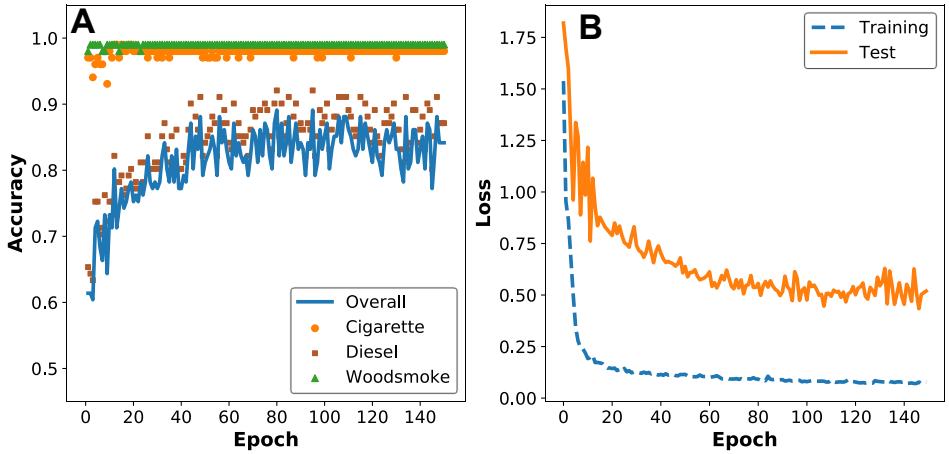


Figure 3.6. Plots of the metrics used for evaluating the CNN vs. training epochs. (A) Plot of classification accuracy as the CNN is trained for 150 epochs. Overall accuracy is shown with a solid blue line and the individual accuracies for cigarette, diesel, and wood smoke are shown with orange circles, brown squares, and green triangles, respectively. I selected the model trained for 80 epochs based on it having the highest overall classification accuracy of 89%. (B) Plot of model loss as the CNN is trained for 150 epochs. A leveling of the test loss at about 80 epochs suggests that training beyond this point results in overfitting.

3.2.3 Evaluation of CNN Architecture

The CNN architecture described in section 3.2.1 was the end result of trying many different architectures and comparing their performance. Figure 3-7 shows an alternative architecture that gave similar performance to the primary network architecture shown in Figure 3.5. The main difference between the two network architectures is the size of the first filter. In the alternative network, the first filter is larger (20-by-20) than in the primary network (5-by-5).

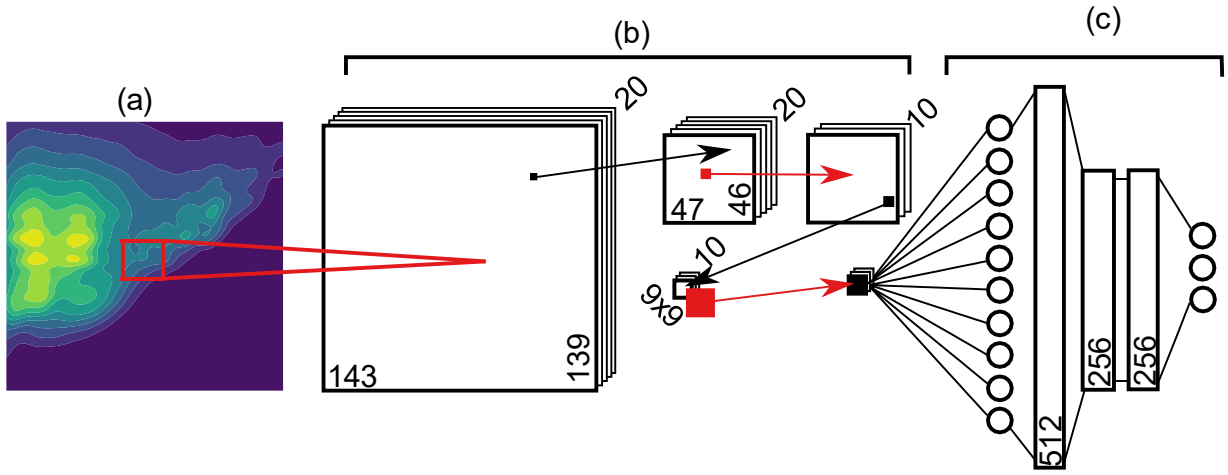


Figure 3-7: Alternative architecture CNN Network Diagram. (a) Input spectra are first convoluted with twenty 20-by-20 filters. (b) convolutional layers and max pooling layers are shown with associated data shapes. Convolutions are shown in red and max-pooling is shown in black. Filters, data, and pooling sizes are shown to scale. (c) Output of the convolutional layers is flattened to a shape of 10-by-1 and fed into fully connected layers followed by the 3-by- output layer (not to scale).

The alternative architecture with the large first filter was of interest because when the first filters of the network were visualized, they resembled peaks and valleys as shown in Figure 3.8. Filters that resemble peaks and valleys make intuitive sense, but what is occurring in the remainder

of the network is not accounted for by visualizing these filters. An alternative approach to model interpretation is needed to better understand the inner workings of the entire network.

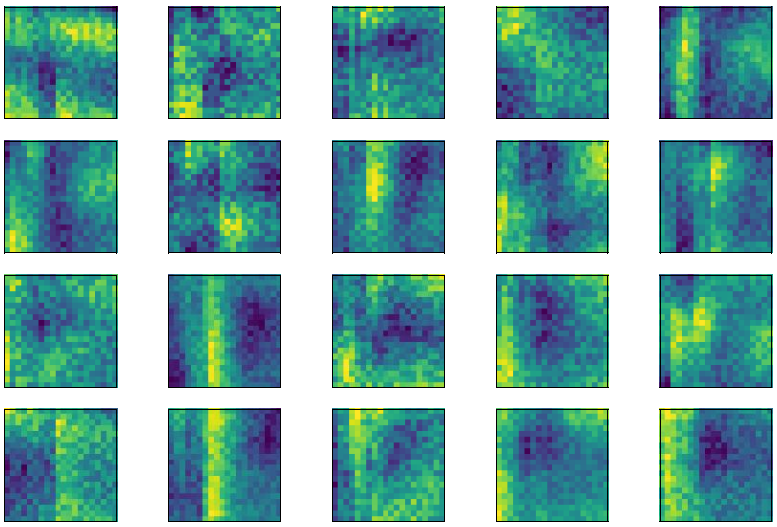


Figure 3.8: Visualization of first layer filters from the alternative network architecture. These 20-by-20 filters resemble peaks and valleys of different sizes and shapes.

To provide insight into what features of the input spectra the entire CNN is using to identify the sources of pollution I used a technique called saliency maps. Saliency maps estimate the sensitivity of the CNN output with respect to each area of an input spectra by calculating the derivative of the network output with respect to each input value. It is assumed when the value of this derivative is high at one location relative to another location, the location with a higher value is more important in determining the output. Tools for computing saliency maps are available in the literature, in this work I computed saliency maps using `SmoothGrad`, an open-source Python package.[87]

Saliency maps are typically applied to image classification, for example in an image containing a dog and a soccer ball, one expects the dog to be highlighted in the saliency map if the image is identified as a dog. In the case of EEM spectra containing mixtures of cigarette, diesel

and wood smoke, it is expected that areas corresponding to the fluorescence of each source will be important for estimating the respective concentrations.

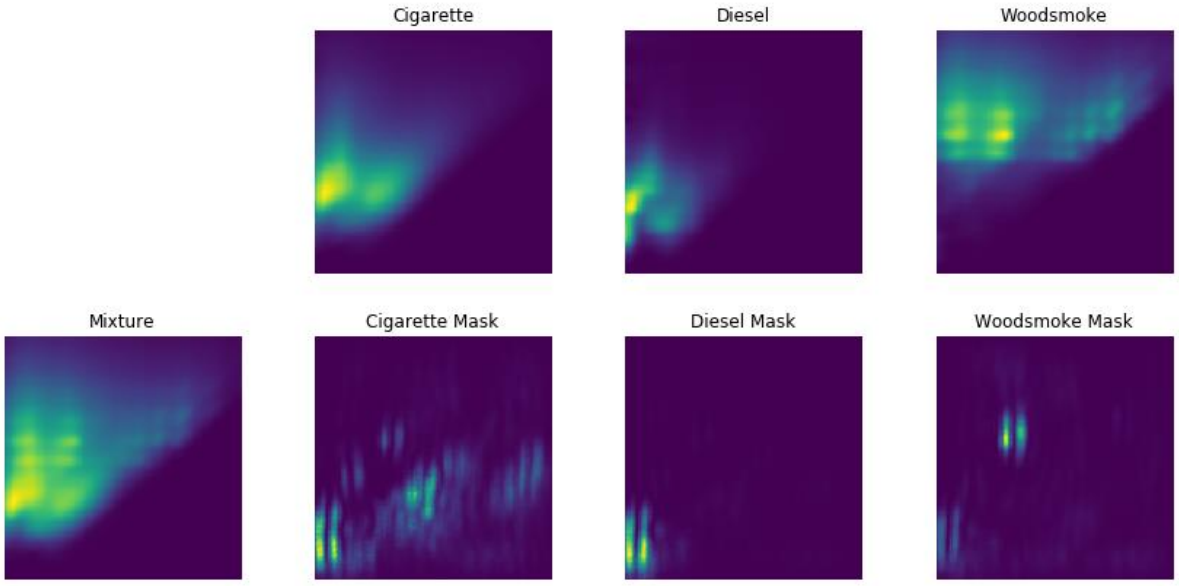


Figure 3.9 Saliency Masks for the alterative architecture CNN. EEM Spectra from cigarette smoke, diesel exhaust, and wood smoke are shown in the top row for reference. Saliency masks computed for the mixture shown at the far left are shown in the second row. The areas highlighted by the saliency maps show an unexpected band like structure suggesting the alternative network architecture may not be learning the most relevant features to identify the PM sources.

When I computed saliency maps for a mixture using the alternative CNN architecture (Figure 3-7) the results were different than expected. Instead of highlighting areas that match with areas of fluorescence shown in the original spectra, the saliency maps show a pattern of vertical bands. This unexpected result shows that although visualization of the first layer filters of this network suggested it was recognizing the expected features of the spectra, the network as a whole was not working as expected.

I computed saliency maps, in the same manner, using the primary architecture CNN (Figure 3.5). The resulting saliency maps are shown in Figure 3.10. These maps show the CNN is looking at areas of the mixture spectrum where the pure source spectra show their unique fingerprints giving me confidence the primary architecture CNN is working properly. The results discussed in all other sections of this dissertation refer to the primary architecture CNN (Figure 3.5).

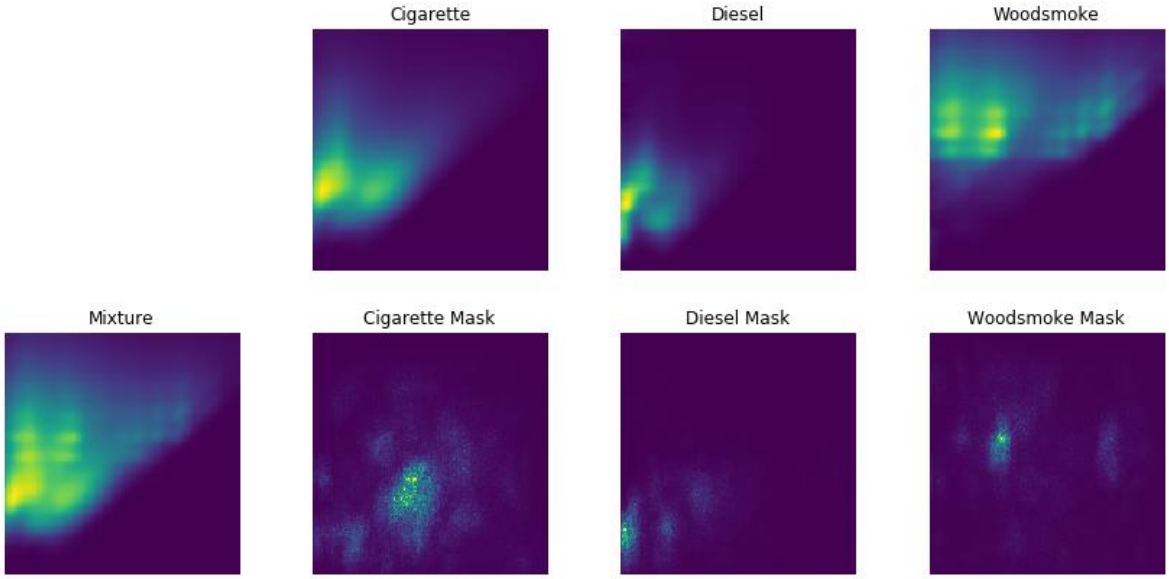


Figure 3.10 Saliency masks for the primary architecture CNN. EEM Spectra from cigarette smoke, diesel exhaust, and wood smoke are shown in the top row for reference. An EEM spectrum containing a mixture of the three PM sources is shown on the far left. Saliency masks computed for the mixture are shown in the second row. The areas highlighted by the saliency maps correspond with areas of unique fluorescence in the spectra of the underlying sources suggesting the CNN is learning relevant features to identify the PM sources.

3.2.4 Source Apportionment Results

The CNN was trained for 80 epochs using the augmented dataset described in 3.1. Figure 3.11 plots the prediction results for the CNN on both the training and test data in parity plots.

Prediction results on the training data are shown as light blue points. The diagonal line represents the perfect prediction of the samples where the CNN prediction values are equal to the values provided during training. The data points that result from the analysis of the original training data roughly follow the diagonal. The R^2 value for the fit to the training data for cigarette, diesel, and wood smoke are 0.99, 0.97, 0.97 respectively. One reason for scatter in the training data is extracts at the same particulate matter concentration show variation in fluorescent signal strengths as discussed in section 3.1 and shown in Figure 3.4.

I then predict the concentration of the 101 test spectra which are shown in the parity plots of Figure 3.11. The results generally follow the diagonal trend, but there are significant under- and over-predictions. This can be attributed to the fact that total fluorescent intensity from a given source varies from sample to sample at the same concentration. The R^2 value for the fit to the test data for cigarette is 0.86, for diesel, it is 0.79, and for wood smoke, it is 0.89. The lower R^2 values for the test data are due to the overlap of the signals (Figure 2.3) that makes mixtures difficult to quantify and variation in fluorescent signal intensity among samples at the same concentration (Figure 3.4).

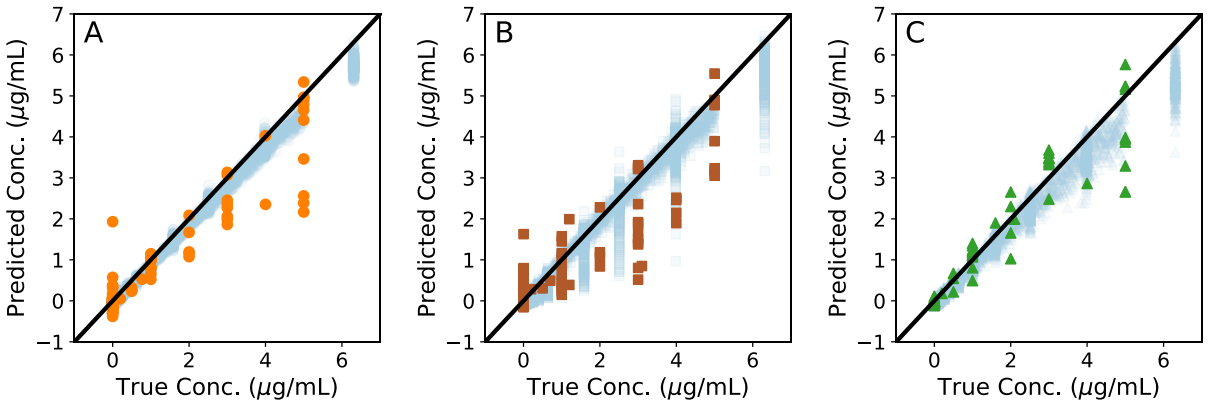


Figure 3.11: Parity plots showing predicted concentration vs. true extract concentrations for (a) cigarette (R^2 training = 0.99, R^2 test = 0.86), (b) diesel (R^2 training = 0.96, R^2 test = 0.79) and (c) wood smoke (R^2 training = 0.97, R^2 test = 0.89). The data points shown as solid colors (orange, brown and green) are from 101 test spectra, the points shown in light blue are the results for the augmented training data.

3.2.5 Limit of Detection for the CNN Model

I repeated the LoD analysis discussed in section 2.6 using the model-predicted concentrations in place of integrated fluorescence intensity for samples containing only a single source. The model was trained using mixtures so using the model-predicted concentration takes into account the difficulty of predicting mixtures, which is expected to increase the LoD. The model is also capable of ignoring noise present in the EEM spectra so this approach may show a lower LoD for this reason. In fact, both effects are observed when comparing the LoD of the CNN model to the LoD determined using integrated fluorescence. The LoD for cigarette remains the same, for diesel, the LoD decreases and for wood smoke, the LoD increases. As before, diesel has the weakest fluorescence intensity and thus the highest LoD of 2.2 µg/mL cyclohexane or 2.6 µg/m^3 air. The LoD for each source in a 24-hour sampling period is significantly lower than the WHO and USEPA

24-hour mean exposure guidelines of 25 and 35 μg/m^3 respectively.[12,13]

Table 3-1: LoD determined by applying the CNN model to single-source samples. The column reporting LoD in μg/mL is determined using PM mass measurement of filters dispersed in a volume of cyclohexane. The columns reporting μg/m^3 in air are determined by converting the LoD in μg/mL to μg/m^3 assuming a 24-hours sampling time at an air sampling rate of 1.8 liters per minute.

Source	CNN LoD [μg/mL cyclohexane]	CNN LoD [μg/m^3 air]	Integrated Fluorescence LoD [μg/m^3 air]
Cigarette	0.6	0.7	0.7
Diesel	2.2	2.6	3.5
Wood smoke	0.8	0.9	0.4

3.2.6 Source Identification

The ability to identify if PM from a source is present or absent above a threshold level could be a useful tool for clinicians and asthma patients in treating asthma or for asthma research, for example. To this end, I evaluated the ability of the CNN analysis of EEM spectra to detect the presence of individual sources above a threshold of 1 μg/mL. This threshold corresponds to an average exposure of nearly 10 μg/m^3, the WHO annual average guideline, during a three-hour sampling period at 1.8 L/min. In Figure 3.12, I plot the predicted concentration of each source in either a negative or a positive column. Samples are considered positive if they had a true concentration (measured gravimetrically) of 1 μg/mL or greater of any of the single sources, and negative if they are below this concentration. This analysis method is based on the establishment of a cut off value for a qualitative diagnostic health test.[81] The clinical sensitivity, specificity, and overall accuracy of the diagnostic is then determined by choosing a threshold that delineates the

positive from negative results. Depending on the purpose of the diagnostic test, the threshold may be set to achieve a specific outcome. For example, in the case of screening for a deadly but treatable disease, the number of false negatives would be minimized (i.e. maximizing sensitivity).[81] In this work, I choose the threshold that maximizes the accuracy for each source. Figure 5D shows a plot of source detection accuracy as a function of the calibrated threshold value used as a cut off between positive and negative detection. This plot shows that as I increase calibrated threshold value the detection accuracy for each source increases to a maximum and then decreases because as the threshold increases nearly all positive samples are classified as negative. The threshold of maximum accuracy varies with the source. The predicted concentration thresholds for maximum detection accuracy for cigarette, diesel, and wood smoke are 0.6, 0.8, and 0.7 μg/mL, respectively. These thresholds are shown by red horizontal lines in Figure 5A-C and when applied, an overall accuracy of 89% is achieved. The accuracies for identifying cigarette and wood smoke were 98% and 99% respectively. Diesel was more challenging because of its low signal intensity relative to the other sources and had one false positive and seven false negatives giving an accuracy of 92%.

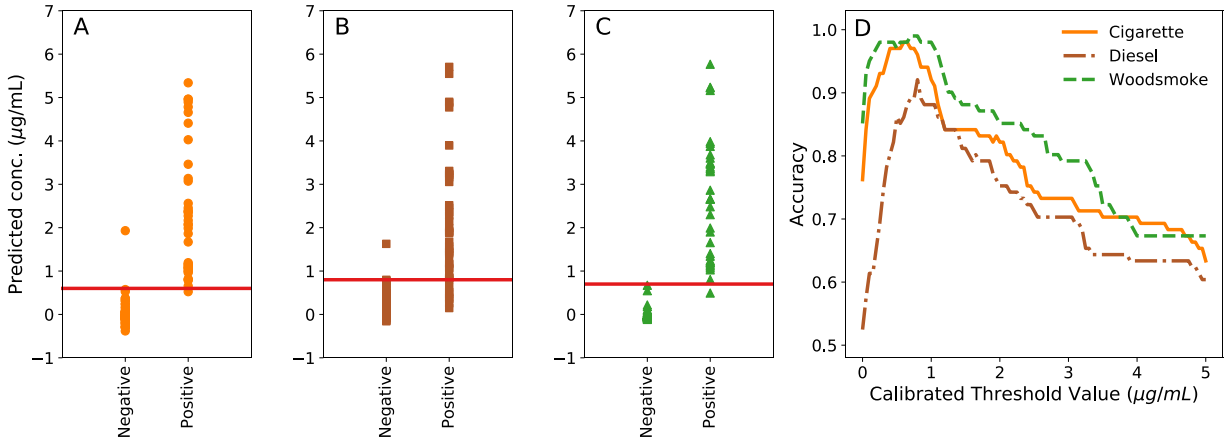


Figure 3.12: Classification plots showing the classification of test data for (a) cigarette, (b) diesel and (c) wood smoke sources as present or absent. Data points above the threshold (red horizontal line) are predicted as positive for the source. The location of the threshold was chosen to give the maximum accuracy for classifying each source individually. The source detection accuracies vs. calibrated thresholds are shown in (d).

After setting threshold values for source classification using all 101 test spectra, I evaluated the model performance on sub-groups of the test set. The sub-group of spectra containing single sources consisted of sixty-nine spectra from sixteen filter samples ranging in concentration from 0.2 µg/mL to 10 µg/mL. Within this group, samples were classified with an overall accuracy of 91%. Cigarette and wood smoke spectra were identified with the best results, while diesel was the most often misclassified with a sensitivity of 0.91 and a specificity of 0.98. Next, I tested the algorithm on the sub-group of test spectra containing two or more sources. Twenty-one samples were generated by mixing liquid extracts together, and five were from exposing an individual filter to multiple PM sources. The CNN algorithm was able to identify the sources present in mixed samples with an overall accuracy of 81%. The sensitivity and specificity for cigarette and were unity; however, diesel continued to show challenges with a specificity of unity and a sensitivity of

0.75. The relatively low sensitivity of diesel is a result of the diesel spectra being weaker than and overlapping with the other sources resulting in five false-negative results. Finally, I evaluated six process blank spectra and the algorithm correctly identified them all as not containing any of the sources. The results for the classification of sub-groups are summarized in Table 3-2.

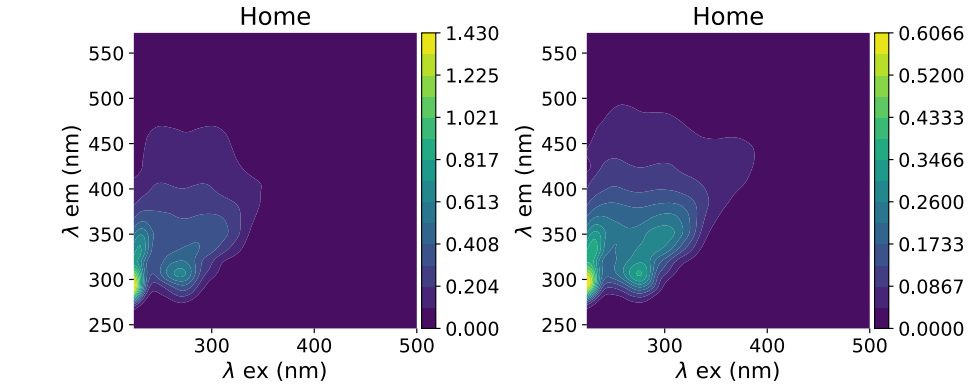
Table 3-2: Classification results for sample sub-groups containing spectra with only one PM source and mixtures (two or three sources). The overall accuracy for the single source and mixtures groups were, 91% and 81%, respectively.

	Single Source			**Mixtures**			**Process Blanks**		
	Cigarette	Diesel	Wood	Cigarette	Diesel	Wood	Cigarette	Diesel	Wood
True +	16	21	17	21	16	20	0	0	0
True -	51	45	51	5	6	6	6	6	6
False +	1	1	0	0	0	0	0	0	0
False -	1	2	1	0	5	0	0	0	0
Accuracy	0.97	0.96	0.99	1.00	0.81	1.00	1.00	1.00	1.00
Sensitivity	0.94	0.91	0.94	1.00	0.75	1.00	1.00	1.00	1.00
Specificity	0.98	0.98	1.00	1.00	1.00	1.00	1.00	1.00	1.00

3.3 Application to environmental samples

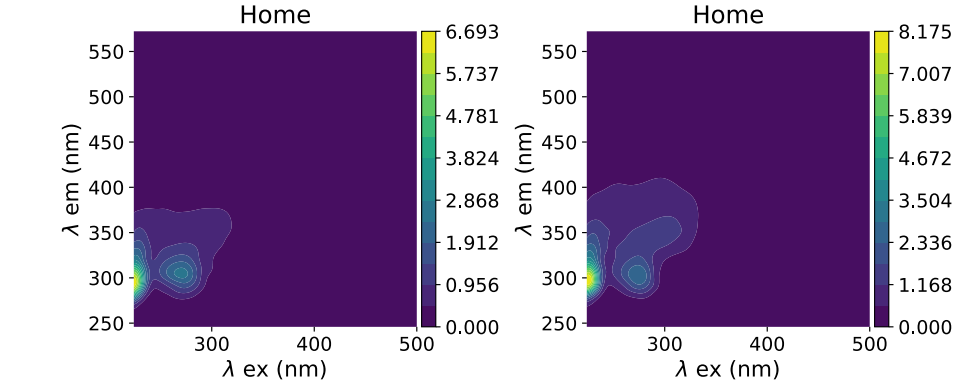
In addition to the samples collected in the laboratory, I collected twelve environmental samples to better understand how the method could be applied to real-world samples. Eight field samples were taken in Seattle homes and inside campus buildings. These samples were taken to get an understanding of typical PM present in indoor air and will be referred to as “background” samples. I also collected samples in areas expected to be dominated by cigarette, diesel, and wood

smoke. These samples will be referred to as “expected primary source” samples.



	Cig	Diesel	Wood
Pred. Conc.	-0.0	1.4	0.1
Class	0	1	0

	Cig	Diesel	Wood
Pred. Conc.	0.1	0.2	-0.1
Class	0	0	0



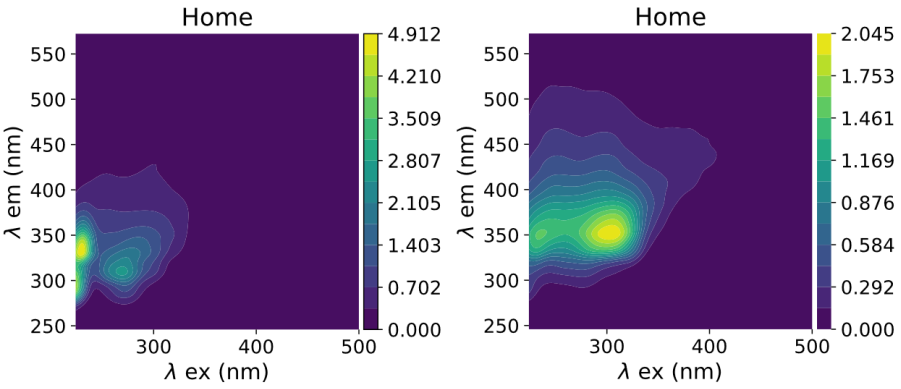
	Cig	Diesel	Wood
Pred. Conc.	-0.6	3.4	1.1
Class	0	1	1

	Cig	Diesel	Wood
Pred. Conc.	-0.2	2.1	1.3
Class	0	1	1

Figure 3.13. First four “background” environmental samples. Samples were taken in the homes of researchers and inside campus buildings. The spectra look most similar to cigarette and diesel. The table below each spectrum shows the model-predicted concentration of each source (µg PM/mL solvent) and the associated classification. A one indicates the spectrum was classified as containing the source while a zero indicates not present.

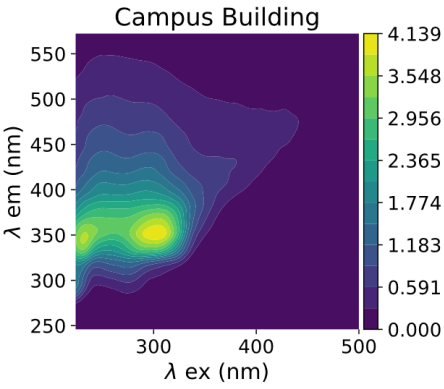
The purpose of analyzing these twelve environmental samples was to test the limits of the algorithm. The algorithm was optimized for the laboratory samples, therefore the results are

limited in scope to understanding potential difficulties when applying this method to a larger set of environmental samples. EEM spectra from the environmental samples were mathematically normalized to an extract concentration of 10 µg/mL for ease of interpreting results: the classification threshold used to train the CNN was 1 µg/mL so a spectrum will be classified as positive for a source if that source makes up 10% or more of the normalized EEM. I expect many sources of PM to contribute to the environmental samples, such as crustal dust and biological material. The EEMs from the eight *background* field samples looked most similar to diesel and cigarette spectra (Figure 3.13 and Figure 3.14). The CNN predicted cigarette as present in three *background* samples, diesel as present in four and wood smoke as present in two. I did not expect any of the *background* samples to contain cigarette smoke, as they came from non-smoking households and buildings, but this source was detected in three samples that had spectra of similar appearance to cigarette smoke. This suggests that some sources of PM have similar EEM spectra. Diesel may have been present as all samples were collected in urban areas of Seattle so the prediction of diesel in four of the background spectra was not a surprising result. Wood smoke was detected in two *background* samples. The spectra where wood smoke was detected looked most similar to diesel but had higher fluorescent intensity than diesel at the same concentration (10 µg/mL). This illustrates that the CNN may give unexpected results when analyzing spectra that are different from the spectra used in training. An algorithm trained with a limited number of underlying sources should be limited in application to samples where only those expected underlying sources are present.

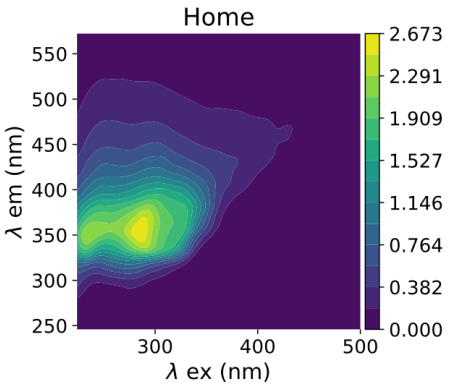


	Cig	Diesel	Wood
Pred. Conc.	-0.3	9.2	0.5
Class	0	1	0

	Cig	Diesel	Wood
Pred. Conc.	2.8	-0.3	0.1
Class	1	0	0



	Cig	Diesel	Wood
Pred. Conc.	5.1	0.0	0.5
Class	1	0	0



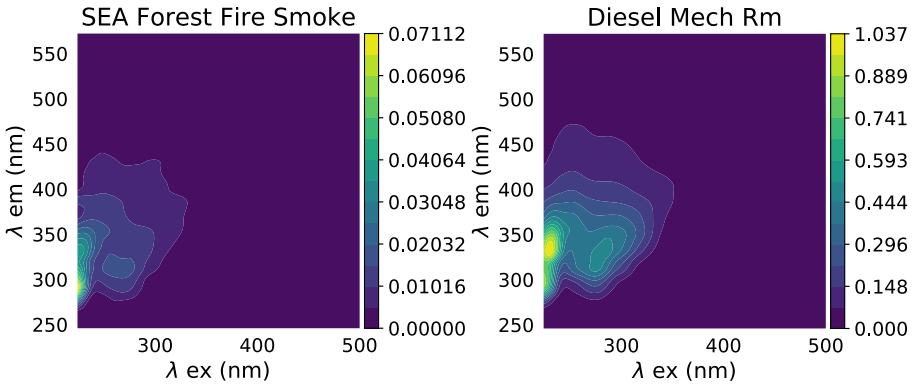
	Cig	Diesel	Wood
Pred. Conc.	3.7	-0.2	0.3
Class	1	0	0

Figure 3.14. Additional "background" environmental samples. (See Figure 3.13 caption for details)

Wood smoke was the expected primary source in an ambient sample taken in the UW clean cookstoves lab and in a sample taken from an open window during a time when forest fire smoke was causing air pollution in Seattle. Wood smoke was detected in the cookstove lab sample as expected and the EEM resembled other wood smoke spectra (Figure 3.15). I believe this was due to small amounts of wood smoke escaping the ventilation system during stove testing. Wood smoke was not detected in the sample taken during the forest fire smoke episode. I believe this is

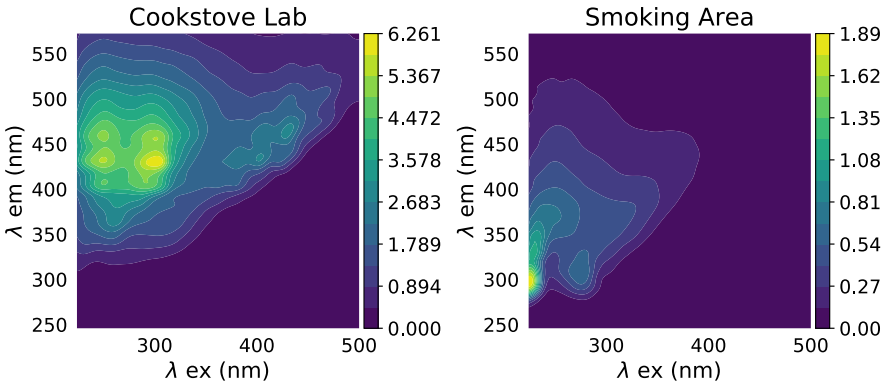
due to the forest fire smoke having a different composition than the laboratory-generated wood smoke due to a combination of aging during atmospheric transport and different combustion conditions in a forest fire compared to a cookstove. Diesel exhaust was expected and detected as present in a sample taken in a mechanical room at the diesel exhaust exposure facility. I believe this was due to fugitive emissions of diesel exhaust as with the cookstove sample. Cigarette smoke was expected in a sample taken outdoors near a smoking area, but only diesel was detected. In retrospect, I believe this was due to minimal amount of cigarette smoke present in the sample as the PM concentration measured by the filter was 6.3 $\mu g/m^3$, while the average concentration measured over the same time period at two nearby air monitoring sites in Seattle was 5.5 $\mu g/m^3$ showing that this outdoor sample likely consisted of a typical mixture of urban PM that would be expected to include diesel exhaust and was not heavily influenced by cigarette smoke.[88]

Analysis of the *expected primary source* samples suggests that the method of collecting samples from known sources and applying an EEM-CNN approach to source apportionment may be useful for occupational monitoring applications. For example, if a workplace had multiple sources of combustion PM that could be sampled individually, a model using these known sources could be trained and used to identify how much of each of the combustion sources a worker was exposed to using personal monitoring equipment.



	Cig	Diesel	Wood
Pred. Conc.	-0.0	0.2	-0.1
Class	0	0	0

	Cig	Diesel	Wood
Pred. Conc.	0.1	2.1	-0.1
Class	0	1	0



	Cig	Diesel	Wood
Pred. Conc.	-0.1	1.1	5.9
Class	0	1	1

	Cig	Diesel	Wood
Pred. Conc.	0.2	2.2	0.2
Class	0	1	0

Figure 3.15. *Expected primary source* environmental samples were taken in locations where there was an a priori expectation that a primary source would contribute to the spectra. The table below each spectrum shows the model-predicted concentration of each source (µg PM/mL solvent) and the associated classification. A one indicates the spectrum was classified as containing the source while a zero indicates not present.

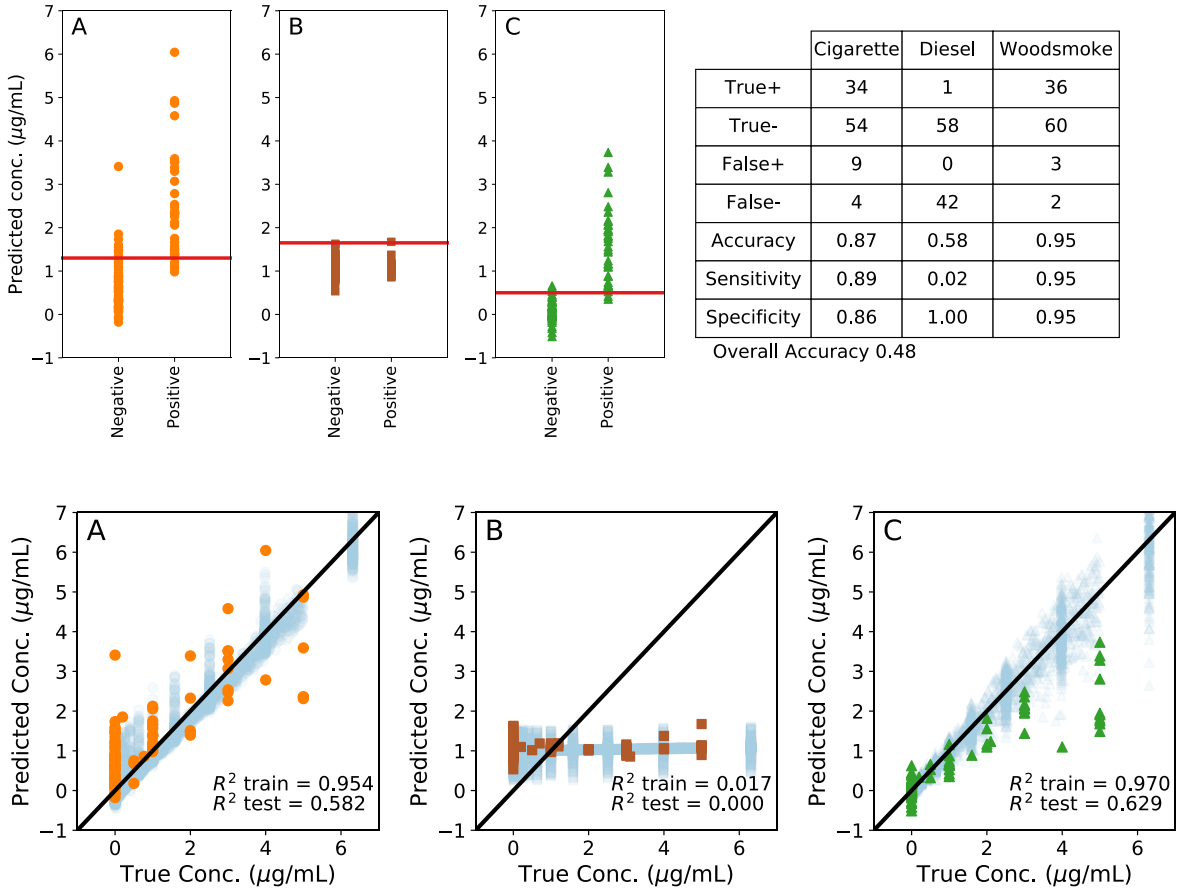
3.4 Other Modeling Results

I evaluated a partial least squares (PLS), a linear model, and a principal component regression (PCR) model to interpret the EEM spectra in the same manner as the CNN. The PCR model performed the best of these three models with an overall classification accuracy of 70% The linear model achieved an overall classification accuracy of 68% and the PLS model had an overall accuracy of only 48%. These alternative models performed poorly largely due to an inability to accurately predict diesel concentration.

3.4.1 Partial Least Squares

PLS, also known as "Projection to Latent Structure" is a modeling technique that projects the original predictors (fluorescent intensities) and responses (concentrations) into a lower-dimensional space. The new representation of the predictors and responses are used to fit a linear model. The conceptual framework of PLS is described by G. James et al. in *An Introduction to Statistical Learning: with Applications in R* and Wegelin provides a detailed discussion of the PLS algorithm.[89,90] I implemented the PLS algorithm using Scikit-learn in Python 3.[91]

The number of dimensions in the lower dimensional space (referred to as components) is a tuning parameter for a PLS model. I fit models with 1 to 6 components. Of these models, the 2-component model had the best results which are shown in Figure 3.16. The PLS model performance for cigarette and wood smoke was acceptable with an accuracy of 87% and 95% for these sources, respectively, but it was unable to accurately identify diesel resulting in 58% accuracy for diesel and overall accuracy of 48%. When additional components are included the model is able to fit the diesel training data, but the R^2 value for the diesel test data remains at zero.



	Cigarette	Diesel	Woodsmoke
True+	34	1	36
True-	54	58	60
False+	9	0	3
False-	4	42	2
Accuracy	0.87	0.58	0.95
Sensitivity	0.89	0.02	0.95
Specificity	0.86	1.00	0.95

Overall Accuracy 0.48

Figure 3.16 PLS classification and regression results for a two-component model. The PLS model performs acceptably well on cigarette and wood smoke but fails to classify diesel with any accuracy.

3.4.2 Simple Linear Model

I also evaluated a linear model. First, I determined the average EEM spectra ($\boldsymbol{S}_{avg}$) for each source using the 4 calibration spectra from each source as:

$$\boldsymbol{S}_{avg} = \sum_{i=1}^{4} \boldsymbol{S}_i / c_i,$$

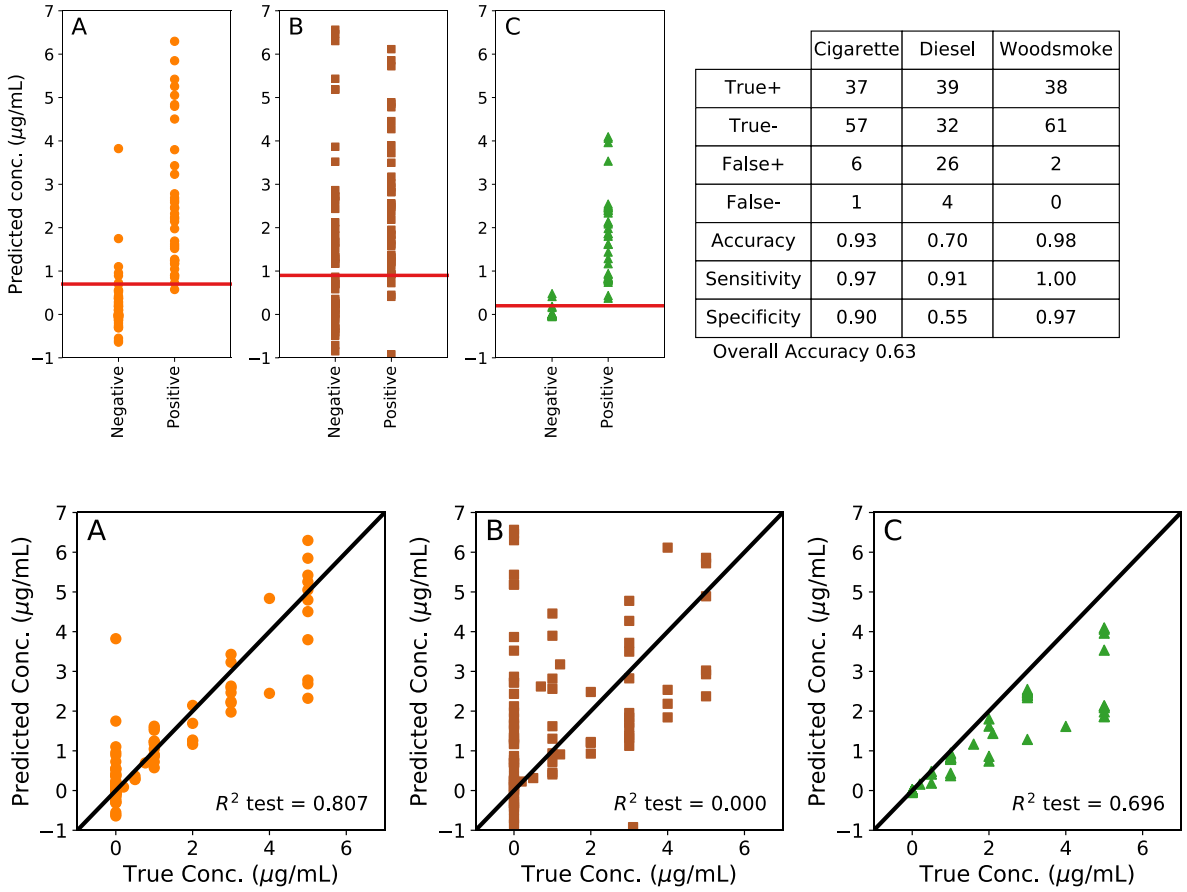
where $\boldsymbol{S}_i$ is a single source calibration-spectra and c_i is the concentration of the calibrant. I

assumed the 3 average spectra profiles could be used to reconstruct a test spectrum according to,

$$\boldsymbol{S} = c_C\boldsymbol{S}_C + c_D\boldsymbol{S}_D + c_W\boldsymbol{S}_W + \boldsymbol{E},$$

Where $\boldsymbol{S}$ is a test-spectra, c_C, c_D, and c_W are the predicted concentrations of cigarette, diesel and wood smoke, respectively, $\boldsymbol{S}_C$, $\boldsymbol{S}_D$, and $\boldsymbol{S}_W$ are the average EEM spectra for cigarette, diesel and wood smoke, respectively and $\boldsymbol{E}$ is the error. To predict the three concentration values, I used the Needler-Mead optimization algorithm in SciPy to minimize the root-mean-square of $\boldsymbol{E}$.

The results of the linear model are summarized in Figure 3.17. The performance of this model was superior to the PLS model, with an overall accuracy of 63%, but was worse than the CNN model that had an overall accuracy of 89%. The performance of the linear model on cigarette and wood smoke was close to the performance of the CNN with accuracies of 93% and 98% compared to 98% and 99%. The linear model performed poorly on diesel with an accuracy of only 70% compared to an accuracy of 92% for the CNN model.



	Cigarette	Diesel	Woodsmoke
True+	37	39	38
True-	57	32	61
False+	6	26	2
False-	1	4	0
Accuracy	0.93	0.70	0.98
Sensitivity	0.97	0.91	1.00
Specificity	0.90	0.55	0.97

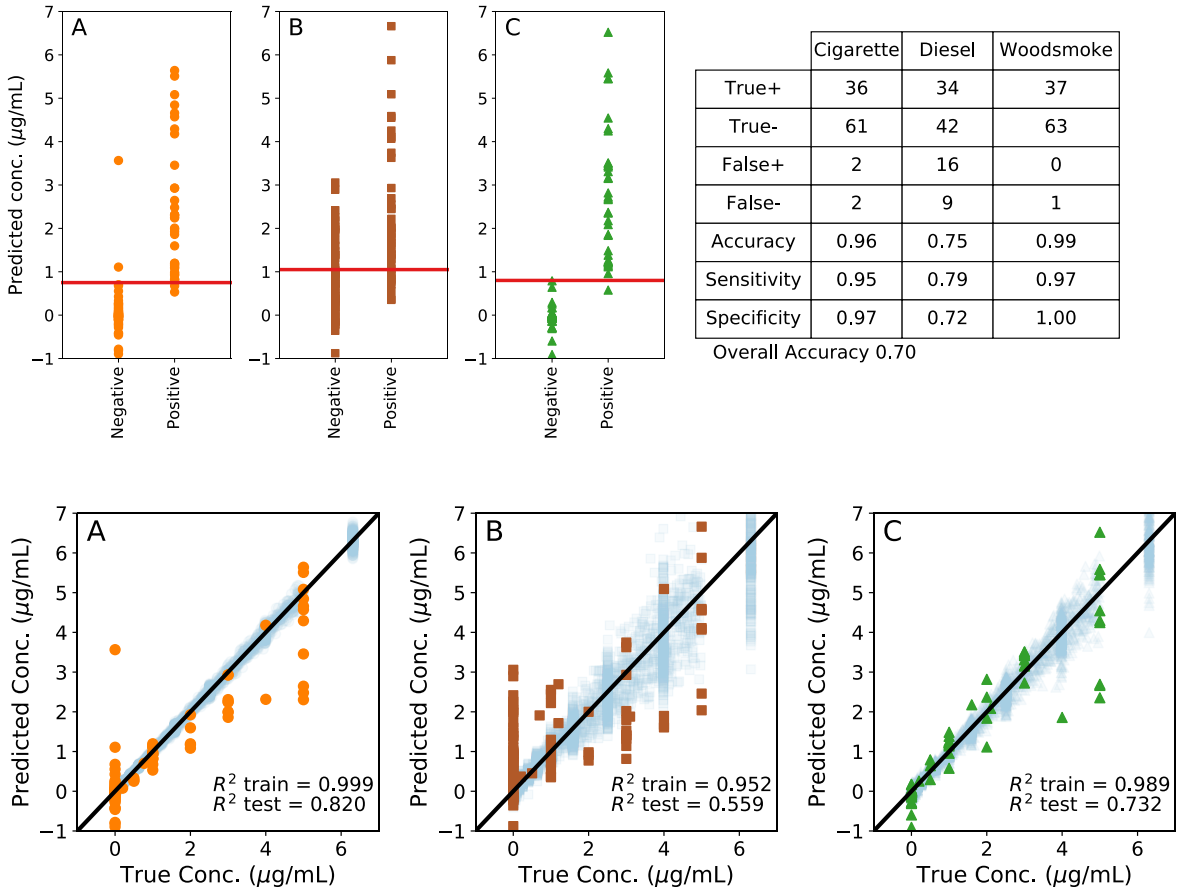
Overall Accuracy 0.63

Figure 3.17. Linear model classification and regression results. The linear model performed well on cigarette (a) and wood smoke (c) but had poor performance on diesel (b). The poor performance on diesel led to a low overall classification accuracy of 63%. The linear model was not fit using the augmented training dataset like all other models so only the test data are shown in the parity plots.

3.4.3 Principal Component Regression

PCR models were evaluated using between 2 and 10 principal-components. Of these models, the 6-principal-component model had the best results which are shown in Figure 3.18. The performance of the PCR model was superior to the PLS and linear models, with an overall accuracy of 70%, but was worse than the CNN model that had an overall accuracy of 89%. The

performance of the PCR model on cigarette and wood smoke was nearly identical to the performance of the CNN with accuracies of 96% and 99% compared to 98% and 99% for the CNN. The PCR model performed marginally on diesel with an accuracy of 70% compared to an accuracy of 92% for the CNN model.



	Cigarette	Diesel	Woodsmoke
True+	36	34	37
True-	61	42	63
False+	2	16	0
False-	2	9	1
Accuracy	0.96	0.75	0.99
Sensitivity	0.95	0.79	0.97
Specificity	0.97	0.72	1.00

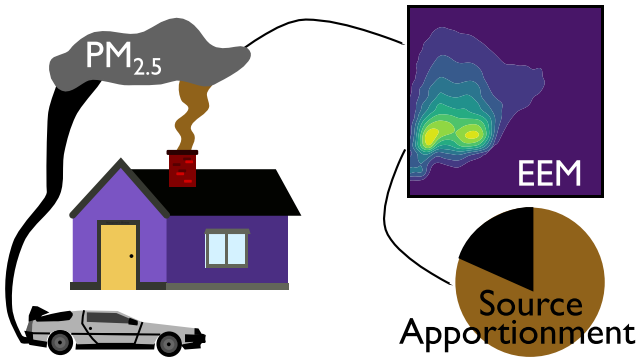
Overall Accuracy 0.70

Figure 3.18 PCR classification and regression results using six principal components. The PCR model performs acceptably well on cigarette (a) and wood smoke(c) and marginally for diesel.

3.5 Summary

I used a CNN model to successfully classify cigarette, diesel, and wood smoke sources as present or absent in a series of laboratory samples. The limit of detection for the method is 0.7, 2.6, and 0.9 μg/m^3 in air for cigarette, diesel, and wood smoke respectively. The CNN was able to identify cigarette and wood smoke individually with 98% and 99% accuracy respectively, while the classification of diesel was less accurate with an accuracy of 92%. The overall classification accuracy for classifying all three sources correctly simultaneously was 89%. When testing the limits of our algorithm by classifying environmental samples, some samples were classified as expected while in others, sources were detected as present even when they were not expected. This illustrates the need for a training data set that includes the expected underlying sources for the successful application of the EEM-CNN method described in this chapter. Additionally, I expect atmospheric aging of PM may change EEM spectra in intensity and or overall appearance. A controlled study of these effects using an atmospheric chamber would be valuable to further evaluate this method.

Chapter 4: Source Apportionment of Environmental Combustion Sources



4.1 Motivation and background

In chapters 2 and 3 I have shown the utility of EEM spectroscopy coupled with machine learning for identifying and apportioning sources of combustion generated particulate matter from controlled sources in the laboratory. I applied the EEM-machine learning approach to a small set of environmental samples collected in homes of researchers and at buildings at the UW showing that without a priori knowledge of the underlying sources the approach is of limited utility. In order for this method to be useful in source apportionment studies, it must be capable of reproducing results obtained by orthogonal source apportionment methods. In this chapter, I apply the EEM-machine learning approach to a set of filed samples and associated source apportionment results obtained by an orthogonal method.

4.1.1 Environmental samples from the Seattle Panel Study

I obtained a set of archived filter samples from a panel study examining the health effects of $PM_{2.5}$ exposure on susceptible populations (e.g. asthmatics, coronary heart disease) in the Seattle WA metropolitan area referred to as the Seattle Panel Study (SPS) henceforth.[36,92,93] In brief, samples were collected for 24 hours using 37-mm polytetrafluoroethylene filters (model 225-1709; SKC, Inc, Eighty Four, PA) and a 10-L/min impactor for $PM_{2.5}$. PM mass concentration was measured gravimetrically and samples were analyzed for trace elements using XRF. Light absorbing carbon (LAC) was determined by measurement of the particle light-absorption coefficient at 525 nm using the integrating plate method followed by regression of this measurement against elemental carbon measured by thermal optical transmittance on collocated quartz filters. Further details of sampling and elemental analysis are described in the original publications.[36,92,93] All analysis performed on the filter samples was non-destructive and the filters were archived in a freezer.

4.1.2 PMF Source Apportionment of SPS Samples

Source apportionment was performed using PMF and results were reported for indoor, outdoor and personal samples.[36] In the present work, I consider the results from the outdoor samples collected either outside subjects' homes (n=198) or at a central monitoring site (n=96). In the outdoor samples, six sources were identified using PMF: vegetative burning, crustal dust, Cl-rich (marine), mobile (traffic), secondary sulfate and fuel oil. The sources were identified by comparing post hoc the elemental profiles generated by PMF to previously published source profiles. For example, the vegetative burning source shows characteristically high carbon content along with potassium. Vegetative burning was responsible for the highest contribution (62%) to total $PM_{2.5}$ mass of all the sources identified. This was expected due to the number of samples collected in the heating season and the prevalent use of woodstoves as a heating source. As an additional verification of the vegetative burning source contribution, the authors compared a subset of the samples to levoglucosan, a wood smoke tracer compound, as determined by GC-MS[94] and found a strong correlation ($R^2 = 0.87$) between the vegetative burning source determined by PMF and the levoglucosan measurements. A detailed discussion of the source profiles and identification is provided by Larson et. al.[36]

4.2 EEM analysis of Environmental Samples

I selected 45 filter samples from the central monitoring site, 50 from outside subjects' homes, and 8 method blank filters for analysis by EEM. I used a specialized cutting tool[94] to remove the polyolefin ring from the filters prior to placing them in 20 mL glass vials, immersing them in 3 mL of cyclohexane, sonicating for 30 mins and allowing them to soak for a total of 24 hours before transferring them to 4 mL glass vials (Cat # 66009-876 VWR, Edison, NJ) for storage until analysis. I collected EEM data from these samples as described in section 2.4 with the

following variations: Data were recorded for all wavelengths between 200 to 500 nm, and all emission values below the excitation wavelength were zeroed after scatter removal was applied.

4.3 EEM Fluorescence Spectra

Fluorescence was observed from all PM extracts. The minimum sample extract concentration was 19.5 µg $PM_{2.5}$ per mL cyclohexane. In Figure 4.1 I plot integrated fluorescent intensity (IFI) vs. $PM_{2.5}$ extract concentration. The data show a roughly linear relationship between PM mass and fluorescence with an average fluorescent intensity of 390 IFI units per µg of PM (i.e. slope of the best fit line). The limit of blank (Equation 2-1) for these samples is 220 IFI units and the lowest sample is at 1100 IFI units illustrating that the PM samples show fluorescence greater than the blanks.[95] In chapter 3, I found diesel soot, cigarette, and wood smoke collected in the laboratory had fluorescent intensities of 200, 2400, and 4600 IFI units per µg of PM in the extract (Figure 3.4). I believe the much higher values of fluorescence per mass of PM for the cigarette and wood smoke laboratory samples are due to these sources being collected fresh with little dilution and the presence of more fluorescent compounds on a per mass basis. The diesel soot samples were collected in a facility designed to simulate environmental exposure to diesel exhaust[96] and show IFI similar to the environmental samples on a per mass basis. The data plotted with blue squares in Figure 4.1 are samples taken outside of subjects' homes and show more scatter than the data from the central monitoring site shown by green circles. The larger scatter in the samples from subjects' homes is expected due to varying sample locations and the associated variation in sources due to local contributions such as wood-burning stoves and nearby roadways.

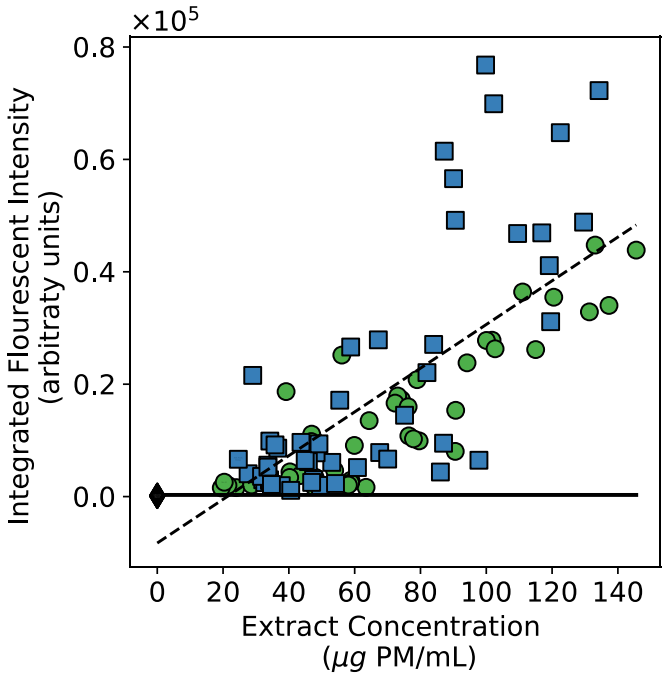


Figure 4.1: Total fluorescence intensity vs. $PM_{2.5}$ extract concentration. Central monitoring site samples are shown as green circles, home samples are shown as blue squares, and process blanks are shown as black diamonds. The solid horizontal line is at the *limit of blank* and the line of best fit for all samples is shown as a dashed line ($R^2 = 0.77$).

Figure 4.2 shows three exemplary EEMs and the associated PMF source contributions having the highest percentage of mobile, vegetative, and fuel oil. These spectra show peaks at approximately (~225 nm excitation, ~350 nm emission) and (~300, ~350) that vary in relative intensity. The spectra in Figure 4.2A and B look similar to the typical cigarette spectra (Figure 2.3). Cigarette smoke is not a significant contributor to these samples, but spectra of this general appearance could be typical of smoldering (low temperature) combustion from sources like cigarettes and poorly tended woodstoves. The spectra plotted in Figure 4.2C has an additional peak at (~220, ~275), this spectrum appears very similar to the typical diesel soot spectra (Figure 2.3). The qualitative differences in these spectra and the associated differences in the source

contribution estimates are an illustration of how the EEM spectra may be useful in determining source contributions.

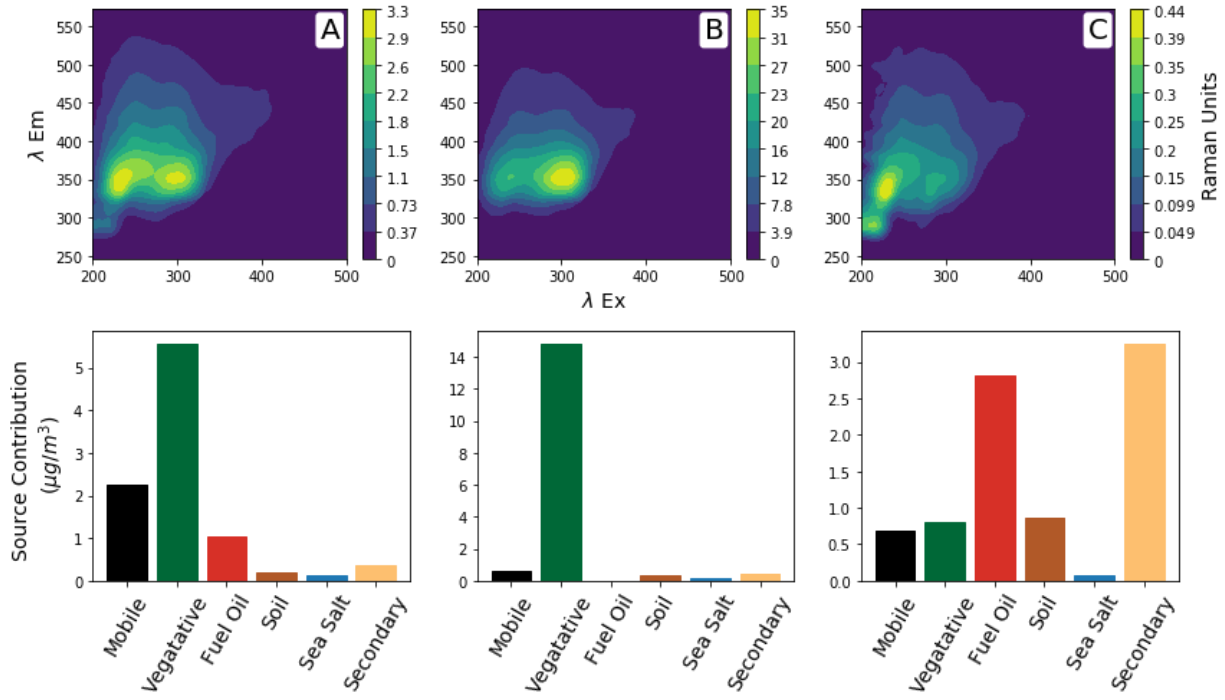


Figure 4.2: Example EEMs and associated PMF source contributions. The variation in the EEM spectra and the associated variation in source contributions suggest EEM may be a useful tool for source apportionment. The spectrum plotted in Panel A shows the sample with the highest percentage mobile contribution, Panel B shows the highest percentage vegetative burning, and Panel C shows the highest percentage fuel oil.

4.4 Machine Learning for Source Apportionment

I matched the 103 EEM spectra to the original source apportionment study data to get source assignments for each sample. For the 8 method blank samples, I assigned source contributions of zero for all sources. These source-EEM pairs were used to test the ability of machine learning to identify combustion sources from EEM data. I split the data into training and

testing sets to fit machine learning models. The training data is used to adjust model parameters to achieve the best fit by minimizing the error between expected and predicted values. The testing set is not used for model fitting and after model fitting is complete, values are predicted for the test data. The quality of fit on the test data provides an assessment of model validity not biased by overfitting. The data chosen for the testing and training sets may influence the results. To provide an assessment that is less influenced by the samples selected cross-validation is employed. I used 5-fold cross-validation for fitting the PCR and CNN models presented in this work. The 103 EEM spectra and associated source contributions were randomly split into 5 groups (three groups of 21 and two of 20). I used the same five groups of data for both model types to ensure the random splitting of the data did not influence the comparison of the models. The results of each fold are aggregated when calculating r-squared values for each type of machine learning applied.

4.5 PCR Analysis

Principal component regression was performed in Python 3 using Scikit-learn, version 0.19.2.[97] EEM data were unfolded into rows for application of PCA using the `decomposition.PCA` module and the resulting PCs were used to perform ordinary least squares regression using the `linear_model.LinearRegression` module.

I performed PCA on the entire dataset (103 spectra) for the purpose of visualization. The resulting PCs and the mean of all EEM spectra are plotted in Figure 4.3. The PCs have both negative and positive values because they represent how to adjust the mean to reconstruct individual fluorescent spectra. The first five PCs shown account for 99.96% of the variance in the dataset showing that although the original data contain over twenty-thousand fluorescent values the spectra can be effectively represented in a lower dimensional space. The mean and PC1 appear

roughly similar to the EEMs plotted in Figure 4.3A and B, which represent the most common appearance of spectra in the dataset showing peaks at (~225 nm, ~350 nm) and (~300, ~350). Principal components 2-5 show finer details of the spectra that are generally not visually apparent when plotting spectra due to their small magnitude compared to the dominant features shown in PC1 (B). PC2 is qualitatively similar to the wood smoke spectra from the laboratory as shown in (Figure 2.3).

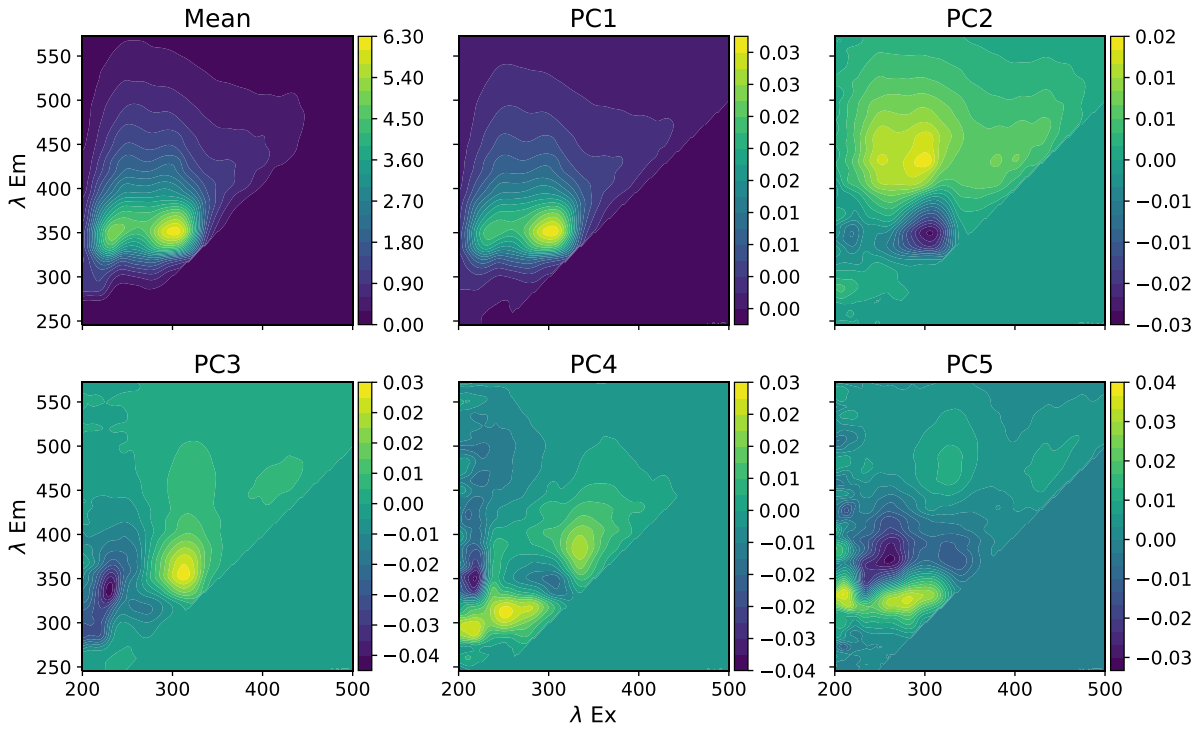


Figure 4.3: The mean of all EEM spectra and the first five principal components. The mean and PC1 represent the most common appearance of spectra in the dataset. The remaining PCs show finer details of the spectra. The five PC shown here account for 99.96% of the variance in the dataset.

I performed PCR with 5-fold cross-validation using from one to fifty principal components. Figure 4.4 plots the R^2 values for mobile and vegetative burning sources vs. the number of PCs used for PCR. The R^2 value for the training data continually increases as the number of PCs is

increased, but the R^2 value for the test data begins to decrease as more principal components are added as a result of overfitting as shown in Figure 4.4. I selected seven PCs for the model as this gave the highest sum of R^2 values for the mobile and vegetative burning sources.

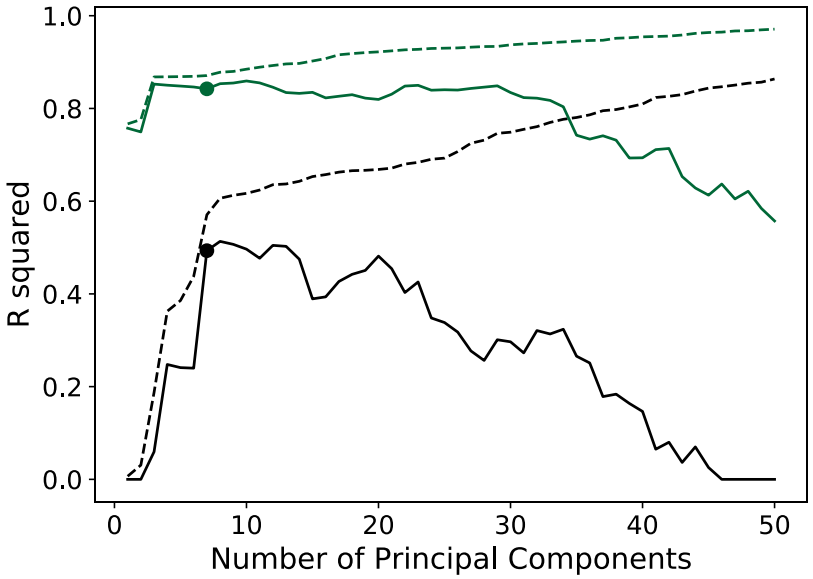


Figure 4.4: PCR R^2 vs. number of PCs used in PCR for mobile (black) and vegetative burning (green). As the number of PCs is increased from one to fifty the fit to the training data continually improves (dashed lines), however, the fit to the test data decreases after reaching a maximum (solid lines). Seven PCs (marked with circles) are used for PCR because this gave the highest combined R^2 value for the mobile and vegetative burning sources.

The prediction results for vegetative burning and mobile are shown along with results for all other sources by the parity plots in Figure 4.5. Training data are shown as light blue points behind the solid color test data. Perfect prediction is when all points fall on the one-to-one line (diagonal line) in the parity plots. Points above this line are over-predictions and below are under-predictions. PCR was able to identify the mobile and vegetative burning sources successfully. The R^2 values on the training and test data for mobile are 0.57 and 0.49, respectively and for vegetative burning the train and test R^2 values are 0.87 and 0.84. For

vegetative burning and mobile emissions, the data roughly follow the one-to-one line showing some over- and under-predictions across all concentrations. The PCR model failed to predict the other sources resulting in R^2 values less than 0.10 on the test data for the remaining sources. The poor prediction for the soil, sea salt, secondary and fuel oil sources is shown in Figure 4.5 by the fact that the data do not follow the one-to-one line. Instead, for low concentrations, the model generally over-predicts and for higher concentrations the model generally under-predicts resulting in the data following a generally horizontal trend near the mean of the PMF concentration.

I performed PCR on all samples to evaluate the relative importance of the various PCs. All principal component scores were normalized to have a mean of zero and unit variance before applying linear regression. PCs are orthogonal to one another and therefore this normalization allows the magnitude of the regression coefficients to be interpreted as their relative importance.[98] For vegetative burning I found PC1 and PC3 to account for 89% of the magnitude of coefficients. For mobile, I found PC3, PC4 and PC7 73% of the magnitude of coefficients. I expected PC2 to be a significant contributor for vegetative burning because of its qualitative similarity to the laboratory wood smoke spectra, however, PC2 was only responsible for 6% of the prediction for vegetative burning.

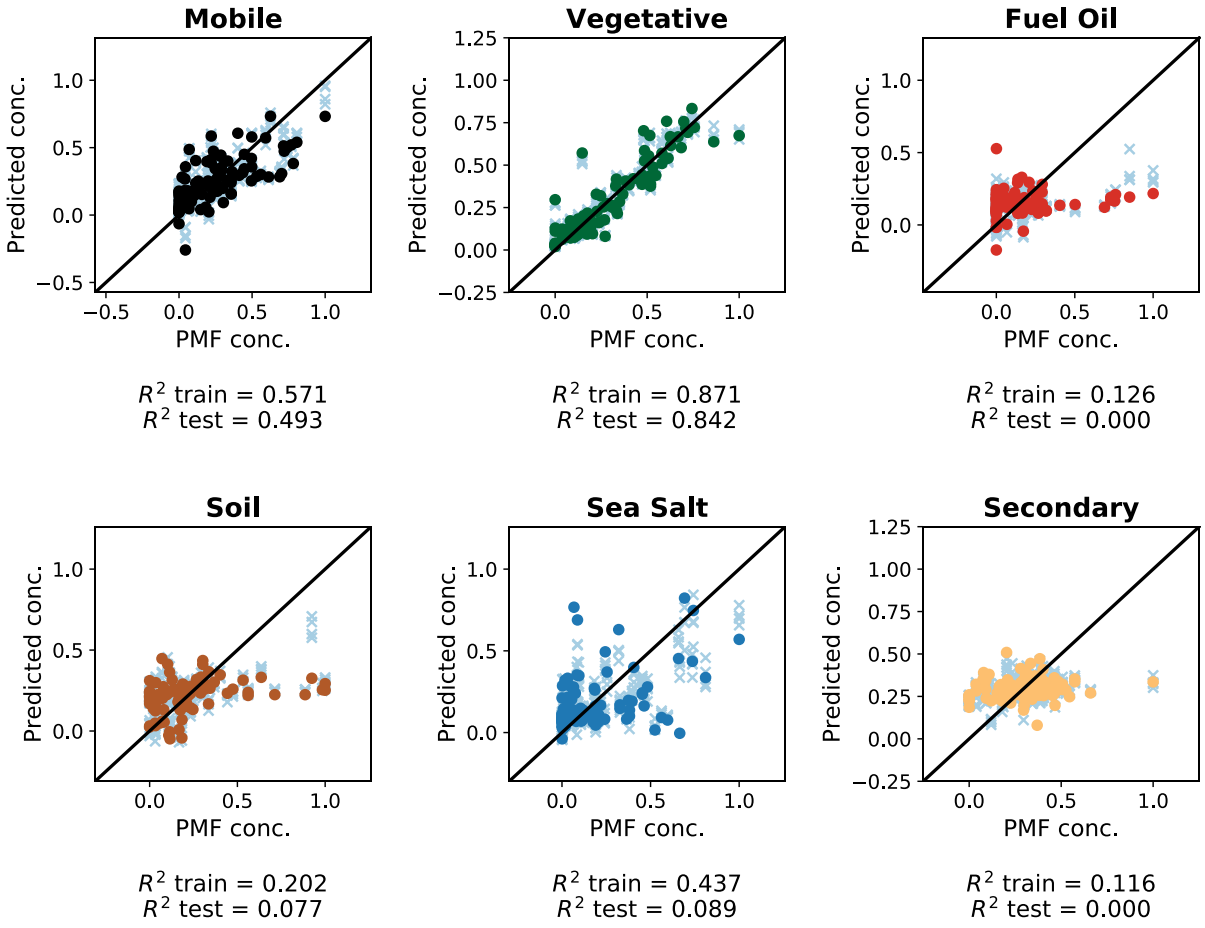


Figure 4.5: Parity Plots showing results of PCR with 7 components using 5-fold cross-validation. Training data are shown by light blue x's and test data are shown with solid color points. The PCR model was able to predict the vegetative burning source with the highest accuracy with training and test R^2 values of 0.87 and 0.84, respectively. Mobile was also successfully predicted with (train, test) R^2 values of (0.57, 0.49). All other sources showed test R^2 values of less than 0.10: fuel oil (0.13, 0.0), soil-(0.20, 0.08), sea salt – (0.44, 0.09), and secondary (0.12, 0.0).

4.6 CNN Analysis

In order to effectively train the CNN on the relatively small dataset of 103 spectra, I implemented a data augmentation approach to increase the number of training spectra. As

discussed in section 3.1, I assume the Beer-Lambert law applies to absorbance, fluorescence quantum yield is constant, and inner filter and matrix effects are negligible, which allows me to generate augmented EEM spectra by linear combination. Taking all possible combinations, with replacement, of the training spectra results in 3240 pairs for a training set of 80 spectra, for example. For each pair, I add the fluorescent intensity values together and divide by two. When a spectrum is paired with itself, the result is the original spectrum. The same approach is applied to the six source contributions paired with each spectrum. Figure 4.6 shows a graphical example of this process.

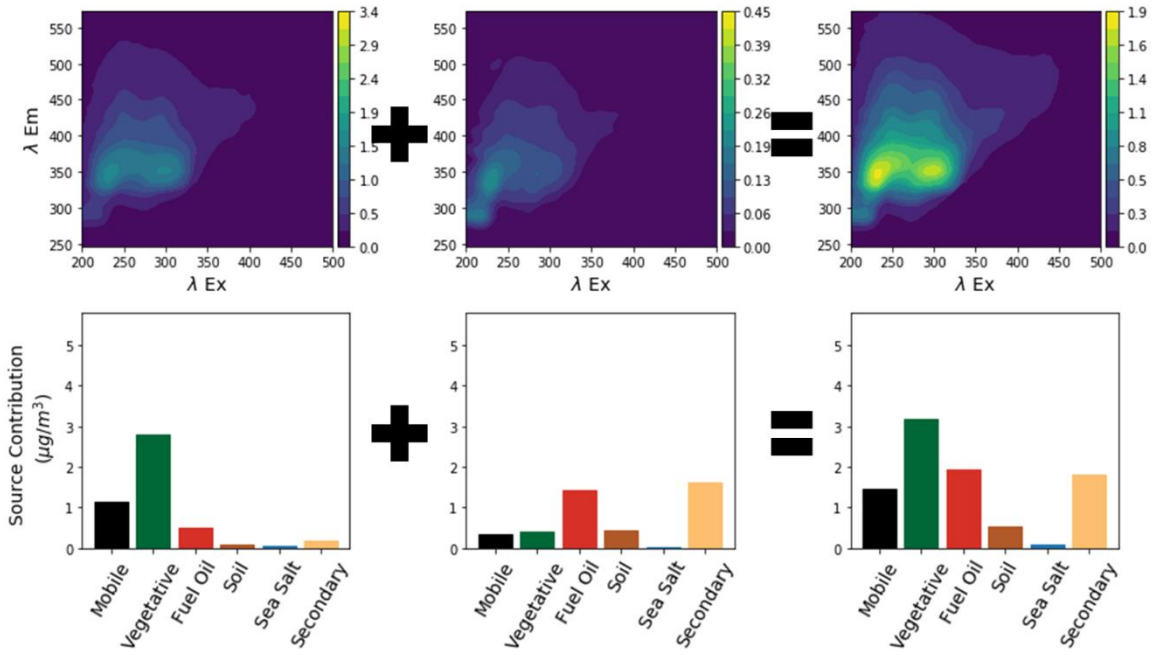


Figure 4.6: Graphical representation of data augmentation by 50-50 combination of spectra and PMF source contributions. The spectra and PMF source contributions shown on the left and center are the same as those shown in panels A and C of Figure 4.2 after scaling by 50%. These spectra are added together to generate a new spectrum for the training process.

The architecture of the CNN used is based on the network developed for the laboratory samples consisting of three convolutional layers each followed by max pooling and a dense neural network with three hidden layers. Data dimensions associated with each layer are shown in Figure 4.7. The output of the convolutional layers is connected to a dense neural network with three hidden layers. I used a dropout rate of 20% between all convolutional and fully connected layers.[99] The CNN was implemented in Python 3 using Keras[85] and TensorFlow[86] using default parameters, with the exception of the "learning rate" used with the *Adam* optimizer[100] which I reduced from the default value of 0.001 to 0.0001. I found this to reduce the variability in the fit between epochs. I trained the CNN for 300 epochs.

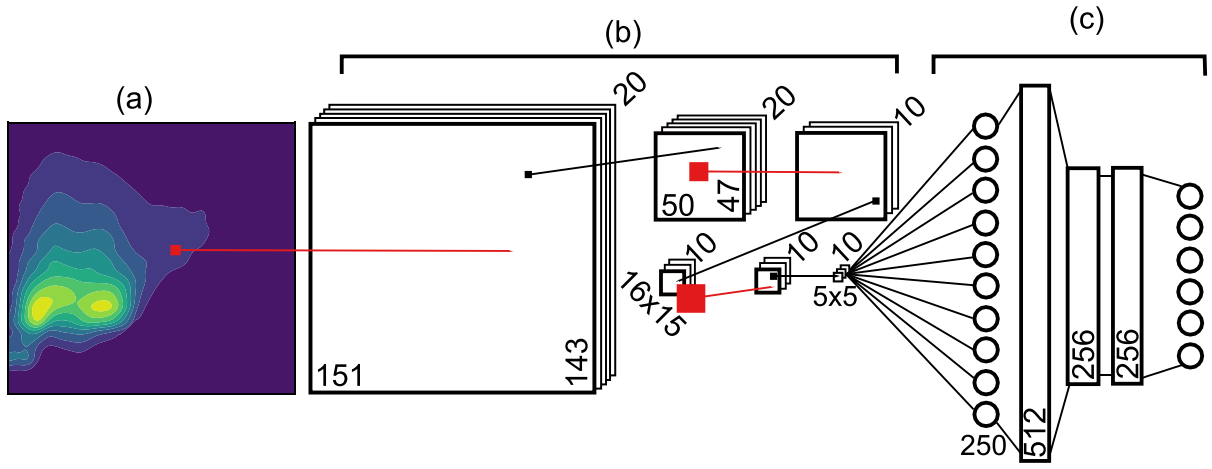


Figure 4.7: Neural Network Architecture. (a) Input spectra are first convolved with twenty 5-by-5 filters. (b) convolutional layers and max pooling layers are shown with associated data shapes. Convolutions (10x10 followed by 15x15) are shown in red and 3-by-3 max pooling is shown in black. Filters, data, and pooling sizes are shown to scale. (c) Output of the convolutional layers is flattened to a shape of 250 by 1 and fed into fully connected layers resulting in six output values (not to scale).

I applied a CNN to the augmented data using 5-fold cross-validation and report aggregate results. Figure 4.8 shows predictions for mobile and vegetative burning in parity plots. The R^2 values on the training and test data for mobile were 0.75 and 0.52, respectively and for vegetative burning the training and test R^2 values were 0.91 and 0.81. Figure 4.8 shows training data in light blue behind the solid test data points. The training data plotted here are more numerous than the training data in the PCR parity plots (Figure 4.5) due to the use of data augmentation. The training and test data roughly follow the diagonal one-to-one line. The R^2 values for the test data on all other sources were below 0.20.

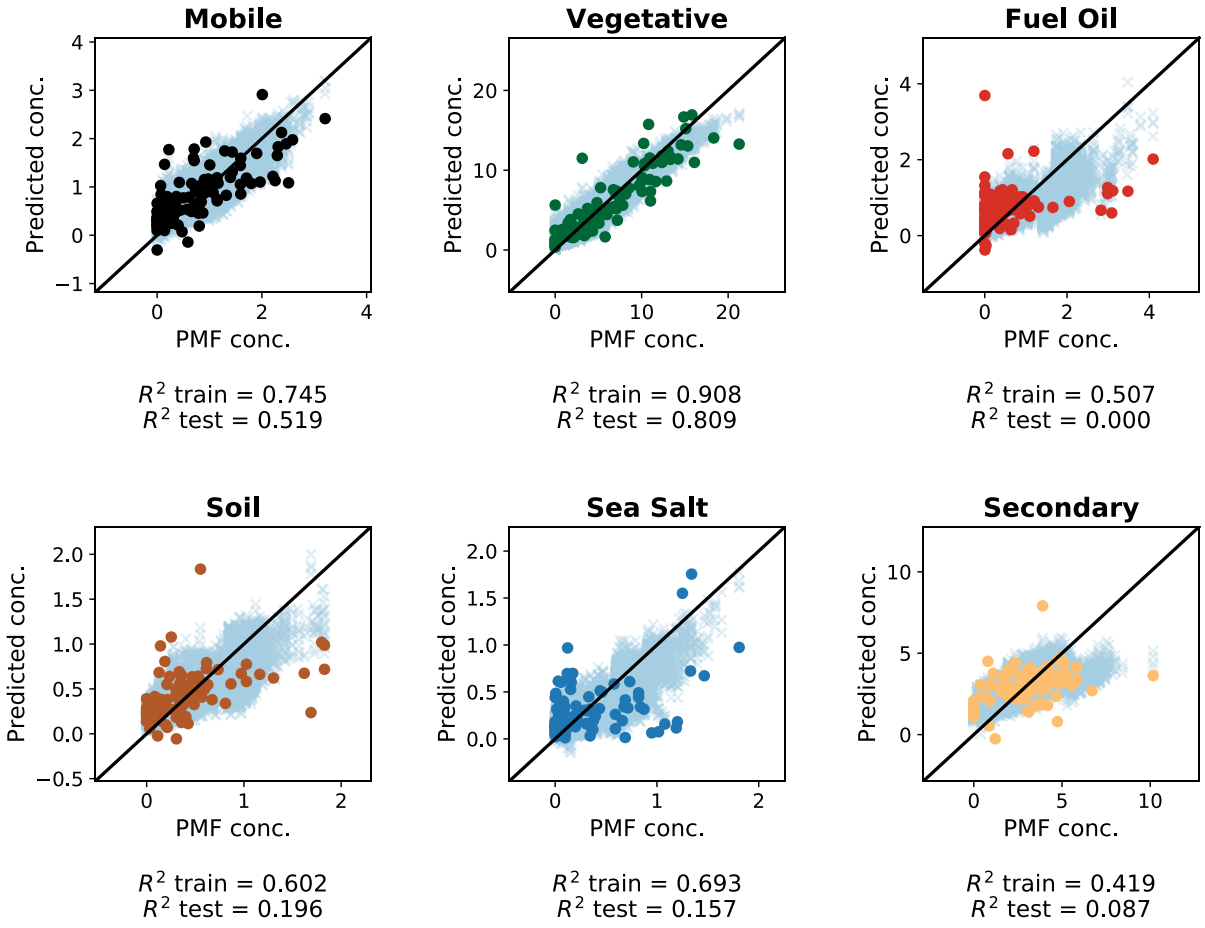


Figure 4.8: Parity Plots showing CNN results using 5-fold cross-validation. Training data are shown by light blue x's and test data are shown with solid color points. The CNN was able to predict the vegetative burning source with the highest accuracy with training and test R^2 values of 0.91 and 0.81, respectively. Mobile was also successfully predicted with train and test R^2 values of 0.75 and 0.52. All other sources showed test R^2 values of less than 0.2.

4.7 Identification of Combustion Sources

The EEM-machine learning method should be capable of detecting all the combustion sources due to the presence of fluorescent compounds such as PAHs in combustion derived PM. For the environmental samples analyzed in this work, it was only successful at predicting two of the three combustion sources assigned by PMF. Figure 4.9 shows histograms of the combustion source contributions for all the samples, excluding the method blanks in our dataset. The vegetative burning source (Figure 4.9A) is the largest contributor to PM mass in the samples and has the most uniform distribution. Mobile (Figure 4.9B) shows lower overall concentrations than vegetative burning and the distribution is slightly skewed towards lower concentration. Fuel oil derived PM (Figure 4.9C) shows a similar concentration range to mobile with a very skewed distribution toward low concentration. Fuel oil PM was an infrequent contributor (nearly zero in 25/95 samples, excluding blanks) and, when present, it was at a low concentration. The limit of detection for diesel soot was 2.2 µg/mL (Table 3-1) and fifty-three of the fuel oil PM samples were below this level. I believe the fact that fuel oil is present at low concentrations is the reason the EEM-machine learning method failed to detect it.

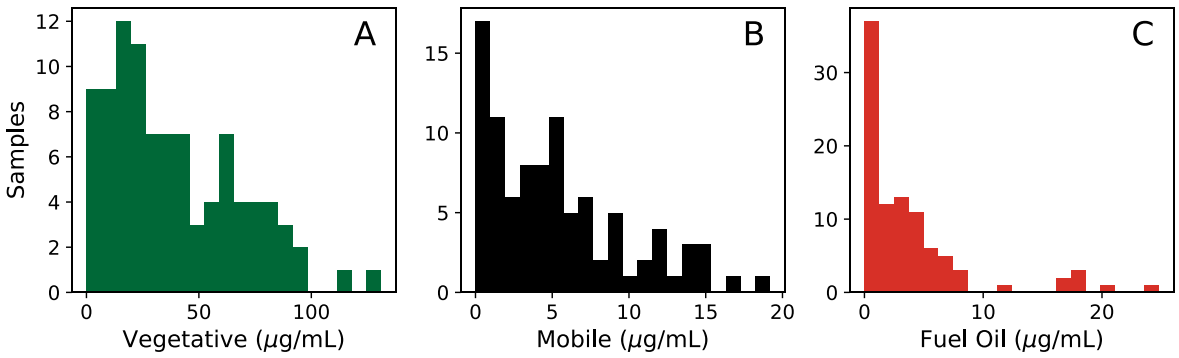


Figure 4.9: Histograms showing the number of samples vs. PM extract concentration (process blanks are not included, n=95). Vegetative burning (A) has the most uniform distribution as well as the highest total concentrations. Mobile (B) has lower overall concentrations a distribution somewhat skewed toward low concentration. Fuel Oil (C) has approximately the same concentration range as mobile but the distribution is significantly skewed towards low concentration. I believe the EEM-machine learning method could detect fuel oil, but it was not often a contributor to our samples and when present its concentration was low.

4.8 Summary

The EEM-machine learning approach was successful in predicting vegetative burning and mobile sources. Table 4-1 summarizes the results of the PCR and CNN modeling approaches. The CNN shows a slight improvement for mobile and slightly worse performance for vegetative burning compared to the PCR model. The CNN model provides more balanced results showing R^2 values closer to one another than PCR, however, the difference is minimal, and the added complexity of the CNN model leads me to recommended PCR in this specific case. However, in other cases, such as the laboratory sources discussed in chapter 3 the CNN approach provided the best results. The EEM-machine learning approach to source apportionment can be used to complement and expand existing source apportionment studies.

Table 4-1: R^2 for PCR and CNN Modeling Approaches

Source	**PCR** R^2	**CNN** R^2
Vegetative Burning	0.84	0.81
Mobile (Traffic)	0.49	0.52
Others	< 0.1	< 0.2

Chapter 5: Conclusions and Future Work

Exposure to particulate matter air pollution is the world's largest environmental health risk accounting for millions of premature deaths and disability-adjusted life years annually. PM originates from natural and anthropogenic sources such as dust from soil, combustion engines, power generation, and forest fires. In some parts of the world, PM pollution is well-controlled but in others, the air is polluted to levels considered hazardous to human health. Quantifying the contributions of individual pollution sources can improve air quality and allows the study of health effects caused by different sources of PM pollution. There are several existing methods for source apportionment of PM air pollution, but no existing one of these methods is used broadly because of their complexity and cost.

In this dissertation, I have presented a new source apportionment method using EEM fluorescence spectroscopy and machine learning. I showed how this method may be used to complement existing source apportionment methods for combustion generated PM. I describe how this method was developed using PM collected from controlled sources in the laboratory. I also applied EEM-machine learning as a source apportionment method to environmental samples illustrating the utility of this new source apportionment tool.

Particulate matter from controlled combustion sources was sampled to show that EEM fluorescence spectroscopy could be used to detect PM from combustion sources at concentrations below those of concern to human health. Using total integrated fluorescence, I showed the LoD for cigarette smoke, diesel exhaust particulate, and wood smoke from clean cookstoves to be 0.7, 3.5 and 0.4 $\mu g/m^3$ of air, respectively. The measured LoDs are well below the USEPA and WHO guidelines for particulate matter exposure showing that EEM fluorescence spectroscopy is a viable tool for measuring combustion generated PM.

The EEM fluorescent spectra observed from the controlled PM sources had unique fingerprints. The uniqueness of the EEM spectra allowed for source apportionment of cigarette, diesel, and woodsmoke sources using a convolutional neural network. I showed that a CNN was able to apportion the three sources with R^2 values of 0.86 for cigarette, 0.79 for diesel, and 0.89 for wood smoke. Using the model-predicted concentration, in place of total integrated fluorescence, I showed that the LoD for cigarette was unchanged, for diesel it decreased (from 3.5 to 2.6 $\mu g/m^3$), and for wood smoke, it decreased (0.4 to 0.9 $\mu g/m^3$). These changes show the opposing effects of improved performance as a result of the model to learning the important features of the spectra and the decrease in performance that is the result of the model being trained to deal with mixtures.

I tested the limits of the model developed using controlled sources by analyzing two sets of environmental samples. One set of environmental samples was obtained from areas expected to be dominated by one of the three sources studied in the laboratory. Analysis of these samples showed the EEM-CNN method for source apportionment to be effective if the underlying sources are known. The other set of environmental samples were obtained from areas where the underlying sources were unknown. In this sample set, the presence of cigarette was often predicted when there was no reason to believe it was present. These samples demonstrate that the measurement of expected underlying sources is not easily generalizable to all ambient samples.

An example of an environmental sample from an area expected to be dominated by one of the controlled laboratory sources is a sample taken in the UW clean cookstoves lab. In this sample, woodsmoke was identified as the primary source. This illustrates how the method could be used for application in occupational exposure. If a workplace had multiple sources of combustion PM, each source could be sampled individually, and a source apportionment model could be trained to

measure the amount of each of the combustion sources a worker was exposed to using personal exposure samplers.

Using the lessons learned from the application of the EEM-CNN model trained on controlled sources, I examined a larger set of environmental samples. These samples were collected during an exposure assessment panel study and previously analyzed for elemental composition. Using the composition measurements, Larson et al. identified six contributing sources with PMF modeling.[36] Three of these PM sources were combustion derived: traffic, vegetative burning, and fuel oil. I expected an EEM-machine learning approach could be used to apportion the combustion sources. I collected EEMs corresponding to the samples analyzed by PMF from extracts of archived filter samples. These EEM spectra and the associated PMF source contributions were used for PCR and training a CNN to identify the sources identified by PMF. Both machine learning methods were capable of identifying the traffic and vegetative burning PM sources but failed to identify the fuel oil PM source.

To apply the EEM-machine learning approach to diverse environmental samples, an orthogonal source apportionment method should be used to train the machine learning algorithms. A potential air pollution study using the EEM-machine learning method could collect air samples at fixed sites and measure elemental composition and EEMs for these samples. In parallel to the fixed site samples, another sample set consisting of more numerous personal samples could be collected and only analyzed by EEM. Using the fixed site samples a source apportionment model could be developed and the personal samples could be efficiently analyzed by EEM to apportion the combustion sources present.

The machine learning source apportionment approach described in this thesis may be extended to include other data sources beyond EEMs. Other spectroscopic measurements that are

based on underlying chemical composition such as infrared and Raman could be used independently or in addition to EEM. Additionally, any other type of air pollution measurement could be used to improve source apportionment tools. Examples include PM size distributions and measurements of gaseous air pollutants

References:

(1) Stanaway, J. D.; Afshin, A.; Gakidou, E.; Lim, S. S.; Abate, D.; Abate, K. H.; Abbafati, C.; Abbasi, N.; Abbastabar, H.; Abd-Allah, F.; et al. Global, Regional, and National Comparative Risk Assessment of 84 Behavioural, Environmental and Occupational, and Metabolic Risks or Clusters of Risks for 195 Countries and Territories, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017. *The Lancet* **2018**, *392* (10159), 1923–1994. https://doi.org/10.1016/S0140-6736(18)32225-6.

(2) Brauer, M.; Freedman, G.; Frostad, J.; van Donkelaar, A.; Martin, R. V.; Dentener, F.; Dingenen, R. van; Estep, K.; Amini, H.; Apte, J. S.; et al. Ambient Air Pollution Exposure Estimation for the Global Burden of Disease 2013. *Environ. Sci. Technol.* **2016**, *50* (1), 79–88. https://doi.org/10.1021/acs.est.5b03709.

(3) World Health Organization Air Pollution Fact Sheet https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health (accessed Dec 5, 2019).

(4) West, J. J.; Cohen, A.; Dentener, F.; Brunekreef, B.; Zhu, T.; Armstrong, B.; Bell, M. L.; Brauer, M.; Carmichael, G.; Costa, D. L.; et al. "What We Breathe Impacts Our Health: Improving Understanding of the Link between Air Pollution and Health." *Environ. Sci. Technol.* **2016**, *50* (10), 4895–4904. https://doi.org/10.1021/acs.est.5b03827.

(5) Pope, C. A.; Ezzati, M.; Dockery, D. W. Fine-Particulate Air Pollution and Life Expectancy in the United States. *N. Engl. J. Med.* **2009**, *360* (4), 376–386. https://doi.org/10.1056/NEJMsa0805646.

(6) Godish, T. *Air Quality*, 4th ed.; Lewis Publishers: Boca Raton, 2004.

(7) US EPA, O. NAAQS Table https://www.epa.gov/criteria-air-pollutants/naaqs-table (accessed Aug 31, 2018).

(8) Technical Assistance Document for the Reporting of Daily Air Quality – the Air Quality Index (AQI). US EPA May 2016.

(9) Tsuda, A.; Henry, F. S.; Butler, J. P. Particle Transport and Deposition: Basic Physics of Particle Kinetics. *Compr. Physiol.* **2013**, *3* (4), 1437–1471. https://doi.org/10.1002/cphy.c100085.

(10) Miller, F. J.; Gardner, D. E.; Graham, J. A.; Lee, R. E.; Wilson, W. E.; Bachmann, J. D. Size Considerations for Establishing a Standard for Inhalable Particles. *J. Air Pollut. Control Assoc.* **1979**, *29* (6), 610–615. https://doi.org/10.1080/00022470.1979.10470831.

(11) Brown, J. S.; Gordon, T.; Price, O.; Asgharian, B. Thoracic and Respirable Particle Definitions for Human Health Risk Assessment. *Part. Fibre Toxicol.* **2013**, *10*, 12. https://doi.org/10.1186/1743-8977-10-12.

(12) US EPA, O. Table of Historical Particulate Matter (PM) National Ambient Air Quality Standards (NAAQS) https://www.epa.gov/pm-pollution/table-historical-particulate-matter-pm-national-ambient-air-quality-standards-naaqs (accessed May 3, 2018).

(13) World Health Organization. Regional Office for Europe. *Air Quality Guidelines: Global Update 2005 : Particulate Matter, Ozone, Nitrogen Dioxide, and Sulfur Dioxide.*; World Health Organization Europe: Copenhagen, Denmark, 2006.

(14) US EPA, O. Particulate Matter (PM) Basics https://www.epa.gov/pm-pollution/particulate-matter-pm-basics (accessed Aug 23, 2018).

(15) Heal, M. R.; Kumar, P.; Harrison, R. M. Particles, Air Quality, Policy and Health. *Chem. Soc. Rev.* **2012**, *41* (19), 6606–6630. https://doi.org/10.1039/c2cs35076a.

(16) Dai, L.; Zanobetti, A.; Koutrakis, P.; Schwartz, J. D. Associations of Fine Particulate Matter Species with Mortality in the United States: A Multicity Time-Series Analysis. *Environ. Health Perspect.* **2014**, *122* (8), 837–842. https://doi.org/10.1289/ehp.1307568.

(17) Washington State Department of Ecology - Air Monitoring Sites - Seattle-Beacon Hill https://fortress.wa.gov/ecy/enviwa/StationInfo.aspx?ST_ID=42 (accessed Sep 25, 2016).

(18) U.S. Department of State Air Quality Monitoring Program - Beijing Embassy http://www.stateair.net/web/post/1/1.html (accessed Sep 25, 2016).

(19) Janssen, N. A. H.; Hoek, G.; Simic-Lawson, M.; Fischer, P.; Van Bree, L.; Ten Brink, H.; Keuken, M.; Atkinson, R. W.; Anderson, H. R.; Brunekreef, B.; et al. Black Carbon as an Additional Indicator of the Adverse Health Effects of Airborne Particles Compared with PM10 and PM2.5. *Environ. Health Perspect.* **2011**, *119* (12), 1691–1699. https://doi.org/10.1289/ehp.1003369.

(20) Bell, M. L.; Ebisu, K.; Leaderer, B. P.; Gent, J. F.; Lee, H. J.; Koutrakis, P.; Wang, Y.; Dominici, F.; Peng, R. D. Associations of $PM_{2.5}$ Constituents and Sources with Hospital Admissions: Analysis of Four Counties in Connecticut and Massachusetts (USA) for Persons ≥ 65 Years of Age. *Environ. Health Perspect.* **2014**, *122* (2), 138–144. https://doi.org/10.1289/ehp.1306656.

(21) Stanek, L. W.; Sacks, J. D.; Dutton, S. J.; Dubois, J.-J. B. Attributing Health Effects to Apportioned Components and Sources of Particulate Matter: An Evaluation of Collective Results. *Atmos. Environ.* **2011**, *45* (32), 5655–5663. https://doi.org/10.1016/j.atmosenv.2011.07.023.

(22) Adams, K.; Greenbaum, D. S.; Shaikh, R.; Erp, A. M. van; Russell, A. G. Particulate Matter Components, Sources, and Health: Systematic Approaches to Testing Effects. *J.*

Air Waste Manag. Assoc. **2015**, *65* (5), 544–558. https://doi.org/10.1080/10962247.2014.1001884.

(23) Hime, N.; Marks, G.; Cowie, C.; Hime, N. J.; Marks, G. B.; Cowie, C. T. A Comparison of the Health Effects of Ambient Particulate Matter Air Pollution from Five Emission Sources. *Int. J. Environ. Res. Public. Health* **2018**, *15* (6), 1206. https://doi.org/10.3390/ijerph15061206.

(24) Koenig, J. Q. Air Pollution and Asthma. *J. Allergy Clin. Immunol.* **1999**, *104* (4), 717–722. https://doi.org/10.1016/S0091-6749(99)70280-0.

(25) National Asthma Education and Prevention Program. Expert Panel Report 3 (EPR-3): Guidelines for the Diagnosis and Management of Asthma–Summary Report 2007. *J. Allergy Clin. Immunol.* **2007**, *120* (5), S94–S138. https://doi.org/10.1016/j.jaci.2007.09.029.

(26) Lewis, S. A.; Antoniak, M.; Venn, A. J.; Davies, L.; Goodwin, A.; Salfield, N.; Britton, J.; Fogarty, A. W. Secondhand Smoke, Dietary Fruit Intake, Road Traffic Exposures, and the Prevalence of Asthma: A Cross-Sectional Study in Young Children. *Am. J. Epidemiol.* **2005**, *161* (5), 406–411. https://doi.org/10.1093/aje/kwi059.

(27) Hopke, P. K. Review of Receptor Modeling Methods for Source Apportionment. *J. Air Waste Manag. Assoc.* **2016**, *66* (3), 237–259. https://doi.org/10.1080/10962247.2016.1140693.

(28) Ward, T. J.; Palmer, C. P.; Noonan, C. W. Fine Particulate Matter Source Apportionment Following a Large Woodstove Changeout Program in Libby, Montana. *J. Air Waste Manag. Assoc.* **2010**, *60* (6), 688–693. https://doi.org/10.3155/1047-3289.60.6.688.

(29) Samburova, V.; Connolly, J.; Gyawali, M.; Yatavelli, R. L. N.; Watts, A. C.; Chakrabarty, R. K.; Zielinska, B.; Moosmüller, H.; Khlystov, A. Polycyclic Aromatic Hydrocarbons in Biomass-Burning Emissions and Their Contribution to Light Absorption and Aerosol Toxicity. *Sci. Total Environ.* **2016**, *568*, 391–401. https://doi.org/10.1016/j.scitotenv.2016.06.026.

(30) Lin, P.; Fleming, L. T.; Nizkorodov, S. A.; Laskin, J.; Laskin, A. Comprehensive Molecular Characterization of Atmospheric Brown Carbon by High Resolution Mass Spectrometry with Electrospray and Atmospheric Pressure Photoionization. *Anal. Chem.* **2018**, *90* (21). https://doi.org/10.1021/acs.analchem.8b02177.

(31) US EPA. Air Toxics Monitoring Methods, Compendium Method TO-13, Determination of Polycyclic Aromatic Hydrocarbons (PAHs) in Ambient Air Using Gas Chromatography/Mass Spectrometry (GC/MS) https://www3.epa.gov/ttnamti1/airtox.html.

(32) Brown, A. S.; Brown, R. J. C. Correlations in Polycyclic Aromatic Hydrocarbon (PAH) Concentrations in UK Ambient Air and Implications for Source Apportionment. *J. Environ. Monit.* **2012**, *14* (8), 2072–2082. https://doi.org/10.1039/C2EM10963H.

(33) Straif, K.; Baan, R.; Grosse, Y.; Secretan, B.; El Ghissassi, F.; Cogliano, V. Carcinogenicity of Polycyclic Aromatic Hydrocarbons. *Lancet Oncol.* **2005**, *6* (12), 931–932. https://doi.org/10.1016/S1470-2045(05)70458-7.

(34) EPA. Compendium Method IO-3.3, DETERMINATION OF METALS IN AMBIENT PARTICULATE MATTER USING X-RAY FLUORESCENCE (XRF) SPECTROSCOPY. EPA/625/R-96/010a. 1999.

(35) Marguí, E. *X-Ray Fluorescence Spectrometry and Related Techniques: An Introduction*; Momentum Press: New York, New York, 2013.

(36) Larson, T.; Gould, T.; Simpson, C.; Liu, L.-J. S.; Claiborn, C.; Lewtas, J. Source Apportionment of Indoor, Outdoor, and Personal PM2.5 in Seattle, Washington, Using Positive Matrix Factorization. *J. Air Waste Manag. Assoc.* **2004**, *54* (9), 1175–1187. https://doi.org/10.1080/10473289.2004.10470976.

(37) Birch, M. E.; Cary, R. A. Elemental Carbon-Based Method for Monitoring Occupational Exposures to Particulate Diesel Exhaust. *Aerosol Sci. Technol.* **1996**, *25* (3), 221–241. https://doi.org/10.1080/02786829608965393.

(38) Watson, J. G.; Chow, J. C.; Lowenthal, D. H.; Pritchett, L. C.; Frazier, C. A.; Neuroth, G. R.; Robbins, R. Differences in the Carbon Composition of Source Profiles for Diesel- and Gasoline-Powered Vehicles. *Atmos. Environ.* **1994**, *28* (15), 2493–2505. https://doi.org/10.1016/1352-2310(94)90400-6.

(39) Kim, E.; Hopke, P. K. Source Apportionment of Fine Particles in Washington, DC, Utilizing Temperature-Resolved Carbon Fractions. *J. Air Waste Manag. Assoc.* **2004**, *54* (7), 773–785. https://doi.org/10.1080/10473289.2004.10470948.

(40) Watson, J.; Chow, J.; Chen, L.-W. Summary of Organic and Elemental Carbon/Black Carbon Analysis Methods and Intercomparisons. *Aerosol Air Qual. Resarch* **2005**, *5* (1), 65–102. https://doi.org/10.4209/aaqr.2005.06.0006.

(41) Simpson, C. D.; Dills, R. L.; Katz, B. S.; Kalman, D. A. Determination of Levoglucosan in Atmospheric Fine Particulate Matter. *J. Air Waste Manag. Assoc.* **2004**, *54* (6).

(42) Moerner, W. E.; Fromm, D. P. Methods of Single-Molecule Fluorescence Spectroscopy and Microscopy. *Rev. Sci. Instrum.* **2003**, *74* (8), 3597–3619. https://doi.org/10.1063/1.1589587.

(43) Elcoroaristizabal, S.; de Juan, A.; García, J. A.; Durana, N.; Alonso, L. Comparison of Second-Order Multivariate Methods for Screening and Determination of PAHs by Total Fluorescence Spectroscopy. *Chemom. Intell. Lab. Syst.* **2014**, *132*, 63–74. https://doi.org/10.1016/j.chemolab.2014.01.005.

(44) Nahorniak, M. L.; Booksh, K. S. Excitation-Emission Matrix Fluorescence Spectroscopy in Conjunction with Multiway Analysis for PAH Detection in Complex Matrices. *Analyst* **2006**, *131* (12), 1308–1315. https://doi.org/10.1039/B609875D.

(45) Johnson, D. W.; Callis, J. B.; Christian, G. D. Rapid Scanning Fluorescence Spectroscopy. *Anal. Chem.* **1977**, *49* (8), 747A-757A. https://doi.org/10.1021/ac50016a769.

(46) Andrade-Eiroa, Á.; Canle, M.; Cerdá, V. Environmental Applications of Excitation-Emission Spectrofluorimetry: An In-Depth Review II. *Appl. Spectrosc. Rev.* **2013**, *48* (2), 77–141. https://doi.org/10.1080/05704928.2012.692105.

(47) Aryal, R.; Lee, B.-K.; Beecham, S.; Kandasamy, J.; Aryal, N.; Parajuli, K. Characterisation of Road Dust Organic Matter as a Function of Particle Size: A PARAFAC Approach. *Water. Air. Soil Pollut.* **2015**, *226* (2), 24. https://doi.org/10.1007/s11270-014-2289-y.

(48) Chen, Q.; Miyazaki, Y.; Kawamura, K.; Matsumoto, K.; Coburn, S.; Volkamer, R.; Iwamoto, Y.; Kagami, S.; Deng, Y.; Ogawa, S.; et al. Characterization of Chromophoric Water-Soluble Organic Matter in Urban, Forest, and Marine Aerosols by HR-ToF-AMS

Analysis and Excitation–Emission Matrix Spectroscopy. *Environ. Sci. Technol.* **2016**, *50* (19), 10351–10360. https://doi.org/10.1021/acs.est.6b01643.

(49) Elcoroaristizabal, S.; Juan, A. de; García, J. A.; Elorduy, I.; Durana, N.; Alonso, L. Chemometric Determination of PAHs in Aerosol Samples by Fluorescence Spectroscopy and Second-Order Data Analysis Algorithms. *J. Chemom.* **2014**, *28* (4), 260–271. https://doi.org/10.1002/cem.2604.

(50) Matos, J. T. V.; Freire, S. M. S. C.; Duarte, R. M. B. O.; Duarte, A. C. Natural Organic Matter in Urban Aerosols: Comparison between Water and Alkaline Soluble Components Using Excitation–Emission Matrix Fluorescence Spectroscopy and Multiway Data Analysis. *Atmos. Environ.* **2015**, *102*, 1–10. https://doi.org/10.1016/j.atmosenv.2014.11.042.

(51) Mladenov, N.; Alados-Arboledas, L.; Olmo, F. J.; Lyamani, H.; Delgado, A.; Molina, A.; Reche, I. Applications of Optical Spectroscopy and Stable Isotope Analyses to Organic Aerosol Source Discrimination in an Urban Area. *Atmos. Environ.* **2011**, *45* (11), 1960–1969. https://doi.org/10.1016/j.atmosenv.2011.01.029.

(52) Mladenov, N.; Lípez-Ramos, J.; McKnight, D. M.; Rechea, I. Alpine Lake Optical Properties as Sentinels of Dust Deposition and Global Change. *Limnol. Oceanogr.* **2009**, *54* (6part2), 2386–2400. https://doi.org/10.4319/lo.2009.54.6_part_2.2386.

(53) Nakajima, H.; Okada, K.; Kuroki, Y.; Nakama, Y.; Handa, D.; Arakaki, T.; Tanahara, A. Photochemical Formation of Peroxides and Fluorescence Characteristics of the Water-Soluble Fraction of Bulk Aerosols Collected in Okinawa, Japan. *Atmos. Environ.* **2008**, *42* (13), 3046–3058. https://doi.org/10.1016/j.atmosenv.2007.12.045.

(54) Muñoz de la Peña, A.; Durán-Merás, I.; Moreno, M. D.; Salinas, F.; Galera, M. M. Resolution of Ternary Mixtures of Salicylic, Salicyluric and Gentisic Acids by Partial Least Squares and Principal Component Regression: Optimization of the Scanning Path in the Excitation-Emission Matrices. *Fresenius J. Anal. Chem.* **1995**, *351* (6), 571–576. https://doi.org/10.1007/BF00322735.

(55) Alostaz, M.; Biggar, K.; Donahue, R.; Hall, gregory. Petroleum Contamination Characterization and Quantification Using Fluorescence Emission-Excitation Matrices (EEMs) and Parallel Factor Analysis (PARAFAC). *J. Environ. Eng. Sci.* **2008**, *7* (3), 183–197. https://doi.org/10.1139/S07-049.

(56) Rawat, W.; Wang, Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. *Neural Comput.* **2017**, *29* (9), 2352–2449. https://doi.org/10.1162/neco_a_00990.

(57) Salman, A. G.; Kanigoro, B.; Heryadi, Y. Weather Forecasting Using Deep Learning Techniques. In *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*; 2015; pp 281–285. https://doi.org/10.1109/ICACSIS.2015.7415154.

(58) Kim, Y. Convolutional Neural Networks for Sentence Classification. *ArXiv14085882 Cs* **2014**.

(59) Swietojanski, P.; Ghoshal, A.; Renals, S. Convolutional Neural Networks for Distant Speech Recognition. *IEEE Signal Process. Lett.* **2014**, *21* (9), 1120–1124. https://doi.org/10.1109/LSP.2014.2325781.

(60) Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *ArXiv14090575 Cs* **2014**.

(61) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*; NIPS'12; Curran Associates Inc.: USA, 2012; pp 1097–1105.

(62) Dumoulin, V.; Visin, F. A Guide to Convolution Arithmetic for Deep Learning. *ArXiv160307285 Cs Stat* **2016**.

(63) Rosenblatt, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychol. Rev.* **1958**, *65* (6), 386–408. https://doi.org/10.1037/h0042519.

(64) Hastie, T. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second edition, [corrected at 7th printing].; Springer series in statistics; Springer: New York, NY, USA, 2013.

(65) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436–444. https://doi.org/10.1038/nature14539.

(66) Prechelt, L. Early Stopping — But When? In *Neural Networks: Tricks of the Trade: Second Edition*; Montavon, G., Orr, G. B., Müller, K.-R., Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; pp 53–67. https://doi.org/10.1007/978-3-642-35289-8_5.

(67) Hinton, G. E.; Srivastava, N.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. R. Improving Neural Networks by Preventing Co-Adaptation of Feature Detectors. *ArXiv12070580 Cs* **2012**.

(68) Di, Q.; Kloog, I.; Koutrakis, P.; Lyapustin, A.; Wang, Y.; Schwartz, J. Assessing PM2.5 Exposures with High Spatiotemporal Resolution across the Continental United States. *Environ. Sci. Technol.* **2016**, *50* (9), 4712–4721. https://doi.org/10.1021/acs.est.5b06121.

(69) Song, X.-H.; Hopke, P. K. Solving the Chemical Mass Balance Problem Using an Artificial Neural Network. *Environ. Sci. Technol.* **1996**, *30* (2), 531–535. https://doi.org/10.1021/es950281o.

(70) Carstea, E. M.; Baker, A.; Bieroza, M.; Reynolds, D. Continuous Fluorescence Excitation–Emission Matrix Monitoring of River Organic Matter. *Water Res.* **2010**, *44* (18), 5356–5366. https://doi.org/10.1016/j.watres.2010.06.036.

(71) Bieroza, M.; Baker, A.; Bridgeman, J. Exploratory Analysis of Excitation-Emission Matrix Fluorescence Spectra with Self-Organizing Maps as a Basis for Determination of Organic Matter Removal Efficiency at Water Treatment Works. *J. Geophys. Res. Biogeosciences* **2009**, *114* (G4). https://doi.org/10.1029/2009JG000940.

(72) García-Reiriz, A.; Damiani, P. C.; Olivieri, A. C.; Cañada-Cañada, F.; Muñoz de la Peña, A. Nonlinear Four-Way Kinetic-Excitation−Emission Fluorescence Data Processed by a Variant of Parallel Factor Analysis and by a Neural Network Model Achieving the Second-Order Advantage: Malonaldehyde Determination in Olive Oil Samples. *Anal. Chem.* **2008**, *80* (19), 7248–7256. https://doi.org/10.1021/ac8007829.

(73) García-Reiriz, A.; Damiani, P. C.; Olivieri, A. C. Analysis of Amoxicillin in Human Urine by Photo-Activated Generation of Fluorescence Excitation–Emission Matrices and

Artificial Neural Networks Combined with Residual Bilinearization. *Anal. Chim. Acta* **2007**, *588* (2), 192–199. https://doi.org/10.1016/j.aca.2007.02.020.

(74) James, G. G. M. *An Introduction to Statistical Learning: With Applications in R*; Springer texts in statistics ; 103; Springer: New York, NY, 2013.

(75) Sullivan, B.; Allawatt, G.; Emery, A.; Means, P.; Kramlich, J.; Posner, J. Time-Resolved Particulate Emissions Monitoring of Cookstove Biomass Combustion Using a Tapered Element Oscillating Microbalance. *Combust. Sci. Technol.* **2017**, *189* (6), 923–936. https://doi.org/10.1080/00102202.2016.1253564.

(76) Gould, T.; Larson, T.; Stewart, J.; Kaufman, J. D.; Slater, D.; Mcewen, N. A Controlled Inhalation Diesel Exhaust Exposure Facility with Dynamic Feedback Control of PM Concentration. *Inhal. Toxicol.* **2008**, *20* (1), 49–52. https://doi.org/10.1080/08958370701758478.

(77) Engelborghs, Y.; Visser, A. J. W. G. *Fluorescence Spectroscopy and Microscopy: Methods and Protocols*; Methods in molecular biology (Clifton, N.J.) ; v. 1076; Humana Press: New York, 2014.

(78) Murphy, K. R.; Butler, K. D.; Spencer, R. G. M.; Stedmon, C. A.; Boehme, J. R.; Aiken, G. R. Measurement of Dissolved Organic Matter Fluorescence in Aquatic Environments: An Interlaboratory Comparison. *Environ. Sci. Technol.* **2010**, *44* (24), 9405–9412. https://doi.org/10.1021/es102362t.

(79) Zepp, R. G.; Sheldon, W. M.; Moran, M. A. Dissolved Organic Fluorophores in Southeastern US Coastal Waters: Correction Method for Eliminating Rayleigh and Raman Scattering Peaks in Excitation–Emission Matrices. *Mar. Chem.* **2004**, *89* (1), 15–36. https://doi.org/10.1016/j.marchem.2004.02.006.

(80) Tholen, D. W. *Protocols for Determination of Limits of Detection and Limits of Quantitation: Approved Guideline*; NCCLS: Wayne, Pa., 2004.

(81) Borysiak, M. D.; Thompson, M. J.; Posner, J. D. Translating Diagnostic Assays from the Laboratory to the Clinic: Analytical and Clinical Metrics for Device Development and Evaluation. *Lab. Chip* **2016**, *16* (8), 1293–1313. https://doi.org/10.1039/c6lc00015k.

(82) Lakowicz, J. R. *Principles of Fluorescence Spectroscopy*, 3rd ed.; Springer: New York, 2006.

(83) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J Mach Learn Res* **2014**, *15* (1), 1929–1958.

(84) Clevert, D.-A.; Unterthiner, T.; Hochreiter, S. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *ArXiv151107289 Cs* **2015**.

(85) Chollet, F. Keras https://keras.io (accessed Sep 21, 2018).

(86) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*; 2015.

(87) Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; Wattenberg, M. SmoothGrad: Removing Noise by Adding Noise. *ArXiv170603825 Cs Stat* **2017**.

(88) Puget Sound Clean Air Agency - Air Graphing Tool https://secure.pscleanair.org/airgraphing (accessed Jan 22, 2019).

(89) James, G. G. M. *An Introduction to Statistical Learning: With Applications in R*; Springer texts in statistics ; 103; Springer: New York, NY, 2013.

(90) Jacob A. Wegelin. A Survey of Partial Least Squares (PLS) Methods, with Emphasis on the Two-Block Case | University of Washington Department of Statistics https://www.stat.washington.edu/index.php/article/tech-report/survey-partial-least-squares-pls-methods-emphasis-two-block-case (accessed Dec 1, 2018).

(91) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(92) Liu, L. J. S.; Box, M.; Kalman, D.; Kaufman, J.; Koenig, J.; Larson, T.; Lumley, T.; Sheppard, L.; Wallace, L. Exposure Assessment of Particulate Matter for Susceptible Populations in Seattle. *Environ. Health Perspect.* **2003**, *111* (7), 909–918.

(93) Koenig, J. Q.; Jansen, K.; Mar, T. F.; Lumley, T.; Kaufman, J.; Trenga, C. A.; Sullivan, J.; Liu, L.-J. S.; Shapiro, G. G.; Larson, T. V. Measurement of Offline Exhaled Nitric Oxide in a Study of Community Exposure to Air Pollution. *Environ. Health Perspect.* **2003**, *111* (13), 1625–1629.

(94) Simpson, C. D.; Dills, R. L.; Katz, B. S.; Kalman, D. A. Determination of Levoglucosan in Atmospheric Fine Particulate Matter. *J. Air Waste Manag. Assoc.* **2004**, *54* (6).

(95) Borysiak, M. D.; Thompson, M. J.; Posner, J. D. Translating Diagnostic Assays from the Laboratory to the Clinic: Analytical and Clinical Metrics for Device Development and Evaluation. *Lab. Chip* **2016**, *16* (8), 1293–1313. https://doi.org/10.1039/c6lc00015k.

(96) Gould, T.; Larson, T.; Stewart, J.; Kaufman, J. D.; Slater, D.; Mcewen, N. A Controlled Inhalation Diesel Exhaust Exposure Facility with Dynamic Feedback Control of PM Concentration. *Inhal. Toxicol.* **2008**, *20* (1), 49–52. https://doi.org/10.1080/08958370701758478.

(97) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(98) Bring, J. How to Standardize Regression Coefficients. *Am. Stat.* **1994**, *48* (3), 209–213. https://doi.org/10.2307/2684719.

(99) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *J. Mach. Learn. Res.* **2014**, *15* (1), 1929–1958.

(100) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. **2014**.