UNIVERSITY OF CALIFORNIA
RIVERSIDE

Development and Application of Statistical Methods for Analysis of Volatile Organic Compounds From Emerging Sources

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Chemical and Environmentnal Engineering

by

Christos Stamatis

December 2021

Dissertation Committee:

Dr. Kelley C. Barsanti, Chairperson
Dr. David R. Cocker III
Dr. William C. Porter

The Dissertation of Christos Stamatis is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I am grateful to my advisor, without whose help, I would not have been here.

To my parents for all the support.

ABSTRACT OF THE DISSERTATION

Development and Application of Statistical Methods for Analysis of Volatile Organic Compounds From Emerging Sources

by

Christos Stamatis

Doctor of Philosophy, Graduate Program in Chemical and Environmentnal Engineering
University of California, Riverside, December 2021
Dr. Kelley C. Barsanti, Chairperson

Non-methane organic compounds (NMOCs) play an important role as air pollutant precursors. For example, NMOCs can form secondary organic aerosol (SOA), which is a major component of fine particulate matter ($PM_{2.5}$). As a significant mass fraction of $PM_{2.5}$, SOA affects air quality, visibility, radiative forcing and cloud droplet formation. NMOCs can be categorized as biogenic (e.g., forest emissions), anthropogenic (e.g., transportation emissions) or pyrogenic (e.g., wildfire emissions). Given the importance of NMOCs for atmospheric chemistry and air pollutant formation, there is ongoing need to track NMOCs and understand how variables such as source types, meteorology, and human behavior affect NMOC mixing ratios and SOA formation. Due to technological advances in analytical and instrumental techniques, including those used to quantify NMOCs, there has been a large increase in NMOC data generated. This poses a challenge for efficient and accurate data analysis, and thus presents an opportunity for statistical data analysis methods that are complementary to more conventional approaches. As presented in this thesis, two tools for faster NMOC data post-processing and chemometric analyses were created: 1) a chromatographic alignment algorithm, and 2) a supervised pattern recognition algorithm for classification and source apportionment analyses of emissions from emerging anthropogenic (e.g., vehicles, personal care products) and pyrogenic (e.g., wildfires) sources of interest in the formation of air pollutants.

The chromatographic alignment algorithm corrects any retention time deviations that occur due to matrix effects during instrumental analysis and is a necessary preprocessing step for the pattern recognition analysis. The alignment is performed using a simple measure of similarity, specifically, the cosine similarity. The supervised pattern recognition algorithm can unveil relationships between samples based on parameters of interest such as source types, and for biomass burning, fuel types. It is comprised of three main parts: 1) feature selection with an analysis of variance (ANOVA) based method, 2) dimensionality reduction via principal components analysis (PCA), and 3) clustering with k-means. The chromatographic alignment algorithm was applied in an analysis of engine tailpipe emissions samples; the algorithm successfully aligned approximately 110 detected NMOCs in 32 emission samples and reduced post-processing time from weeks to minutes. The pattern recognition algorithm was applied in three case studies involving analysis of tailpipe emissions samples (anthropogenic) and biomass burning samples (pyrogenic).

Applied to the tailpipe emissions samples, the algorithm identified 15 NMOCs (out of 110) that effectively separated the tailpipe emission samples among eight different gasoline fuel blends, based on patterns in the emitted NMOCs associated with the different fuel blends. Applied to biomass burning samples, the algorithm was able to successfully differentiate laboratory smoke samples collected from combustion of three different fuel families (firs, pines and spruce) using automated selection of only five compounds (out of 93). Furthermore, using the unique NMOC profiles associated with the different fuel families, a classification model was created that successfully classified smoke samples from prescribed burns based on their dominant fuel source. In a second biomass burning case study, the algorithm was applied to data collected from smoke plume transects during two recent large-scale field campaigns (WE-CAN and FIREX-AQ), in an effort to identify NMOCs that were linked with observed enhancement in organic aerosol mass. Samples were separated based on whether positive or negative/neutral organic aerosol enhancement was observed; no consistent set of NMOCs could be identified that successfully differentiated the samples between the two groups. Additional metrics were evaluated, however they did not provide any further ability to differentiate the samples. It was determined that the plume-to-plume variability, and competing chemical and physical processes, overwhelmed clear patterns in NMOC mixing ratios which thus resulted in poor statistical power for differentiating samples. Collectively, the tools developed and presented here provide a streamlined approach for the analysis of samples from diverse sources that reduces the post-processing time and adds information that is difficult to obtain using traditional analytical approaches. The alignment algorithm can be applied to different samples, and the pattern recognition algorithm can be applied to different observational data, thus linking observations with predictive variables including source type, meteorology and human behaviors.

Contents

List of Figures **xi**

List of Tables **xiv**

1 Introduction **1**

1.1 Organic Emissions and Their Influence on Climate and Atmospheric Chemistry 1

1.2 Aromatics in Tailpipe Emissions Samples 3

1.3 Monoterpenes in Smoke Samples 4

1.4 Precursors of Organic Aerosol in Smoke Plumes 5

2 Methods **7**

2.1 Data pre-processing 7

2.1.1 Missing values 9

2.1.2 Data scaling 10

2.2 Feature selection 11

2.2.1 Analysis of variance based feature selection 12

2.3 Dimensionality reduction 14

2.3.1 Principal component analysis steps 15

2.4 Clustering 17

2.4.1 *k*-means clustering 17

2.5 Pattern recognition algorithm 19

2.6 Linear discriminant analysis 20

2.7 Chromatographic alignment 21

2.7.1 Data files 24

2.7.2 Sample merging 24

2.7.3 Mass spectra formatting 25

2.7.4 Alignment 25

3 Pattern Recognition Analysis of Tailpipe Emissions from Gasoline Fuels Using an ANOVA Based Feature Selection, PCA and *k*-means Clustering **29**

3.1 Introduction 29

3.2 Data 31

3.3 Algorithm Implementation, Results and Discussion 33
3.3.1 Data pre-processing 33
3.3.2 Pattern recognition results 33
3.4 Summary 41

4 Development and Application of a Supervised Pattern Recognition Algorithm for Identification of Source-Specific Emissions Profiles 42
4.1 Introduction 42
4.2 Data 46
4.3 Algorithm Implementation, Results and Discussion 49
4.3.1 Sample and fuel type selection for PR and classification 49
4.3.2 Manual versus automated feature selection 50
4.3.3 PCA and *k*-means clustering 52
4.3.4 Beyond principal components one and two 56
4.3.5 Mixed samples 56
4.4 Classification 58
4.4.1 Synthetic mixtures and FLAME-4 FL12 samples 61
4.4.2 Blodgett samples 63
4.5 Summary 69

5 Application of the Supervised Pattern Recognition Algorithm for Identification of SOA Precursors in Ambient Plume Samples 71
5.1 Introduction 71
5.2 Data 73
5.3 Algorithm Implementation, Results and Discussion 74
5.3.1 Data pre-processing 74
5.3.2 Pattern recognition results 76
5.3.3 Dealing with feature dispersion 82
5.4 Summary 84

6 Conclusions 85

Bibliography 88

List of Figures

2.1 PCA results for the demonstrative UC Irvine data set. Left plot shows PCA with no preprocesing (no standardization) and the right plot results after pre-processing. 9
2.2 Plots a-c show the PCA results for a synthetic data set with a varying number of features selected using the ANOVA based method (see Chapter 2.5). . . 13
2.3 Pattern recognition algorithm flowchart 20
2.4 The plot shows the decision boundary (white line) for linear discriminant analysis between two classes (red and blue). The rest of the LDA space is coloured based on the probability of a sample to belong to one of the two classes. Class centroids are pointed with yellow stars. 22
2.5 Chromatographic alignment algorithm flow chart. 23

3.1 Plots [a]-[c] shows the PCA results for peak area threshold 1%, 2% and $F = 20$. 36
3.2 Plots [a]-[b] shows the PCA and k-means clustering results for $F = 20$. The colour coding corresponds to the different cluster from k-means. 37
3.3 Left plots shows cosine similarity for all fuels among the fours clusters except F7 which has its own cluster. Right plot shows cosine similarity for all fuels across different clusters (not all combinations are shows for figure clarity). . 38
3.4 Plots [a]-[b] shows the PCA and k-means clustering results for $F = 20$. The colour coding corresponds to the different cluster from k-means. The red cross markers are the cluster centroids. 40

4.1 Percentage of variance explained per principal component (PC) for the case of automated (blue) and manual (orange) compound selection using Fisher ratios. 51
4.2 Normalized emission ratio profiles for: Douglas fir, subalpine fir, lodgepole pine, ponderosa pine, and Engelmann spruce based on manual selection of compounds. 52
4.3 Normalized emission ratio profiles for: Douglas fir, subalpine fir, lodgepole pine, ponderosa pine, and Engelmann spruce based on automated selection of 10 compounds. 53

4.4 Normalized emission ratio profiles for: Douglas fir, subalpine fir, lodgepole pine, ponderosa pine, and Engelmann spruce based on automated selection of 5 compounds. 54

4.5 Plot [a] shows the elbow plot for k-means clustering with automated compound selection for the PC1 and PC2 pair. The orange marker indicates the optimum number of clusters. Plot [b] shows the PCA coupled with k-means clustering results for the PC1 and PC2 pair. 55

4.6 Plot [a] shows the elbow plot for k-means clustering with automated compound selection for the PC1 and PC4 pair. The orange marker indicates the optimum number of clusters. Plot [b] shows the PCA coupled with k-means clustering results for the PC1 and PC4 pair. Plots [c] shows the spruce overlap with firs based on the PC combination used. 57

4.7 Plot [a] shows the PCA coupled with k-means clustering results for the PC1 and PC4 pair including the 60%/40% synthetic mixtures. Plot [b] shows the PCA coupled with k-means clustering results for the PC1 and PC4 pair including the 90%/10% synthetic mixtures. 59

4.8 Normalized emission ratio profiles for: Douglas fir, subalpine fir, lodgepole pine, ponderosa pine, Engelmann spruce and the 60/60 pine fir synthetic sample. 60

4.9 Classification results for synthetic mixtures and FLAME-4 samples. 62

4.10 Normalized emission ratio profiles for FIREX FL16 samples: pines and firs; and FLAME-4 FL12 samples: sawgrass, wiregrass, rice straw, and Indonesian peat. The relative abundances of camphene and b-pinene in sawgrass, Indonesian peat, and rice straw were > 0.6, but for figure clarity, the axis limits were not changed. 63

4.11 Plots [a]-[c] shows the classification probability by fuel class for plots 60, 340 and 400. 66

4.12 Plots [a]-[b] show the average overstory composition for plots 60 sites 1, 2, and 3. The fractional contribution of each species was determined by the proportion of basal area among all trees with diameter > 11.4 cm at breast height. Data for the overstory composition were taken from the supplemental material in [1]. 67

4.13 Same as Fig. 4.12 but for plots 340 and 400. 68

5.1 Min max normalized levels of toluene, CO and OA for one of the FIREX-AQ flights. The markers on the toluene plot show the start and end points for the calculation of ΔOA and $\Delta \mathrm{NMOC}_{\mathrm{i}}$. 76

5.2 Plot [a] shows the PCA results for five compounds. Plot [b] shows the PCA results for ten compounds. 78

5.3 Plot [a] shows the average $\Delta \mathrm{NMOC}_{\mathrm{i}}$ for 5 compounds. Plot [b] shows the same as [a] but for 10 compounds. 80

5.4 Left plot shows the F-ratio values for all compounds in the data set. Right plot shows the *p*-values for every compound. The red dashed line shows the threshold of statistical significance (0.05). 81

5.5 Plot [a] shows the PCA results for the 5 groups of compounds. Plot [b] shows the average $\Delta\mathrm{NMOC_i}$ for every group between the two classes. 83

List of Tables

3.1 Raw Fuel properties and total aromatic peak area % in emissions. 32

3.2 Selected compounds with peak area threshold and ANOVA. 34

3.3 Clustered fuels and total aromatic peak area % 39

4.1 Fuels Burned and Smoke Analyzed From FLAME-4 2012 Laboratory Fires, FIREX 2016 Laboratory Fires, and Blodgett Forest Research Station Prescribed Burns. 48

4.2 Data sets used for developing the PR algorithm and testing and training the classification algorithm. 49

5.1 Selected compounds for the two cases. 77

Chapter 1

Introduction

1.1 Organic Emissions and Their Influence on Climate and Atmospheric Chemistry

Non-methane organic compounds (NMOCs) play an important role as air pollutant precursors. For example, NMOCs can form secondary organic aerosol (SOA)[2]. As a significant mass fraction of particulate matter[3][4][5], SOA affects air quality, visibility, radiative forcing and cloud droplet formation[6][7][8][9]. NMOCs can be categorized as biogenic (e.g., forest emissions), anthropogenic (e.g., transportation emissions) or pyrogenic (e.g., wildfire emissions). In recent years, wildfire emissions and volatile chemical products (shampoos, surface cleaners, etc.) have become sources of emerging interest[10][11][12][13]. Given the importance of NMOCs for atmospheric chemistry and air pollutant formation, there is ongoing need to track NMOCs and understand how variables such as source types, meteorology, and human behavior affect NMOC mixing ratios and SOA formation.

Due to technological advances in instrumental techniques, including those used to quantify NMOCs, there has been a large increase in data generated. This poses a challenge for efficient analysis of data in their entirety, while also presenting an opportunity to apply statistical data analysis methods that are complementary to more conventional approaches. The primary goal of the work presented in this thesis was to create tools for more efficient and more insightful analysis of two dimensional gas chromatography with time-of-flight mass spectrometry (GC×GC-TOFMS) data. The specific objectives were to: 1) develop a pattern recognition (PR) algorithm that can unveil relationships between samples based on parameters such as source contributions, and for biomass burning, fuel type; and 2) develop a classifier that identifies source types based on emission patterns or measured mixing ratios. The PR algorithm can also be used to explore linkages between emissions and SOA formation, using parameters such temperature and organic aerosol mass loadings. For objective 1, a four step PR algorithm was developed that includes: 1) data pre-processing, 2) feature selection using an analysis of variance based method, 3) dimensionality reduction using principal components analysis (PCA) and 4) clustering using k-means. In the feature selection step, several hundreds of compounds are evaluated for their potential to separate samples based on sampling conditions (e.g., emission source(s), fuel type(s), temperature). In the PCA and k-means steps, the samples are grouped based on their similarity. To increase efficiency, a pre-processing algorithm was also developed for chromatographic alignment, in which retention times and mass spectra are used to correct for small retention time shifts between samples. Together these algorithms reduce the time required for data processing and can be applied to a wide range of sample types.

This thesis is comprised of six chapters in total. This chapter, Chapter 1, introduces each of the applied projects and the motivation for those projects. Chapter 2 describes in detail the theory, the algorithm components, and any other additional major analysis methods that were used in each of the projects. Chapter 3 focuses on the first application of the PR and chromatographic alignment algorithms for analysis of tailpipe emissions samples. Chapter 4 focuses on the updated and final form of the PR algorithm and its application to laboratory and field smoke samples from several different individual (lab) and mixed (field) fuel types common in Western forests. Chapter 5 focuses on the application of a modified version of the PR algorithm for analysis of smoke plume transects from two recent field campaigns. Finally in Chapter 6 the major findings from all projects and future directions are summarized.

1.2 Aromatics in Tailpipe Emissions Samples

During the summer of 2017 at the Bourns College of Engineering Center for Environmental Research and Technology (CE-CERT), our group collaborated with the Emissions and Fuels group to study eight different blends of gasoline fuels, with varying levels of aromatic and ethanol content, on the emissions from a light-duty vehicle equipped with a gasoline direct injection (GDI) engine [14]. The purpose of the project was to better understand the effects of fuel composition on emissions, and subsequently, on secondary aerosol formation and composition. While the effects of fuel composition on primary vehicle emissions have been well studied [15][16][17], less is known about the effects on secondary aerosol formation and composition; only one study was available at the time of this project[18].

The data set acquired from this study was ideal for the initial development and testing of the PR algorithm. First there was no mixing between the tailpipe emission samples from different fuel blends, as is usually the case in smoke samples from the field where multiple fuel types (different plant and tree types) are burning at the same time. Second the emissions profile of each fuel was much simpler-only 100 to 110 compounds per sample-compared to smoke samples with several hundred compounds present from each fuel type. Therefore this data set offered a reasonable starting point to develop and test the algorithm, prior to application for smoke sample analysis, which was the primary focus of the research presented. The algorithm applied in the analysis of the tailpipe emissions data was an earlier version of the algorithm presented in Chapter 2.

1.3 Monoterpenes in Smoke Samples

Wildfires have increased in frequency, duration and size in the Western United States (U.S.) over the past decades [10][11][12]. These trends are projected to continue, with negative consequences for air quality across the U.S [11][12][19]. Wildfires emit large quantities of particles and gases that serve as air pollutants and their precursors, and can lead to severe air quality conditions over large spatial and long temporal scales. Characterization of the chemical constituents in smoke as a function of combustion conditions, fuel type, and fuel component is an important step towards improving the prediction of air quality effects from fires and evaluating mitigation strategies.

Building on the comprehensive characterization of gaseous non-methane organic compounds (NMOCs), specifically monoterpenes, identified in laboratory and field studies, the supervised PR algorithm from Chapter 3 was further developed and successfully used to identify unique chemical speciation profiles among similar fuel types common in forests in the Western U.S. Using the results from the PR algorithm, a classification model based on linear discriminant analysis was trained to differentiate smoke samples collected from laboratory and field studies based on the contribution(s) of dominant fuel type(s).

1.4 Precursors of Organic Aerosol in Smoke Plumes

Western wildland fires constitute a significant part of biomass burning (BB) in the US along with agricultural and prescribed burning. Overall, BB is globally the largest source of combustion-related source of NMOCs [20][21]. NMOCs have a significant influence on the atmosphere, which includes a major contribution to the formation of SOA. Given the complex chemistry and dynamics of SOA formation from fires[22] [23] there is a need for better understanding fire emissions and their contribution to SOA formation. Laboratory studies of controlled burns have shown that NMOC oxidation results in SOA production and an enhancement in OA mass but with increased variability. Many studies have tried to narrow down the variability in observed OA enhancement by focusing on specific SOA precursor groups: 1) monoterpenes, 2) oxygenated aromatics, 3) heterocyclic compounds, and 4) polyaromatic hydrocarbons [24][25][26]. Ahern et. al. [24] found that monoterpenes are important SOA precursors for black spruce and ponderosa pine while furans are important for grasses. Lim et. al. [27] showed that there is a strong correlation between

SOA produced and the initial NMOC mass, with stronger correlations observed for NMOCs more volatile than monoterpenes. These previous studies suggest that there is significant variability among dominant SOA precursors based on fuel type and that there is need to be able to better understand SOA from biomass-burning sources.

In this work the PR algorithm developed in Chapters 3 and 4 was applied to data collected during WE-CAN and FIREX-AQ field campaigns. The goal of this application was to expand the application of the algorithm from linking emissions with sources to linking emissions with processes. Specifically it was an effort to identify the most important SOA precursors for fires sampled in the field. For this application a slightly modified version of the algorithm was used to try to connect changes in compound mixing ratios with extents of OA enhancement.

Chapter 2

Methods

The data analyses presented in this thesis are linked by the application of a pattern recognition (PR) algorithm that was the key development of this research. The parts of the algorithm are described in detail in the following sections: section 2.1 focuses on the data pre-processing, section 2.2 on the feature selection method via analysis of variance (ANOVA), section 2.3 on the dimensionality reduction via principal component analysis, section 2.4 on clustering via k-means and section 2.5 focuses on the entire algorithm flow. Section 2.6 describes a classification model that uses the results of the PR algorithm and was applied in the analysis of the smoke samples (Chapter 4).

2.1 Data pre-processing

Data pre-processing is a broad term used to describe any operation performed on data prior to providing them as input to a model. Depending on the model to be used certain requirements must be met by the data set. Some examples include replacing missing values, checking for normality of the data, and bringing variables of different scales to the same scale. Other examples that are more case specific include correction for background (ambient samples), correction for collected sample volume, or correction for dilution (plume samples). If the data are not properly processed prior to using them in a model they can bias the outcome.

An example of the importance of pre-processing prior to principal component analysis (PCA)[28], is shown in Fig. 2.1 using an example data set from the University of California Irivine machine learning repository [29]. In PCA we are interested in the components that maximize the variance [28][30]. If one component (e.g. wine pH) varies less than another (e.g. alcohol content) because of their respective scales (pH units vs. alcohol %), and if those features are not scaled, PCA might determine that the direction of maximal variance more closely corresponds with the 'pH' axis. As a change in pH of one unit can be considered much more important than the change in alcohol content of one %, this would clearly be an incorrect conclusion.

In this thesis there were two major pre-processing steps that were performed in all of the applications. The first was to handle any missing values and the second was to standardize the data so that all variables were brought to scale. Any additional pre-processing steps that were performed were project specific and will be discussed with the results for each project.

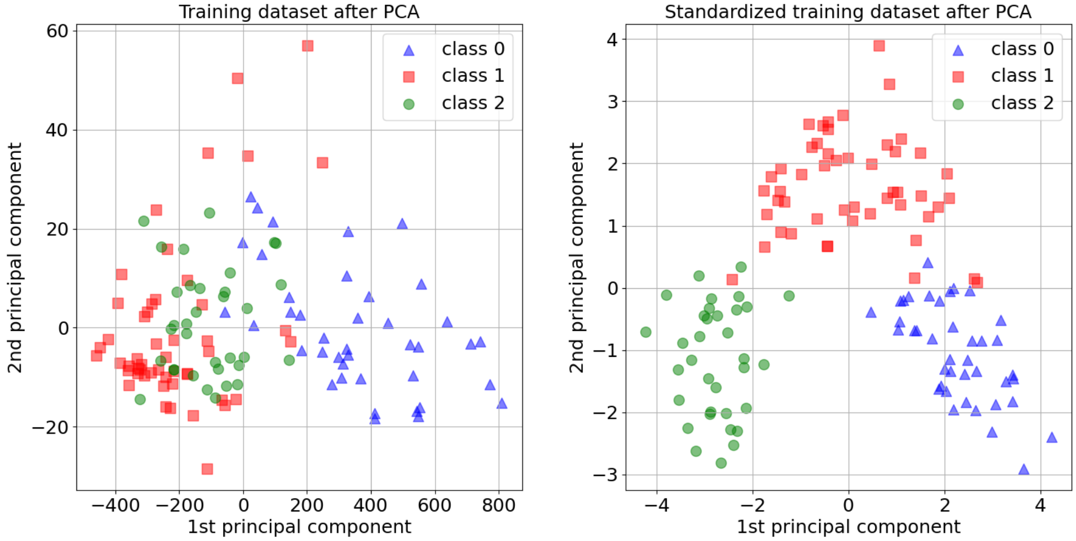


Figure 2.1: PCA results for the demonstrative UC Irvine data set. Left plot shows PCA with no preprocesing (no standardization) and the right plot results after pre-processing.

2.1.1 Missing values

Missing values in experimental data sets pose a relatively difficult problem to solve when analyzing such data. Regardless of the types of data, the missing values can be described under three categories [31]: 1) missing at random (MAR), 2) missing completely at random (MCAR) and 3) missing not completely at random (MNAR). With MAR, the data are not missing across all observations but only within sub-samples of the data. It is not known if the data should be there; instead, data are missing given the data in the observational data set. The missing data can be predicted based on the complete observational data set. With MCAR, the data are missing across all observations regardless of the expected value or other variables. Data may be missing due to test design, failure in the observations or failure in recording observations. These missing data are labeled as MCAR because the reasons for the absence are external and not related to the value of the observation. The MNAR category applies when the missing data have structure. In other words, there appear to be reasons the data are missing. For example, certain classes of compounds might be missing entirely from an aged smoke sample due to chemical losses of those compounds during transport.

For the research presented in this dissertation, for every case in which there were missing values, the same rule was applied to handle them. The rule states that if the missing values are more than 30% by number among the samples then that particular variable should be omitted from the analysis [31]. If the missing values are less than 30% by number then they are imputed. The omission of a variable past the 30% threshold is preferable because imputation in that case lacks natural variation that can help to build an effective model. Regarding imputation there are different methods depending on the type of data. Some examples include regression methods for time series data, mean or median imputation, stochastic imputation, and multiple imputation [32] [33] [34]. In all cases the main disadvantage is reduced variability, depending on the size of the data set, that can bias the model used each time [34]. In all of the applications presented in this thesis the missing values were replaced by zeros unless stated otherwise. Even though the zero filling is not a usual treatment for missing values, it was used because the variables (chemical species) were determined to be truly zero after correction for background.

2.1.2 Data scaling

There are different methods for data scaling [35][36][37] such as min max normalization, quantile scaling, standardization, and power and log transformations. Even though all transformations bring the data to the same scale not all methods are appropriate for every data set. For example log transformations cannot be used with data sets that contain negative values. Some methods such as min max normalization are more sensitive to outliers than others, which are more robust such as quantile scaling and standardization. In this work, standardization was used as the scaling method unless otherwise stated. Mathematically, scaling can be done by subtracting the mean and dividing by the standard deviation for each value of each variable (Eq. 1)

$$z = \frac{v_i - \mu_i}{\sigma_i} \qquad (1)$$

where z is the standardized value or else known as z-score, v is the value of the variable that is being standardized, μ and σ are the mean and standard deviation of the variable to be standardized.

2.2 Feature selection

When building a machine learning model, whether it is as simple as linear regression or more complicated such as neural networks it is almost rare that all the variables in the data set are useful and/or informative. Retaining redundant variables can reduce the generalization capability of a model by over-fitting or increasing the overall complexity and therefore computational resources needed to deploy it. Thus, feature selection becomes an indispensable part of building a robust model. The techniques used for feature selection can be broadly classified into the following categories [38][39]: 1) filter methods, 2) wrapper methods, and 3) embedded methods. Filter methods pick up intrinsic properties of the features among the different classes of the problem by using univariate statistics such as analysis of variance (ANOVA) and the chi-square test. Then using a threshold value of the statistic used, a number of features is selected for the model. Filter methods are generally fast and computationally cheap. Wrapper methods require a search of all possible features to find the subset that best minimizes the error of a specific classification or regression model. They usually are more computationally expensive and sometimes not implementable due to the computational cost. Finally the embedded methods encompass the benefits of both the wrapper and filter methods, by including interactions of features but also maintaining reasonable computational cost.

To demonstrate the importance of feature selection a synthetic data set was created with 1000 observations, 4 classes, and 90 variables from which only 20 were informative and the rest were redundant or noise. Figure 2.2 [a]-[c] shows the separation of the four classes in the PCA space with increasing number of informative features selected using ANOVA (section 2.2.1). The results show that increasing the number of informative features the class separation becomes better.

2.2.1 Analysis of variance based feature selection

In this work a filter method was used, specifically one way analysis of variance (ANOVA) [40]. One way ANOVA is a statistical test used to analyze the difference between the means of more than two groups for variable/s of interest. The null hypothesis (H_o) of ANOVA is that there is no difference among group means. The alternate hypothesis (H_a) is that at least one group differs significantly from the overall mean of the dependent variable. ANOVA uses the F-test for statistical significance. This allows for comparison of multiple means at once, because the error is calculated for the whole set of comparisons rather than

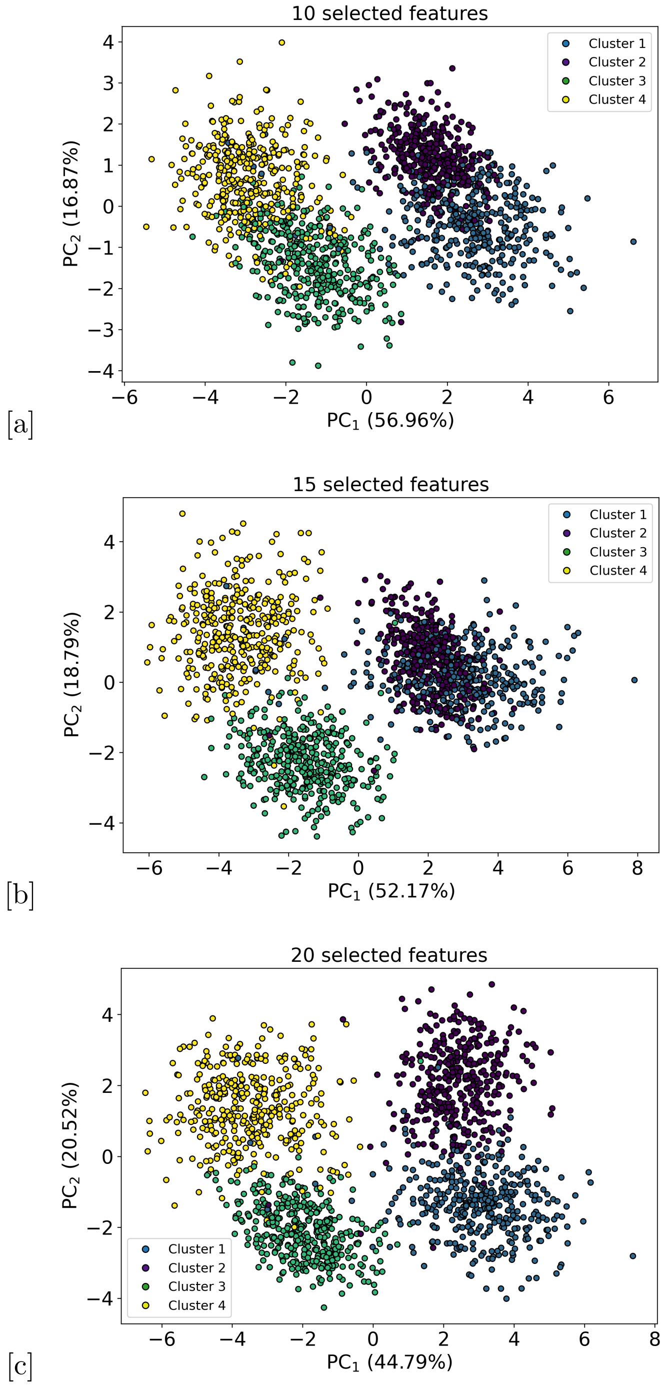


Figure 2.2: Plots a-c show the PCA results for a synthetic data set with a varying number of features selected using the ANOVA based method (see Chapter 2.5).

for each individual two-way comparison (which would happen with a t-test). The F-test compares the variance in each group mean from the overall group variance. If the variance within groups is smaller than the variance between groups, the F-test will find a higher F-statistic (Eq. 2), and therefore a higher likelihood that the difference observed is real and not due to chance. Using an F-value threshold the variables of interest can be selected to be used as input for the model. The exact method for variable selection with ANOVA is described in section 2.5.

$$F - \text{statistic} = \frac{\mathrm{MS_B}}{\mathrm{MS_W}} \quad (2)$$

where MS_B is the mean sum of squares between groups/classes and MS_W is the sum of squares within the same class/group. More details on ANOVA can be found in [40].

2.3 Dimensionality reduction

Dimensionality reduction or low rank approximations are a group of matrix factorization techniques that aim to create a lower dimensional representation of the original data cloud. Conceptually, for n observations/samples that exist in a p-dimensional space, not all of the dimensions are equally important or informative. Depending on the nature of data there are different dimensionality reduction methods such as non-negative matrix factorization (NNMF) [41], isometric mapping (ISOMAP) [42], and principal component analysis (PCA) [28]; PCA was used in this work. PCA seeks a small number of new dimensions that preserve as much variability as possible of the initial data when projected to the lower dimensional space. The new dimensions found by PCA are called principal components (PCs) and are linear combinations of the p original features.

2.3.1 Principal component analysis steps

The first step of in PCA involves the decomposition of the standardized data matrix. For the decomposition there are a few different methods available such as nonlinear iterative partial least squares (NIPALS) [43], eigen decomposition [44], and singular value decomposition (SVD)[44]. Like with the different scaling methods, each decomposition method has certain advantages and disadvantages. For example, NIPALS can be faster than SVD [45]; NIPALS also can accommodate for missing values[45] while SVD cannot. In this work, SVD was used (Eq. 3-4) because missing values were handled prior to PCA in the pre-processing step. Also the data sizes did not pose any numerical problems requiring faster alternatives to be explored. The matrix notation for SVD and the calculation of the PCs are given below.

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}} \quad (3)$$

X is the original (n x d) standardized data matrix, **U** (n x r) is the matrix of left singular vectors, Σ (r x r) is a diagonal matrix of singular values for **X**, and **V** (r x d) is the matrix of right singular vectors that correspond to the new principal directions. r is the rank of matrix **X**. The principal component scores which are the projection of **X** to a lower dimensional space are given by:

$$\text{scores(PCs)} = \boldsymbol{X}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{U}\boldsymbol{\Sigma} \quad (4)$$

As mentioned in the beginning of this section the PCs are a linear combination of the original variables. Equation 4 in a reduced format gives:

$$\mathrm{PC_i} = \mathrm{X} * \mathrm{u_i} = ((\mathrm{x_i}-\mu)\mathrm{u_i}, \ldots, (\mathrm{x_n}-\mu)^{\mathrm{T}}\mathrm{u_i})^{\mathrm{T}} \quad (5)$$

were x_i, ... x_n are the original features. Finally the reduced dimensionality is achieved by using an appropriate number of PCs (m ¡ d) to represent the projected data. The selection of components is based on the minimization of the reconstruction error (Eq. 6)

$$\text{error} = \min|\boldsymbol{X}_{\text{original}} - \boldsymbol{X}_{\text{reconstructed}}| \quad (6)$$

The error shows how much information was lost during the dimensionality reduction using fewer components than the number of dimensions to reconstruct the original data matrix. Because the calculation of the error might be computationally expensive, especially for large matrices, there is another metric which is related to the reconstruction error and is much easier to calculate. That metric is called the explained variance ratio [46] and is equal to:

$$\text{EV\%} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{k}} \lambda_{\mathrm{i}}}{\sum_{\mathrm{i}=1}^{\mathrm{p}} \lambda_{\mathrm{i}}} * 100 \quad (7)$$

k is the number of retained components, p is the number of original variables, λ are the eignevalues of the decomposition and can be obtained from:

$$\boldsymbol{\Lambda} = \frac{\boldsymbol{\Sigma}^2}{(n\text{–}1)} \quad (8)$$

or in reduced format

$$\lambda = \frac{\sigma_i^2}{(n\text{–}1)} \quad (9)$$

where σ are the singular values of the SVD decomposition and n the number of samples in the data. As a rule of thumb a 75% - 80% explained variance ratio and over is considered a good threshold for minimal loss of information.

2.4 Clustering

Clustering or cluster analysis is the task of grouping a set of objects/samples in a way that the objects/samples that belong to a group (called a cluster) are more similar than the objects/samples tat belong to another group (cluster). Clustering is not one specific algorithm but rather a general problem to be solved. There are different algorithms that can be applied to solve clustering problems. Clusters include groups with small distances (e.g. Euclidean) between cluster members, dense areas of the data space or intervals. The appropriate clustering algorithm and parameter settings (including parameters such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but rather an iterative process of optimization that involves trial and error. Often it is necessary to modify data pre-processing and model parameters until the result achieves the desired properties.

2.4.1 *k*-means clustering

In this work whenever clustering was used it was the k-means algorithm [47]. k-means clustering is a method that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster. k-means clustering minimizes within-cluster variances (squared Euclidean distances) [47]. Mathematically the k-means algorithm is formulated as following. Given a set of observations (x_1, x_2 , x_3, ... x_n) where each observation is an d-dimensional real vector, k-means tries to partition the n observations into $k \leq$ n sets S = $\{S_1, S_2\ , S_3\ , ..., S_k\}$. The objective is to find:

$$\underset{S}{\arg\min} \sum_{i=1}^{k} \sum_{x \in S} ||x - \mu_i||^2 \quad (10)$$

where μ_i is the mean of points in S_i

Finally, in order to determine the optimal number of clusters, the elbow plot method was used. The elbow plot is a graphical method in which the cumulative distance is calculated for all points/samples from their respective centroids and then plotted against the number of clusters. In this study Euclidean distance was used (Eq. 3):

$$d = \sum_{k=1}^{k_{max}} \sum_{j=1}^{n} \sqrt{(q_j - c_k)^2} \quad (11)$$

where q and c are multidimensional vectors with the coordinates for each centroid and sample. The indices j and k correspond to centroid number and sample number. The optimum number of clusters is found after a steep decrease in the total Euclidean distance, followed by trivial changes, with increasing number of clusters.

2.5 Pattern recognition algorithm

For the purposes of this work a four-step PR algorithm (Fig. 1) was developed to select a subset of variables/compounds that captured the variance between the fuel types and then use those marker compounds to differentiate fuel types based on the selected variables/compounds; the algorithm steps are: 1) data pre-processing, 2) feature selection, 3) principal component analysis, 4) and k-means clustering. The algorithm was implemented using the Python package scikit-learn [48]. The same algorithm, with small modifications, was used in all applications presented in this thesis. The algorithm steps are explained in more detail in the following paragraph.

In step one, all features are evaluated using the missing values threshold criteria as described in section 2.1.1. After handing any missing values, the second step is the feature selection procedure. Feature selection takes place as follows: 1) the samples are separated to their respective classes and for every variable/compound, ANOVA is performed (section 2.2.1) until the calculation of the F-statistic, also known as the Fisher ratio (F-ratio). No further inference takes place. Following the F-ratio calculation the compounds are ranked in ascending order based on their F-ratio values. Then, in an iterative fashion, PCA (section 2.3.1) is performed and the explained variance from the first two PCs as well as the cluster separation (k-means) are evaluated as a function of the number of compounds (with the highest F-ratios) retained. Provided that the separation in the PCA space is satisfactory the process is terminated and the compounds that provide maximum separation between the classes have been found.

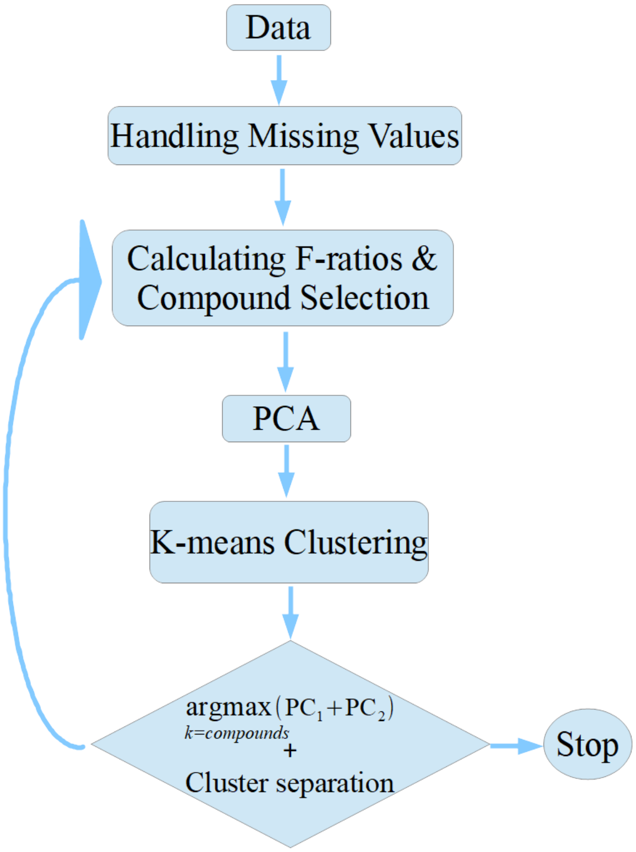


Figure 2.3: Pattern recognition algorithm flowchart

2.6 Linear discriminant analysis

In addition to the PR algorithm, the purpose of which was to find features that can help differentiate smoke samples, a classification model was also trained and used. The purpose of the classification model was to test the applicability of the PR algorithm results for differentiating smoke samples from the field (Chapter 3). The classification algorithm used was linear discriminant analysis (LDA) [38]. LDA is a supervised learning method that is similar to PCA. Both LDA and PCA are linear transformation techniques: LDA is supervised, whereas PCA is unsupervised and ignores class labels. While PCA tries to find a subspace of features in order to maximize variance among samples, LDA attempts to find a feature subspace that maximizes class separability. The output of LDA is a probability score for every sample that is being tested for its likelihood to belong to a particular fuel type (Eq. 12).

$$\log P(y=k|x) = -\frac{1}{2}(x-\mu_k)^t \sum{}^{-1}(x-\mu_k) + logP(y=k) + C_{st} \quad (12)$$

where k is the class of sample x, μ is the vector of the means for each class based on the selected features and Σ is the common covariance matrix for the three classes in the training set. Figure 2.4 shows an example of LDA applied to synthetic data where the decision boundary (white line) and probability space (colour intensity of blue and red) are visualized.

2.7 Chromatographic alignment

Chromatographic alignment involves correcting retention time shifts for compounds between samples that occurred during the instrumental analysis. During instrumental analysis, even if the methods (e.g., columns, temperature ramp, modulation time) are the same, the same compounds across samples might have slightly different (to the second or third decimal digit) retention times. Even though such a difference is minuscule compared to the modulation time (5 seconds for the GCxGC data used in these applications) the result can be that a compound that is in fact present in all samples may not be identified as they same compound across all samples due to these retention time shifts. One way to resolve this

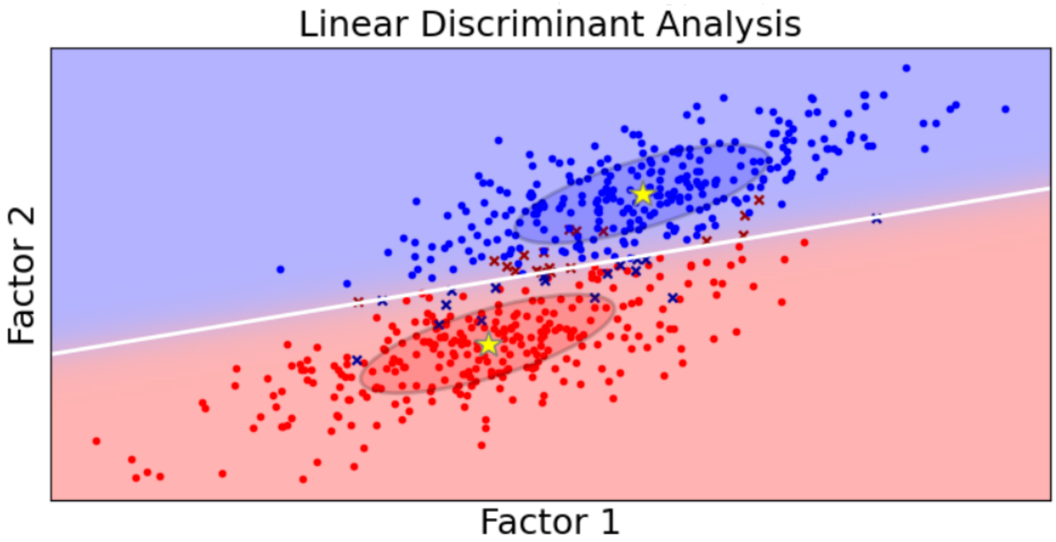


Figure 2.4: The plot shows the decision boundary (white line) for linear discriminant analysis between two classes (red and blue). The rest of the LDA space is coloured based on the probability of a sample to belong to one of the two classes. Class centroids are pointed with yellow stars.

issue is to set a very narrow retention time window for compounds. With complex mixtures, such as smoke, even narrowing the retention time window may not be enough to guarantee that peaks within that window are or are not the same challenge, a chromatographic alignment algorithm (Fig. 2.5) was created that combines the retention time information for every compound across samples along with the relative mass spectra information. The algorithm is described in detail in the following paragraphs.

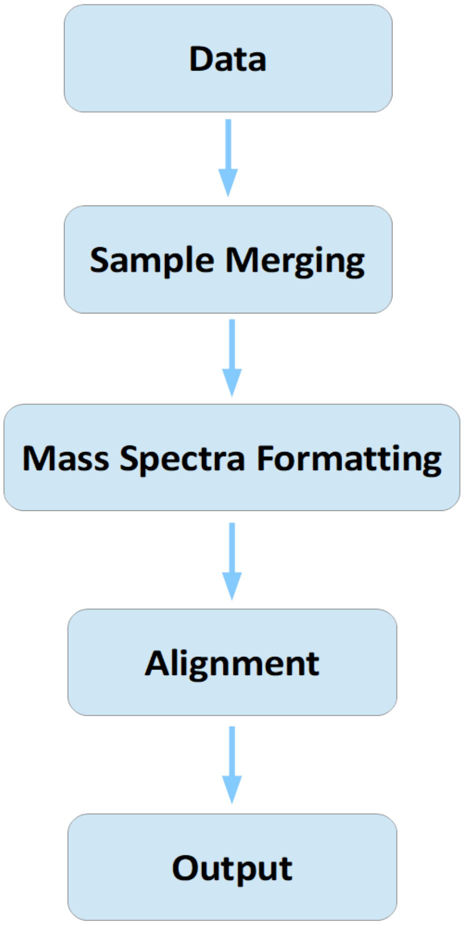


Figure 2.5: Chromatographic alignment algorithm flow chart.

2.7.1 Data files

The data files that are used as an input to the algorithm need to have the following format. Each file's name consists of four fields separated by underscores (e.g [location]_[tubeNo]_[sampleType]_[time_info].csv). The first field is used for the location that the sample was taken, the second for the tube number, the third for the type of sample (e.g. background, blank, etc) and the last one for the time that the sample was taken. The information from the file name is stored as metadata and used later by the script in the output. The file naming convention has not been finalized at this point and could change in future versions of the algorithm. Tabs are used to delimit data fields within records (lines) of data in a file. Each file should contain the following data fields in the order showed here. Name, Group, 1st Dimension Time (s), 2nd Dimension Time (s), Area, Formula, Exact Mass, UniqueMass, Quant Masses, Quant S/N, BaselineModified, Comment, Spectra (relative Abundances). The naming of the data fields comes directly from Chromatof were they are exported.

2.7.2 Sample merging

In order to avoid opening and handling multiple files at a time, which is error prone, the first step of the algorithm involves combining all the sample files into a single file. This is done in an iterative fashion by stacking each file. In order for the algorithm to be able to separate the files later, during the merging each file is being assigned with a unique number ranging from 1 to k where k is the number of samples being merged. The order of the files being merged and has no influence on the alignment.

2.7.3 Mass spectra formatting

Because Chromatof exports the mass spectra of a compound in a format that is not useful for the algorithm, there is one additional step before the alignment begins, and that is to reformat the spectral information. Chromatof provides the spectra in a string format in the form of m/z values ranging from 0 to 550. Instead the use of a list format is more appropriate and efficient in storing that information. The list uses the format given in lists in Python where the values represent the relative abundances for every m/z ordered from 0 to 550.

2.7.4 Alignment

After the spectral reformatting the alignment process can begin. The process involves two major steps: 1) initial alignment and 2) relaxed alignment. The difference between the first the and second is the threshold value being used for whether or not a compound should be matched. Both steps will be discussed in detail in the following paragraphs.

After all the samples have been merged into one file the algorithm starts by selecting the very first compound from the stack. Using the retention time information from that compound the algorithm then creates a retention time window (RT1 ± k, RT2 ± m, k and m are provided by the user) and searches for all compounds that are within that window. When the initial selection of compounds is finished a new search is initiated within the selected compounds list to determine the reference compound. The reference compound is selected by comparing the peak area of each selected compound to each other in order to find the compound with the median peak area between the maximum and the minimum among the selected ones. The reasoning behind this selection is to avoid problematic compounds (located at high and low peak area values) that Chromatof might not deconvolute correctly and thus produce inaccurate mass spectra.

After the reference compound has been selected the rest of the compounds from the initial selection are compared using the dot product method[49]. Briefly each compound's spectrum is considered a vector with the m/z values corresponding to the vector coordinates. Using formulas (13), (14) and (15) the dot product (cos) and then the angle is calculated for every pair of compounds (reference compound and a compound from the list).

$$r_a = \sqrt{\sum_i {a_i}^2} \quad (13)$$

$$r_b = \sqrt{\sum_i {b_i}^2} \quad (14)$$

$$cos\theta = \frac{a_i b_i}{\sqrt{r_a r_b}} \quad (15)$$

where a and b are vectors that hold the mass spectrum values ranging from m/z 0 to 550.

The angle of a pair is being compared to the threshold value and if it is smaller or equal then the compound is retained and flagged. If the angle is greater then the compound is discarded. This process continues until all the compounds in the list are processed. The retained compounds then have their retention times (first and second dimension) replaced by the most frequently occurring first dimension retention time and the average second dimension retention time from the retained compounds. The reasoning for not averaging the first dimension retention time is that some compounds do not elute in one modulation period so they have higher than usual RT_1s. Averaging these compounds adds error to RT_1 for that compound. By choosing the most frequently occurring value instead that error is minimized. To avoid potential similarities in the averaged retention time between different compounds the second retention time average is combined with a randomly generated number (average value + random number) on the order of 10^{-3} using the Random package from Python. After the retention times have been replaced the names of the selected compounds are evaluated for their frequency of occurrence. The name that appears most frequent is selected as the name of the matched compounds. The rest of the properties for the compounds (Formula, Exact Mass, UniqueMass, Quant Masses, Quant S/N, BaselineModified) are taken from the reference compound. After the alignment, compounds that have been aligned are flagged and returned back to the original list. The procedure is repeated until all compounds have been processed. Any compound that is unique among all samples is left as is.

For the relaxed alignment the same procedure is used but the threshold value is larger, thus the name relaxed, to accommodate for potential issues (e.g. bad deconvolution) that might have distorted the mass spectra and caused a compound to get mismatched. The process of alignment is the largely the same, with the only difference being the mass spectra used to match compounds. In the relaxed alignment there is no reference compound being used. Instead the aligned compounds are being grouped and their mass spectra are averaged to produce a more representative spectrum for the group. Then the compounds that are within the retention time windows are being compared against the averaged mass spectrum. Groups of aligned compounds that contain less than three compounds are skipped because the average of only two compounds does not give a representative spectrum. After the relaxed alignment the aligned compounds share the same Formula, Exact Mass, UniqueMass, Quant Masses, Quant S/N, BaselineModified and average mass spectrum.

During the alignment process the script also stores and prints in the final data file comments that the researcher made for a specific group of aligned compounds. The script also finds the most prominent Quant mass used in every aligned group along with the deviations from it made manually by the user during initial analysis of the samples.

Chapter 3

Pattern Recognition Analysis of Tailpipe Emissions from Gasoline Fuels Using an ANOVA Based Feature Selection, PCA and *k*-means Clustering

3.1 Introduction

While this thesis focuses predominantly on the development of chemometric tools for analysis of smoke samples the initial development and testing of the algorithms was performed on tailpipe emissions samples. During the summer of 2017 at the Center for Environmental Research and Technology (CE-CERT), our group collaborated with the Emissions and Fuels group to study eight different blends of gasoline fuels, with varying levels of aromatic and ethanol content, on a light-duty vehicle equipped with gasoline direct injection (GDI) engine [14]. The purpose of the project was to better understand the effects of fuel composition on emissions, and on secondary aerosol formation and composition. While the effects of fuel composition on primary vehicle emissions have been well studied [15][16][17], less is known about the effects on secondary aerosol formation and composition with only one study available at the time of this project[18].

This chapter focuses exclusively on the use of the eight fuel study data to develop the pattern recognition (PR) algorithm. While the PR algorithm in its final form is fully described in Chapter 2, section 2.5, here it is presented in its initial, less refined, form. The purpose of this application was to test whether or not the algorithm can separate the different fuel types by selecting an appropriate number of NMOCs present in the tailpipe emissions of each fuel.

The data set acquired from this study was ideal to start the development and testing of the PR algorithm. First there was no mixing between the emission samples as it is usually the case on smoke samples from field campaigns where multiple fuel types (different plant and tree types) being burned at the same time. Second the emissions profile of each fuel was much simpler with only 100 to 110 compounds per sample compared to several hundred present in smoke. Therefore if the algorithm could pass the benchmark with samples much less complex than smoke samples it would have the potential to be effective when deployed for biomass burning data analysis.

3.2 Data

This study utilized a 2017 model year vehicle that was certified to meet the Federal Tier 3 or the California low-emission vehicle (LEV) III, super ultra-low emission vehicle (SULEV) emission standards. The test vehicle was equipped with a turbocharged 1.5 L centrally-mounted direct injection engine and a three-way catalyst (TWC). Testing was performed over duplicate LA92 cycles on a chassis dynamometer. The LA92 cycle, also known as the California Unified Cycle (UC) is dynamometer driving schedule for light-duty vehicles developed by the California Air Resources Board (CARB). It has a similar three-bag structure to the certification Federal Test Procedure (FTP) cycle, but is a more aggressive cycle with higher speed and acceleration, and less idle time than the FTP. All tests were cold-starts. For the tailpipe emissions of each fuel two replicates were taken and tested in a random order. The emissions from the tailpipe were first diluted using a 10:1 ratio. After the dilution the emissions line split in two. One line was injected into a mobile atmospheric chamber (max volume 29.63 m^3) with additional dilution of 85:1 and the second line was directed though a constant volume sampler (CVS). Samples for chromatographic analysis were collected, using two phase coated stainless steel sorbent tubes, from the CVS and the chamber along with backgrounds (empty chamber and CVS with no tailpipe emissions). The sampling procedure yielded 32 samples and a total of 110-120 identified compounds per sample after the chromatographic analysis. All samples were analyzed using an automated thermal desorption unit coupled to a two dimensional gas chromatograph with a time-of-flight mass spectrometer (GC × GC-TOFMS). The raw chromatograms were processed using the commercially available software ChromaTOF.

The fuels, named as F1 through F8, were designed to meet nominal total aromatics of 30% and 20% (v)̌ and ethanol levels varying between 0% to 20% (v)̌ (Table 3.1). Also additional name identifiers were given based on their ethanol and aromatic content (e.g. F1(E0, 20), F2(E0, 30), etc.). More details about the fuels in [16]. Briefly, F5 (E15, 20) and F8 (E20, 19) were created by splash blending denatured ethanol with F3 (E10,20), which was Tier 3 E10 certification fuel. The other five were match blended to meet high and low aromatic targets with varying ethanol levels. The fuels were also blended to represent lower PM indices [50](fuels F1, F3, F5, F6, and F8 with PMI values ranging from 1.613 to 1.888) and higher PM indices (fuels F2, F4, and F7 with PMI values 2.093 to 2.330). Splash blending involves dilution of gasoline fuel with a specified volume of ethanol. Gasoline in the US is predominantly produced through splash blending. Match blending is when the gasoline blend stock is adjusted to meet specific one or more fuel properties.

Table 3.1: Raw Fuel properties and total aromatic peak area % in emissions.

Fuel Name	Aro. (vol.) %	Eth. (vol.) %	Emissions: Aro. Peak Area (%)
F1	21.2	0	31
F2	29.4	0	28.8
F3	21.4	10	29.6
F4	29.1	15	40.5
F5	20.3	15	30.2
F6	21.8	15	42.2
F7	29.3	15	35.8
F8	19.1	20	27.2

3.3 Algorithm Implementation, Results and Discussion

3.3.1 Data pre-processing

In this project the pre-processing was slightly altered from the usual procedure (Chapter 2, section 2.1). First, all of the samples were chromatographically aligned as discussed in Chapter 2, section 2.7. After alignment, for all compounds that were identified, the true peak area in the chromatogram was calculated by subtracting the peak areas from the background samples. Because several backgrounds were taken during the project, for each compound the highest value among the different background samples was used each time. Following the background calculation the peak areas for all compounds were normalized to the total peak area of each chromatogram. This was done in order to correct for small differences in the collected volume for each sample and to also bring all compound peak areas from the different samples to the same scale. The last step of pre-processing entailed handling missing values for compounds among samples. In this case the missing values were handled as described in Chapter 2, section 2.1.1.

3.3.2 Pattern recognition results

In this project two different feature selection methods were tested. Method one was to filter compounds in the samples using a threshold value for the average normalized peak among samples and the second one was to use a statistical approach, specifically an ANOVA based method. For the peak area cutoff values, two cases were tested with threshold values of 1% and 2%. The threshold values are relatively small but the the average peak area % range among the samples was between 0.01% and 16%. The peak area filtering

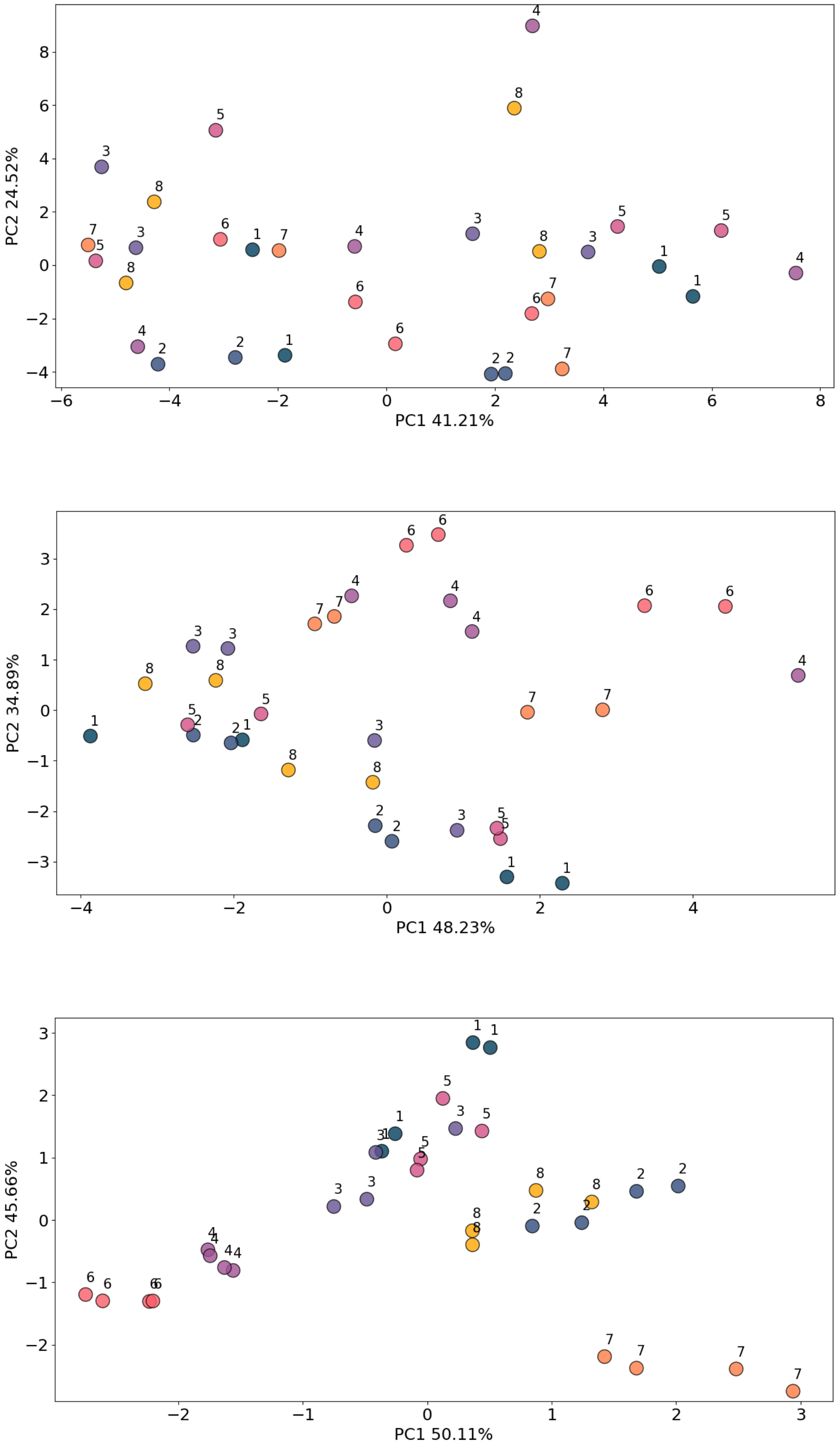
Table 3.2: Selected compounds with peak area threshold and ANOVA.

1% threshold	2% threshold	ANOVA ($F = 20$)
toluene	toluene	3-methyl, hexane
benzene	1,3-dimethyl, benzene	1,3-dimethyl-cis, cylcopentane
C8-paraffin	2,2,3-trimethyl, pentane	C7-olefin
1,3-dimethyl, benzene	1-ethyl, 2-methyl-cis, pentane	heptane
2,3,3-trimethyl, pentane	2,3,4-trimethyl, pentane	methyl, cyclohexane
1-ethyl, 2-methyl-cis, pentane	*n*-hexane	3-methylene, heptane
2,2,3-trimethyl, pentane		2,3,3-trimethyl, pentane
methyl, cyclobutane		2,3-dimethyl, hexane
n-hexane		4-ethyl, 5-methyl, pentane
pentane		Nonane
ethylbenzene		2,6-dimethyl, octane
2,3,4-trimethyl, pentane		1,2,3,4-tetramethyl, benzene
2-methyl, butane		1-ethyl, 2,4-dimethyl, benzene
2,4-dimethyl, hexane		(*E* 1-phenyl, 1-butene)
2,2,4-trimethyl, hexane		2-ethyl, 1,3-dimethyl, benzene
2-methyl pentane		
o-xylene		

yielded 17 and 6 compounds respectively. For the ANOVA based method the compound selection was performed slightly different than the final procedure in Chapter 2, section 2.2.1. Specifically different F values (5, 10, 15, 20, 30, and 35) were used as a threshold and then PCA was peformed and the fuel separation was inspected visually. The compounds selected are presented in Table 3.2. For ANOVA only the F that resulted in the optimal separation is presented with 15 compounds.

Figure 3.1 [a]-[c] shows the PCA results for the compounds selected using the peak threshold and the F statistic. From the plots 3.1 [a] and [b] it is obvious that there is almost no separation between the different fuel types. On the contrary the ANOVA based method with a threshold of $F = 20$ (Fig. 3.1 [c]) provided a much better separation between the classes. The rest of the F values that were tested (not shown here) yielded either the same or worse separation results similar to the peak area filtering method. The ANOVA based method resulted in better separation of the fuels (Fig. 3.2 [b]) and coupled with the k-means grouped the eight fuels in four optimal clusters as shown from the elbow plot (Fig. 3.2 [a]). The fuel clusters are; 1) (F1, F3, F5), 2) (F4, F6), 3) F7 on its own, and 4) (F2, F8).

Because the separation of the samples in the PCA space is connected with selected compounds, their normalized peak area profiles were investigated using the cosine θ measure of similarity (Chapter 2, section 2.7.4). The comparison was performed pairwise for all fuels that were clustered together first and then for all the fuels across different clusters. The results are shown in Fig. 3.3. The cosine similarity for fuels 1,3,5 and 2,8 which are clustered together (Fig. 3.2 [b]) is almost perfect with an average value of 0.98 (Fig. 3.3 left plot). Interestingly the 4,6 fuel cluster has relatively low similarity with a value of 0.77. The right plot in Fig. 3.3 shows the pairwise similarity of fuels across different clusters. As expected the fuels that did not cluster together have lower cosine values ranging from 0.25 to 0.83. The results shown in Figs. 3.2 and 3.3 support that the algorithm successfully selected compounds, that have enough discriminatory power to separate dissimilar fuel types and group together similar ones. While the separation is attributed to all the selected compounds, the aromatics might have played a significant on the clustering and separation of the samples. Table 3.3 shows the clustered fuels and the total aromatic peak are in the emissions. All clustered fuels have very similar total aromatic peak area. From the ANOVA based selection, while only three compounds are aromatics (see Table 3.2) they account for about 5% - 34% of the total peak area among the selected compounds. Therefore the similarity of the emissions based on aromatics could be reflected on the selected compounds.

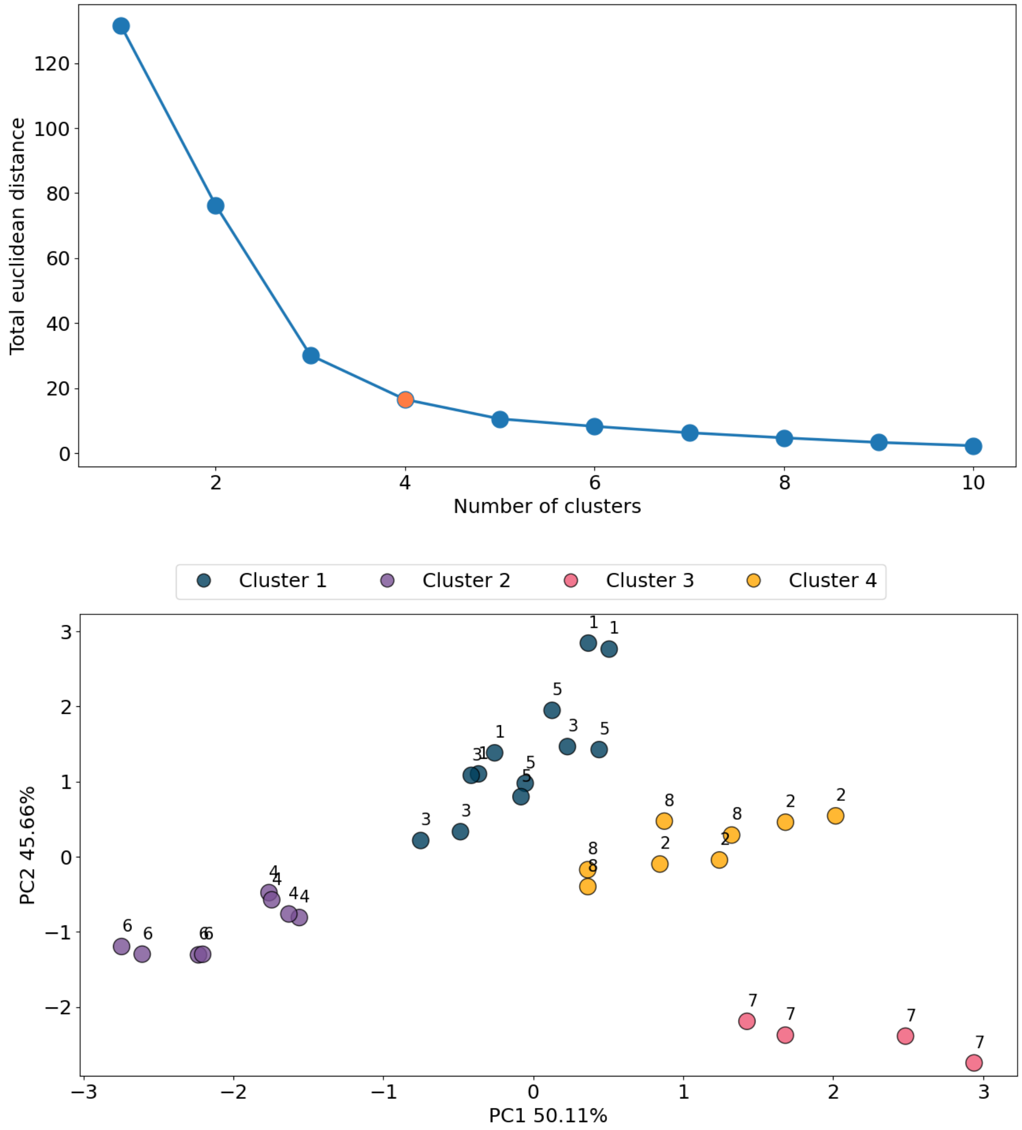


[a]

[b]

[c]

Figure 3.1: Plots [a]-[c] shows the PCA results for peak area threshold 1%, 2% and $F = 20$.



[a]

[b]

Figure 3.2: Plots [a]-[b] shows the PCA and k-means clustering results for $F = 20$. The colour coding corresponds to the different cluster from k-means.

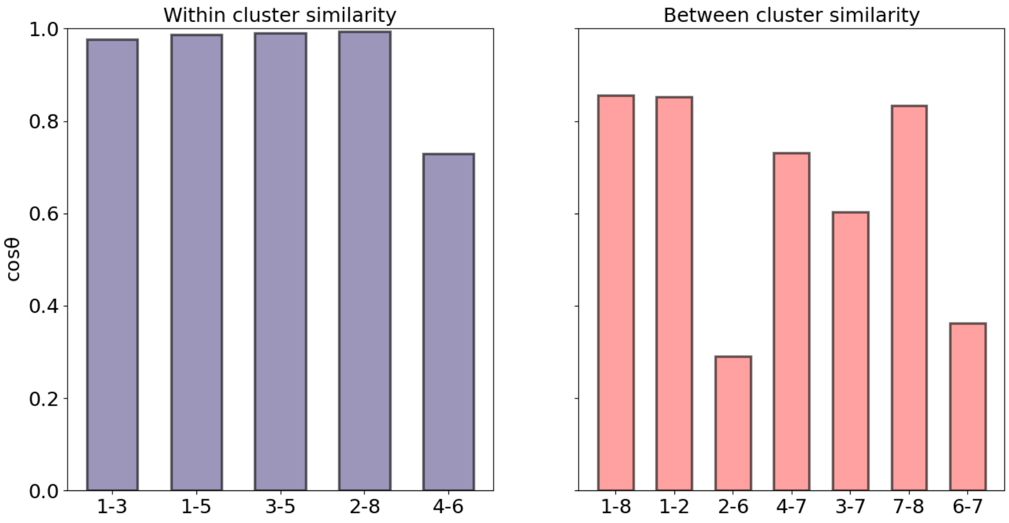


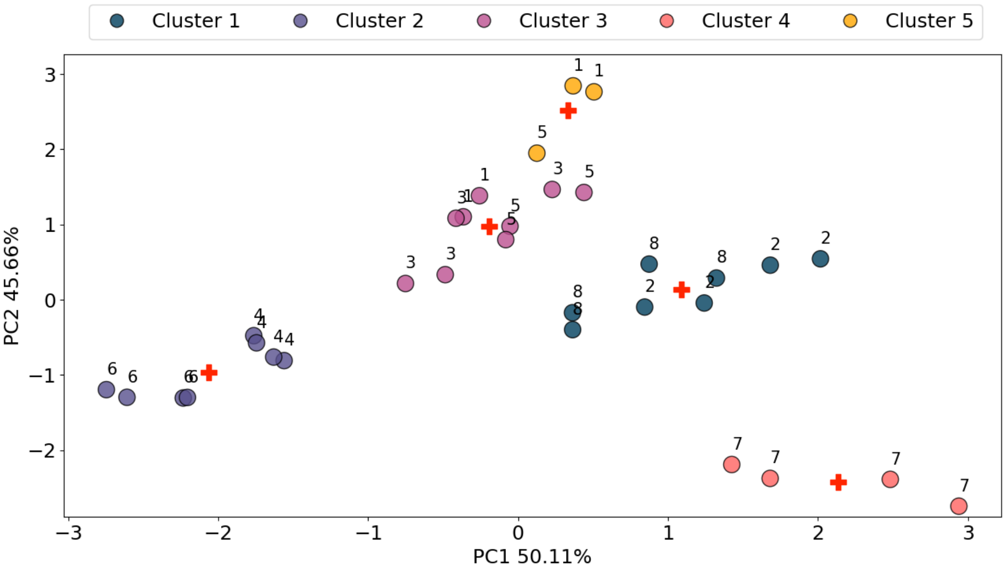
Figure 3.3: Left plots shows cosine similarity for all fuels among the fours clusters except F7 which has its own cluster. Right plot shows cosine similarity for all fuels across different clusters (not all combinations are shows for figure clarity).

While all fuels were clustered based on their tailpipe emissions similarity, the pair (4,6) had a similarity value 0.77. Such level of similarity should have resulted in fuels 4 and 6 to group by themselves (e.g. fuel 7). Instead the clustering algorithm grouped them together. In order to investigate the reason for that grouping, the number of clusters in the PCA space was increased from the optimal solution of four to five and six. While the elbow method provides an optimal number of clusters, sometimes an intuitive selection based on a expected outcome (in this case separation of fuels 6 and 4) is reasonable. Figure 3.4 [a]-[b] shows the clustering results with five and six clusters.

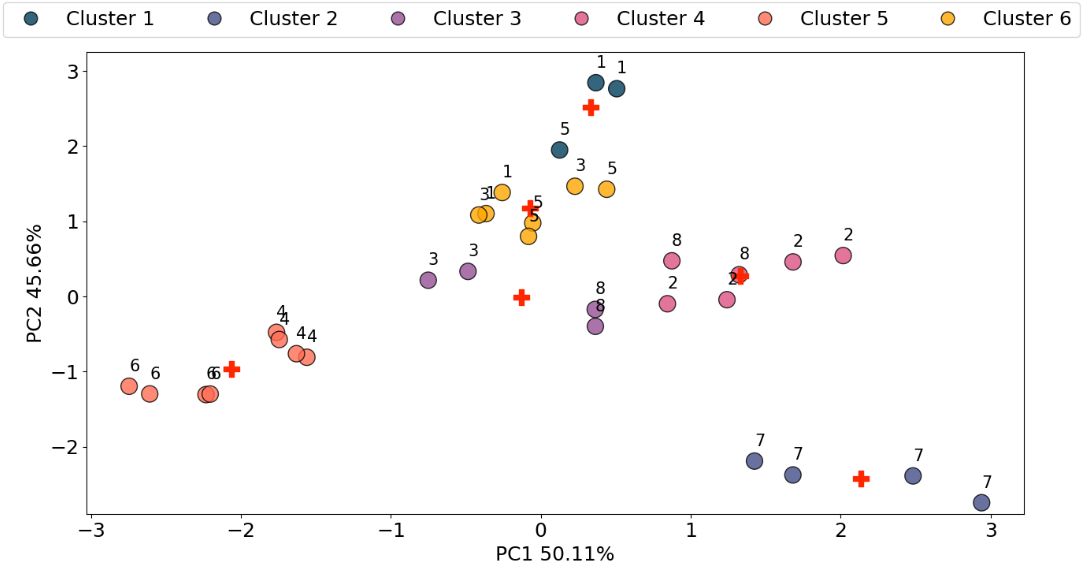
In both cases fuels 6 and 4 are still clustered together. The clustering performance for the rest of the fuels though degraded. Fuels 3 and 8 formed a new cluster (Fig. 3.4 [b]) and some of the samples from the initial cluster of fuels 1,3, and 5 formed another cluster by themselves (Fig. 3.4 [a]-[b]). While increasing the number of clusters did not help further separate fuels 4 and 6 the two tests provided further insight on why they cluster together. As seen in Fig. 3.4 [a]-[b], the optimal placement of cluster centroids (red crosses), in order to minimize the objective function for the k-means algorithm (Chapter 2, section 2.4.1), does not allow for further separation of the two fuels. This result could be due to a convergence of the k-means algorithm to a local minimum. To test that possibility the number of iterations was increased from 300 to 1000 and the initial random placement of centroids from 10 to 1000. This round of tests resulted in the same exact solution shown in Fig. 3.4 [a]-[b]. This suggests that the PR algorithm in its current form cannot separate fuels 4 and 6 further.

Table 3.3: Clustered fuels and total aromatic peak area %

Fuel Clusters %	Emissions: Aromatic Peak Area (%)
F1	31
F3	29.6
F5	30.2
F4	40.5
F6	42.2
F2	28.8
F8	27.2
F7	35.8



[a]



[b]

Figure 3.4: Plots [a]-[b] shows the PCA and k-means clustering results for $F = 20$. The colour coding corresponds to the different cluster from k-means. The red cross markers are the cluster centroids.

3.4 Summary

In this project a PR algorithm and a chromatographic alignment algorithm were created to process the tailpipe emissions from eight gasoline fuels with varying levels of ethanol and aromatics. The chromatographic alignment algorithm successfully aligned all the detected compounds and reduced the amount of time that would have been required for manual alignment from months to just a few minutes. Finally the PR algorithm was applied with two different approaches for feature selection; 1) peak area filtering 2) ANOVA based selection, with the ANOVA being the approach that successfully separated the eight fuels. The major difference between them was reflected on the selected compounds. While the peak area filtering inherently chose prominent compounds in the tailpipe emissions the ANOVA did not. Instead the ANOVA selected compounds reflected important differences/similarities, specifically the aromatic content, between the tailpipe emissions that were not intuitive based on the fuel composition or the major compounds emitted from each fuel.

While most fuels were separated based on their tailpipe emission similarity, one pair of fuels (4,6) was not. The pair was further investigated and the results showed that the algorithm, in its current format, could not fully separate the two fuels. A future potential improvement would be to replace the k-means component (distance based) of the algorithm with a density based clustering approach, such as DBSCAN.

Chapter 4

Development and Application of a Supervised Pattern Recognition Algorithm for Identification of Source-Specific Emissions Profiles

4.1 Introduction

Research has showed that the Western U.S. has seen an increase in the frequency and intensity of wildfires, over the last three decades ([10], [11] and [12]), which is projected to continue in the coming decades ([11], [12] and [19]). One of the consequences of wildfires is extremely poor air quality ([6], [7], [8], and [9]). Emissions from wildfires include carbon monoxide (CO), carbon dioxide (CO_2), and methane (CH4); several hundreds of gas-phase non-methane organic compounds (NMOCs); and particulate matter (PM). While CO_2 and CH_4 are important greenhouse gases, NMOCs are of particular importance in the context of air quality because they serve as precursors to secondary air pollutants including photochemical ozone (O_3) and secondary organic aerosol (SOA) ([2]). The latter of which, SOA, is a major constituent of atmospheric PM ([3], [4], [5]). In order to predict the air quality impacts of wildfires, differences in emissions and their effects on chemistry and pollutant formation must be represented in models ([10], [51], [52], [53], [54]). Wildfire emissions are dependent on a number of factors such as combustion conditions (e.g., flaming vs. smoldering), fuel conditions (e.g., moisture content), and fuel type (e.g., species and component) ([1], [55], [56], [57], [58], [59], [60], [61], [62]). Differences in these factors can affect the total flux of emissions as well as the profile of emissions, i.e., the identities and quantities of individual constituents. Permar et. al. [63] recently reported that combustion conditions, specifically modified combustion efficiency (MCE), explained approximately 70% of the variability in observed trace gas emissions from wildfires. Consistent with some existing modeling approaches, they suggested total NMOCs could be predicted using MCE, and the contribution of individual compounds determined using speciation profiles. Success of that approach requires knowledge of the relevant speciation profiles, and therefore contributing fuel types.

NMOC speciation profiles have been developed from both field and laboratory studies ([62], [64], [65], [66], [67], [68]). Laboratory studies offer some advantages over field studies in the context of controlling fuel species and fuel components; other variables, such as combustion conditions and fuel moisture, can be harder to control and can lead to differences in the identities and quantities of NMOCs emitted between laboratory and field studies ([58], [57], [61], [69]). Yokelson et. al.[69] presented an inter-comparison of laboratory- and field-based emission factors (EFs), and approaches for using laboratory data to enhance the fundamental understanding of fire emissions coupled with field data to evaluate the representativeness of laboratory-based measurements. At that time, they noted that up to 70% of NMOCs remained unidentified for certain fuel types. More recently, due to the application of advanced instrumental techniques, there has been significant improvements in the identification and quantification of NMOCs emitted from fires, particularly in laboratory studies ([58], [59], [60], [70]). Stockwell et. al.[59] detected approximately 80–96 % of the total emitted NMOC mass in experiments during the 2012 fourth Fire Lab at Missoula Experiment (FLAME-4); and Hatch et. al.[1] identified more than 500 individual NMOCs during FLAME-4. The relatively rapid expansion in available NMOC data provides opportunities for developing more detailed speciation profiles (in which a higher fraction of the detected mass is assigned to unique compounds or formulas) and for applying statistical data analysis methods, including to identify unique sets of compounds that allow differentiation of fuel type(s) and estimation of their contributions to smoke samples.

Existing approaches for identifying the contribution of fuel types to smoke include land cover databases or fuel loading models coupled with fuel consumption models (e.g. FOFEM [71] and CONSUME [72]), and the use of marker compounds. One of the limitations of land cover databases or fuel loading models is that they are difficult to update frequently enough to reflect changes in ecosystems ([73], [74], [75], [76]). Marker compounds are emitted in relatively high abundances and can be used to differentiate fuels by component or fuel layer and in some case by species. For example, Wan et. al. [77] showed that p-hydroxybenzoic acid was emitted from combustion of herbaceous plants, while vanillic acid was emitted from combustion of softwoods and hardwoods. It has also been shown that syringic acid is associated with hardwood combustion ([78] and [79]), and dehydroabietic acid with conifers ([80]). Zhang et. al. [81] found that the benzene to toluene ratio in smoke from sugarcane leaves was different than the ratio in smoke from sesame stalk, demonstrating differences among agricultural fuels. In measurements of Western forests and shrublands, [82] showed that hydroquinone was a good marker for manzanita combustion. One of the limitations of using marker compounds to identify fuel types is the lack of specificity, i.e., marker compounds have not been identified that enable identification of a large number of fuel species or closely related fuel species.

In this work a method is presented for identifying fuel types from measured NMOCs in smoke samples. To overcome some of the existing limitations in identifying the contribution of specific fuel types to smoke, the pattern recognition (PR) algorithm from Chapter 3 was further developed and classification was build using the output of the PR algorithm. Both model were developed using data obtained during two laboratory campaigns in 2012 and 2015, and applied to data obtained during a field study in 2017. Machine learning techniques have been applied for source identification in other disciplines. For example, [83] and [84] used principal component analysis (PCA) and linear discriminant analysis (LDA) to differentiate and classify wine varietals based on specific compounds present in wine samples. [85] used PCA and analysis of variance to select marker compounds in gasoline fuel blends and PR to differentiate the blends. In this work, the data sets generated during FLAME-4 and the Fire Influence on Regional to Global Environments Experiment (FIREX) 2016 Fire Lab campaign were leveraged to develop a source identification method using fuel-specific NMOC profiles. The PR algorithm performs an automated selection of compounds that differentiate sources (in this case, fuels) based on measured NMOCs. The classification model then uses the source profiles to identify source contributions to specific samples.

4.2 Data

The NMOC data used in this study were acquired from a variety of fuel types burned in laboratory and field settings during three campaigns: 1) FLAME-4 laboratory campaign in 2012 (FLAME-4 FL12), 2) FIREX laboratory campaign in 2016 (FIREX FL16), and 3) Blodgett Forest Research Station (BFRS) prescribed burns in 2017; both laboratory campaigns took place at the U.S. Forest Service Fire Science Laboratory (FSL). Details of the facilities, sample collection, and data analysis have been discussed in previous publications ([1], [70], [86]). Briefly, during FLAME-4 FL12 and FIREX FL16 a broad variety of biomass fuels were burned ([58] and [87]), including conifers and shrubs (Table 4.1); 80 samples were collected from both room and stack burns as described in [58] and [87]. During the BFRS study, a total of 28 samples [1] were collected from a utility task vehicle parked downwind from three different prescribed burn plots that had different fuel distributions Fig. 4.12 and 4.13 and Table 1). All NMOC samples were collected using dual bed stainless steel sorbent tubes and were analyzed using an automated thermal desorption unit coupled to a two dimensional gas chromatograph with a time-of-flight mass spectrometer

(GC × GC-TOFMS). The raw chromatograms were processed using the commercially available software Chromatof (Leco Corp., St. Joseph, MI). The measured mixing ratios were used to calculate normalized excess mixing ratios (NEMR) versus CO, $\Delta X/\Delta\mathrm{CO}$ ([88]), in which delta represents excess over background. The calculated NEMRs of monoterpenoids ($C_{10}H_{16}$ and $C_{10}H_{16}O$) were used as the starting point for this analysis based on [89] and [1]. Hatch et. al. [1] demonstrated that the variability in NMOC composition could not be attributed entirely to MCE, and that chemical speciation was highly correlated among some fuel types across a range of MCE values, particularly conifers; within conifers, clear differences in monoterpenoid emissions were observed as a function of fuel species.

Table 4.1: Fuels Burned and Smoke Analyzed From FLAME-4 2012 Laboratory Fires, FIREX 2016 Laboratory Fires, and Blodgett Forest Research Station Prescribed Burns.

Fuel Family	Fuel Type	FLAME-4 FL12 (lab)	FIREX FL16 (lab)	BFRS
Conifers				
	Ponderosa pine	x	x	x
	Lodgepole pine		x	
	Engelmann spruce		x	
	Black spruce	x		
	Douglas fir		x	
	Subalpine fir	x	x	
	White fir			x
	Juniper		x	
	Loblolly pine		x	
	Sugar pine			x
	Jeffrey pine		x	
	Incense cedar			x
Shrubs				
	Chamise		x	
	Manzanita		x	
	Sagebrush		x	
	Snowbrush ceanothus		x	
Miscellaneous				
	California black oak	x		
	Tanoak			x
	Excelsior		x	
	Yak dung		x	
	Peat	x	x	
	Rice straw	x	x	
	Bear grass		x	
	Untreated lumber	x		

4.3 Algorithm Implementation, Results and Discussion

4.3.1 Sample and fuel type selection for PR and classification

The PR algorithm (Chapter 2 section 2.5) was applied to the FIREX FL16 data set to identify a group of marker compounds that could be used to differentiate fuel types. Classification (Chapter 2 section 2.6) was then performed using the FIREX FL16 data as the training set, and BFRS data as the testing set. The selection of the training and testing sets was based on the size of each data set; the FIREX FL16 data set had 74 samples, and the BFRS data set had 29 samples. A larger training set ensured more statistically robust parameters for the LDA algorithm. Because the BFRS data span a wide range of complexity in the fuels sampled, a synthetic data set was generated to test the performance of the PR and classification algorithms on mixed fuel samples prior to application on the BFRS data. Five synthetic mixtures were generated with the following compositions: 60% pine/40% spruce, 60% fir/40% spruce, 60% pine/40% fir, 90% pine/10% spruce, and 90% fir/10% spruce. The FLAME-4 FL12 data were used as an independent data set to test the response of the classification algorithm to fuel types that were not included in the training set. The use of each data set in both algorithms is summarized in Table 4.2.

Table 4.2: Data sets used for developing the PR algorithm and testing and training the classification algorithm.

Data Set	Pattern Recognition	Training Set	Testing Set
FLAME-4 FL12			x
BFRS			x
FIREX FL16	x	x	
Synthetic data	x		x

Except from the standard pre-processing (Chapter 2 section 2.1), two selection criteria were applied to the training set to ensure that standard deviations and averages could be computed, which are central features of the PR algorithm. First, only fuel types that had more than 30% (by number) of the 93 monoterpenoids above the limit of detection (LOD) were selected. Since the PR algorithm was based on monoterpenoids, samples with little to no detected monoterpenoids would reduce the ability of the algorithm to differentiate between fuel types and therefore reduce the overall efficiency. Second, only fuel types that had three or more samples were retained. Application of these criteria reduced the number of samples from a total of 74 to 39 and the number of fuel species from 18 to five: pines (ponderosa pine and lodgepole pine), firs (Douglas fir and subalpine fir) and spruce (Engelmann spruce). During the FIREX FL16 study different fuel components were also burned such as canopy, rotten log, composite, litter and duff. While differences in component emissions may be important for differentiating prescribed burn and wildfire emissions in smoke, for this application,based on the selection criteria, 32 composite and canopy samples were retained along with seven litter and duff samples.

4.3.2 Manual versus automated feature selection

Feature selection was performed and evaluated using two approaches: 1) manual selection; and 2) automated selection based on F-ratios (Eq. 2). The percentage of variance explained (Eq. 7) using the first two principal components (PCs) was used as the metric to evaluate the quality of feature selection using the two approaches (see scree plot method section 3.2.2) following application of PCA. For the automated selection the compounds were selected as described in Chapter 2, section 2.2. For manual selection, the compounds

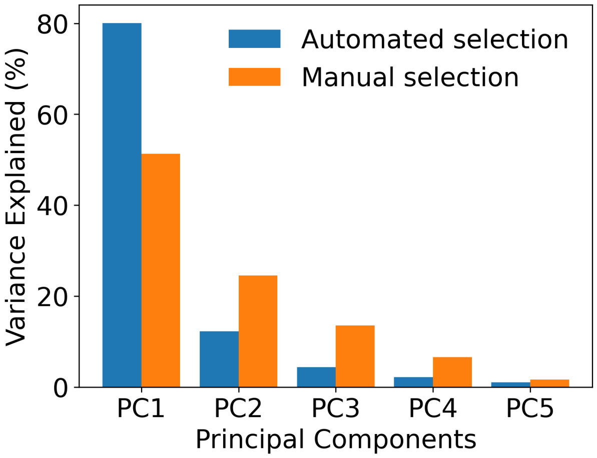


Figure 4.1: Percentage of variance explained per principal component (PC) for the case of automated (blue) and manual (orange) compound selection using Fisher ratios.

were filtered based on a single criterion: whether a compound was present in more than three fuel species.

The application of the manual approach resulted in the selection of the following nine compounds (out of 93): a-pinene, limonene, 3-carene, b-myrcene, camphene, p-cymene, bornyl acetate, b-phellandrene, and tricyclene. Application of the automated approach resulted in selection of the following five compounds: tricyclene, camphene, b-pinene, 3-carene, and bornyl acetate. Figure 4.1 shows the improved performance of automated feature selection over manual feature selection, based on the single highest explained variance across PCs 1-4. To make the feature selection results more intuitive, the normalized emission ratio profiles (ratio of the compound ER to the sum of ERs for the selected compounds) as a function of fuel species are shown for manual selection (Fig. 4.2), automated selection for ten compounds (Fig. 4.3) and five compounds (Fig. 4.4). The automated selection with five compounds results in more distinct and consistent profiles for each fuel

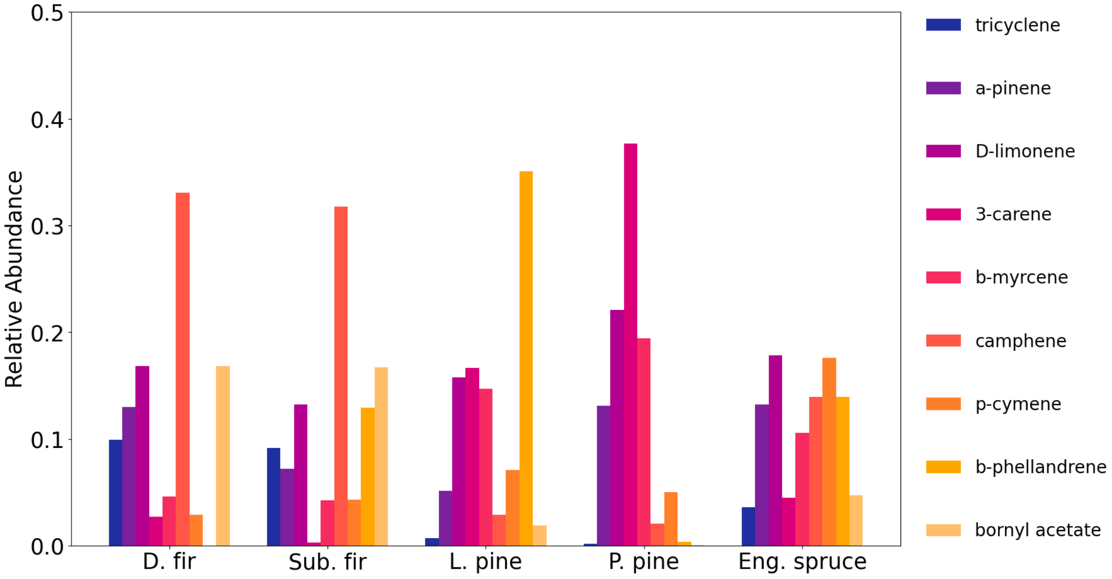


Figure 4.2: Normalized emission ratio profiles for: Douglas fir, subalpine fir, lodgepole pine, ponderosa pine, and Engelmann spruce based on manual selection of compounds.

family, which translates to a higher potential for separation (greater explained variance) in the PCA space. The five compounds selected with the automated approach were thus used for the PR analysis.

4.3.3 PCA and *k*-means clustering

Following data pre-processing and feature selection, PCA was performed on the reduced data set. To determine the number of PCs to be retained, a scree test using a modified version of the Kaiser criterion [46] was performed. Figure 4.1 shows that with automated feature selection two PCs (PC1 and PC2) were adequate to explain 92% of the variance in the data set. The scores from the retained PCs (PC1 and PC2) were then used as input for *k*-means clustering.

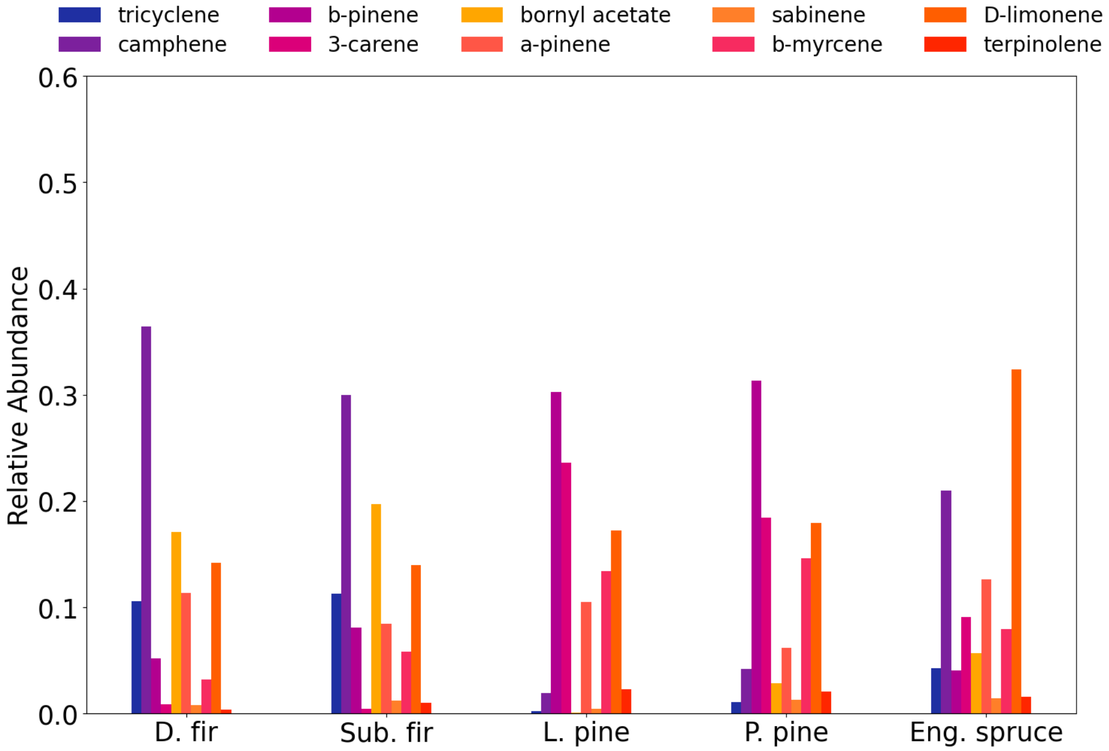


Figure 4.3: Normalized emission ratio profiles for: Douglas fir, subalpine fir, lodgepole pine, ponderosa pine, and Engelmann spruce based on automated selection of 10 compounds.

The coupled PCA and k-means results are shown in Fig. 4.5 [a]-[b]. The algorithm identified four clusters as optimal using the elbow plot method (Fig. 4.5 [a]). The clusters identified by the algorithm are differentiated using marker color while the fuel families are differentiated using marker shape. Cluster one included 15 out of 16 pine samples and one overlapping fir sample. Cluster two included 11 out of 13 fir samples and four overlapping spruce samples. Clusters three and four included the remaining six spruce samples, one overlapping pine sample, and one overlapping fir sample. Generally the algorithm resulted in adequate separation between firs and pines; but poorer separation for spruce, for which four of ten samples overlapped with another fuel family. The difficulty that the algorithm

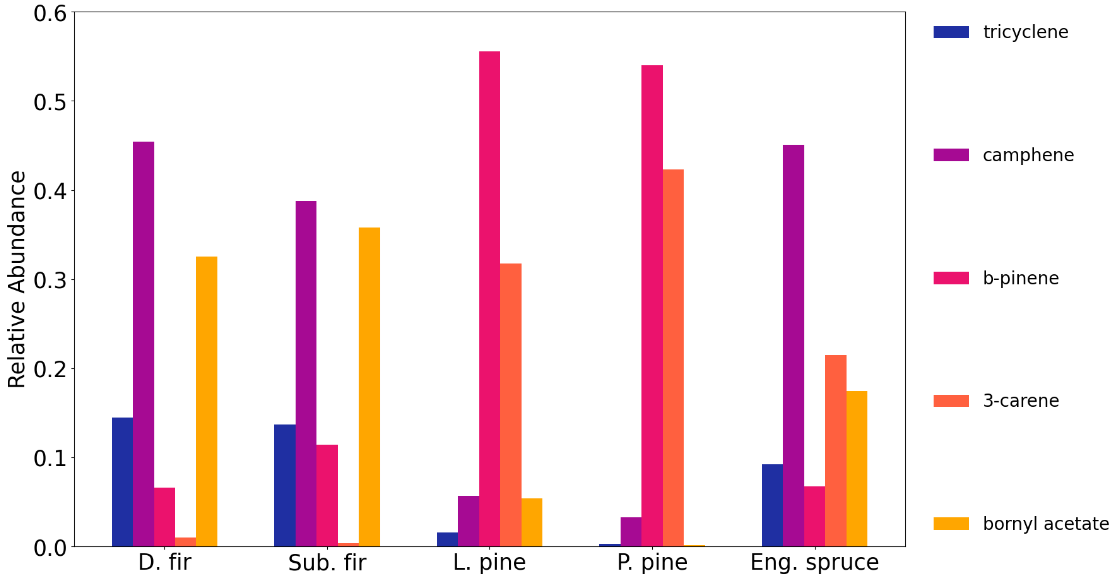


Figure 4.4: Normalized emission ratio profiles for: Douglas fir, subalpine fir, lodgepole pine, ponderosa pine, and Engelmann spruce based on automated selection of 5 compounds.

encounters separating spruce (Fig. 4.5 [b]) effectively can be explained using the elbow plot (Fig. 4.5 [a]). The k-means algorithm identified four clusters as the optimum number, but the steep decrease in the total Euclidean distance actually occurs between one and two total clusters. The Euclidean distance decreases more between two and four but to a lesser extent (smaller slope) compared to the decrease between one and two. The lesser decrease between two and four clusters indicates that the clustering algorithm had difficulty identifying clusters in the PCA space. From (Fig. 4.4), it can be seen that the spruce and fir samples have similar normalized tricyclene, camphene, and b-pinene emission ratios. This limits the ability to fully separate spruce and firs in the PCA space.

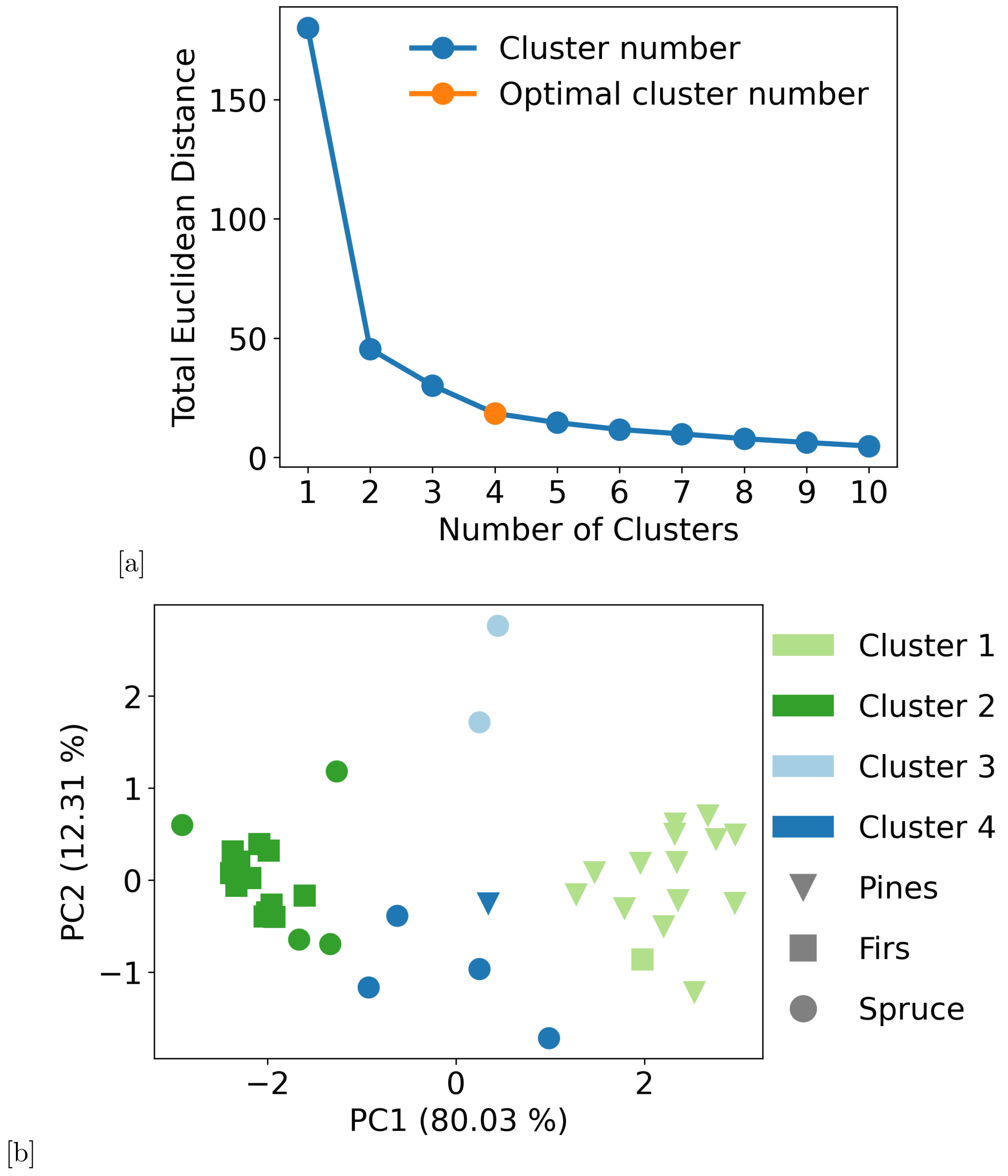


Figure 4.5: Plot [a] shows the elbow plot for k-means clustering with automated compound selection for the PC1 and PC2 pair. The orange marker indicates the optimum number of clusters. Plot [b] shows the PCA coupled with k-means clustering results for the PC1 and PC2 pair.

4.3.4 Beyond principal components one and two

Thus far, only the first two PCs (from a total of five) were used in the analysis since they explained 92% of the variance in the data set. Another 8% is shared between PCs three and four, which could potentially provide better separation for spruce samples. After testing PCs one with three and one with four, the combination of one and four resulted in better performance of the PR algorithm. Even though the optimal number of clusters for the PC1 and PC4 pair (Fig. 4.6 [a]) is the same as with the PC1 and PC2 pair it was found at a lower total Euclidean distance, which is indicative of more dense clustering. Figure 4.6 [b] shows the results for the PC1 and PC4 pair. Cluster one included 13 out of 16 pine samples and one overlapping fir sample. Cluster two included 11 out of 13 fir samples and only one overlapping spruce sample. Clusters three and four included nine spruce samples out of 10, one overlapping fir sample and three overlapping pine samples. With the PC1 and PC4 pair, spruce samples have 30% less overlap with firs (Fig. 4.6 [c]), with only moderate losses in the separation between spruce and pines. These results demonstrate the ability of the PR algorithm to separate firs, pines, and spruce in the smoke samples, with only two PCs accounting for most of the variance in the data set (PC1 and PC4 about 82%).

4.3.5 Mixed samples

The PR algorithm selected compounds that separated smoke samples by the contribution of three individual fuel families (firs, pines and spruce). To test the algorithm for mixed fuel samples, as would be more common in the field, five synthetic fuel mixtures were used: 60% pine / 40% spruce, 60% fir / 40% spruce, 60% pine / 40% fir, 90% pine

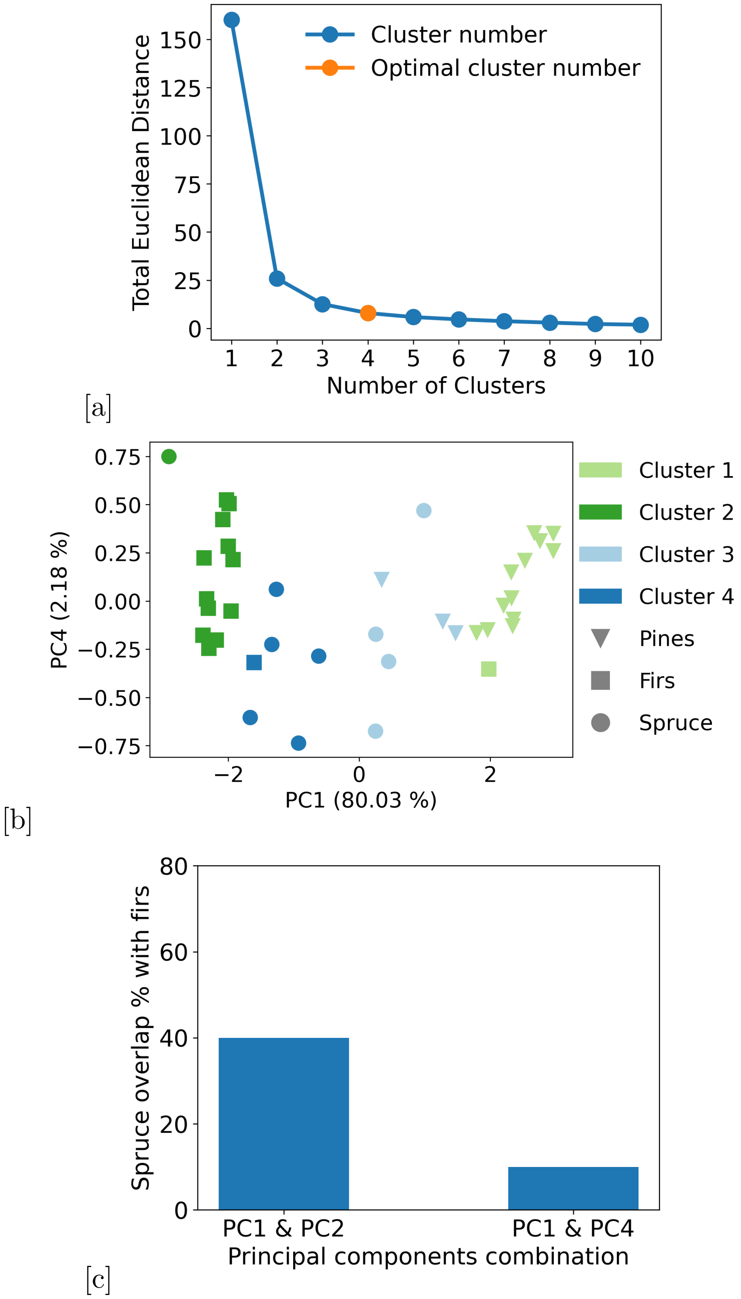


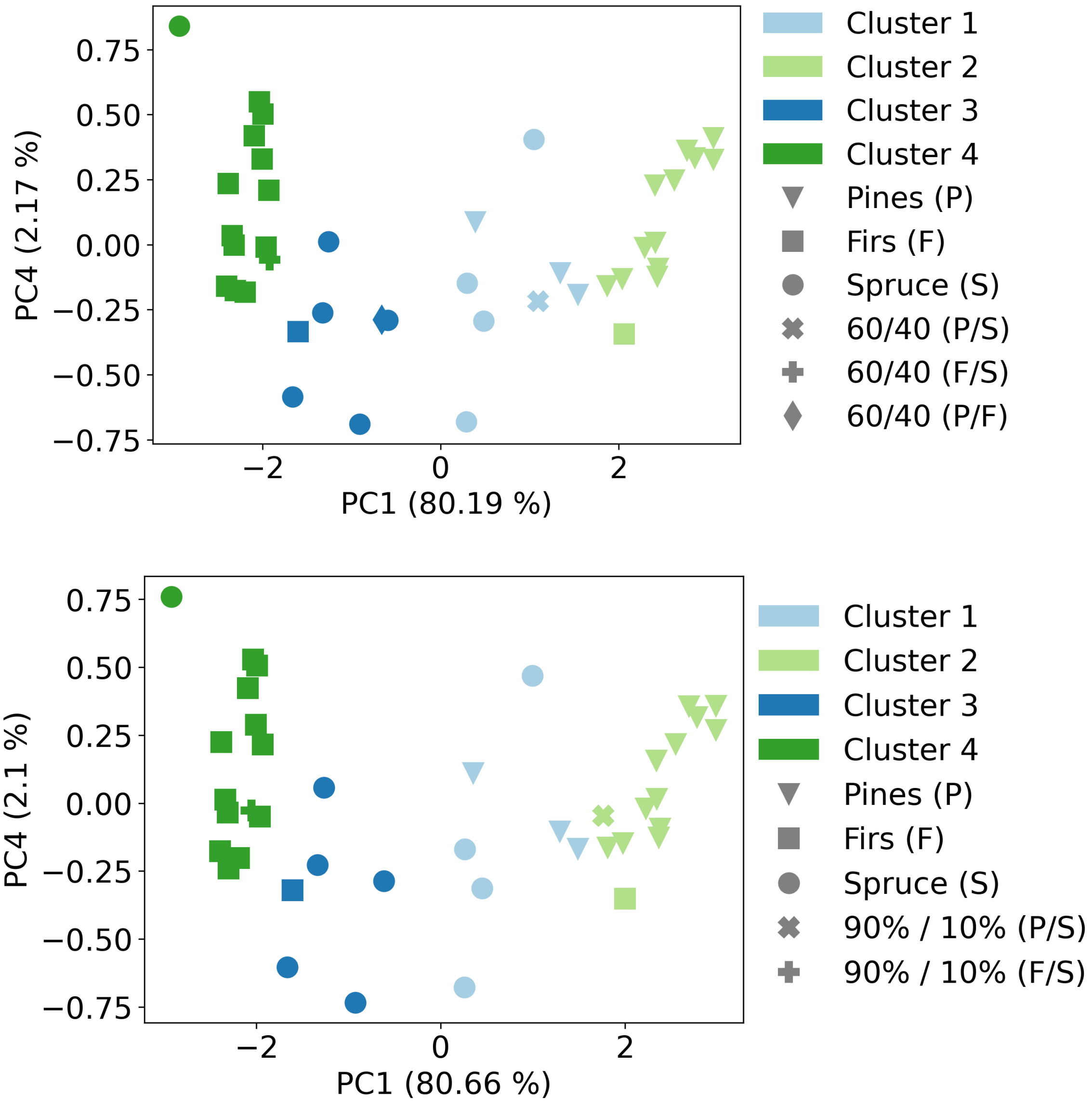
Figure 4.6: Plot [a] shows the elbow plot for k-means clustering with automated compound selection for the PC1 and PC4 pair. The orange marker indicates the optimum number of clusters. Plot [b] shows the PCA coupled with k-means clustering results for the PC1 and PC4 pair. Plots [c] shows the spruce overlap with firs based on the PC combination used.

/ 10% spruce, and 90% fir / 10% pine. From the three 60/40 samples only the fir/spruce synthetic mixture was clustered with the dominant fuel family (fir). The pine/spruce and pine/fir synthetic mixtures were clustered with spruce (Fig. 4.7 [a]), clusters one and three respectively. The grouping of the pine/spruce synthetic mixture as spruce is marginal in the PCA space and is due to the scatter of the spruce samples rather than the similarity of the synthetic mixture with spruce. Of the three 60/40 synthetic mixtures, the pine/fir mixture was grouped with the spruce clusters. The grouping of the pine/fir mixture with spruce is more intuitive after comparing the normalized ER profile with that of spruce (Fig. 4.8), where the ER profile for the pine/fir mixtures is similar to spruce. Figure 4.7 [b] shows the PR results including the 90/10 synthetic mixtures. Both samples were grouped correctly with their respective dominant fuel family.

The results with the synthetic mixtures suggest that the algorithm can select marker compounds that can differentiate fuel types even when they are highly mixed (60/40 cases for pine/spruce and fir/spruce) but for some mixtures (60/40 pine/fir) the differentiation might be poor. One approach for improving separation using the PR algorithm, could be to incorporate more mixed fuel samples in the training and test sets.

4.4 Classification

For the classification algorithm, the scores of the selected PCs were used as input for the LDA training. PC1 and PC4 were selected since they provided better separation across the three fuel families (section 3.2.3), while explaining 82% of the variability in the



[a]

[b]

Figure 4.7: Plot [a] shows the PCA coupled with k-means clustering results for the PC1 and PC4 pair including the 60%/40% synthetic mixtures. Plot [b] shows the PCA coupled with k-means clustering results for the PC1 and PC4 pair including the 90%/10% synthetic mixtures.

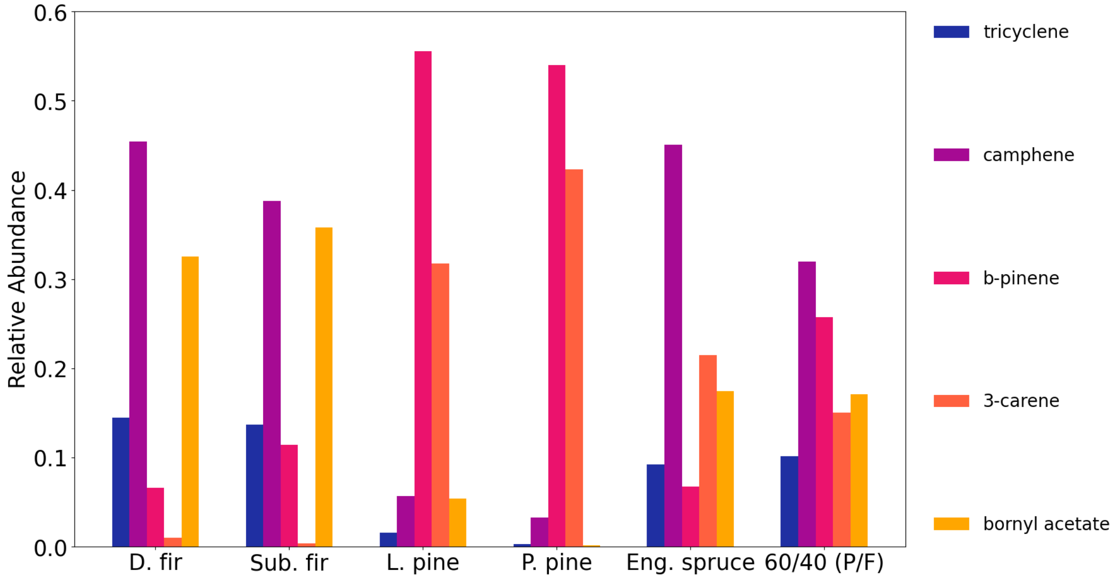


Figure 4.8: Normalized emission ratio profiles for: Douglas fir, subalpine fir, lodgepole pine, ponderosa pine, Engelmann spruce and the 60/60 pine fir synthetic sample.

data set. In this application the probability score is related to the proximity of sample to a class of samples (cluster) in the PCA space (Fig. 4.7 [a]-[b]) which is linked to its similarity with the emission profiles for the three fuel families (Fig. 4.4). The assignment of a sample to a class is based on the class with the highest probability, even if marginally higher. For example a sample with a pine probability score of 70% or more will most likely be inside the pine cluster. Generally, samples with probability scores 60% and higher are most likely in the cluster space of a fuel family. Samples with a probability score 60% and lower are more likely to be adjacent to more than one fuel family in the PCA space. For the training of the classifier all 39 samples from FIREX FL16 were used.

4.4.1 Synthetic mixtures and FLAME-4 FL12 samples

The classification algorithm was tested using the synthetic mixtures and FLAME-4 FL12 samples before testing using the BFRS field data. The classification results for the synthetic mixtures are shown in Fig. 4.9. Two of three 60/40 synthetic mixtures were classified correctly, pine/spruce and fir/spruce, with classification probabilities of 70% and higher for the dominant fuel family. The 60/40 pine/fir synthetic mixture was classified as spruce. Its classification is a result of its clustering in the PCA space with spruce (Fig. 4.7 [a]), which is directly connected to its similarity with the spruce emissions profile (Fig. 4.8). The two 90/10 synthetic mixtures, pine/spruce and fir/spruce, were correctly classified with classification probabilities over 80% for the dominant fuel family. Application of the classifier to the synthetic mixtures demonstrated that the mixtures can be correctly classified based on the dominant fuel family in mixed fuel samples (4 of 5 mixtures); however, incorrect classification can occur when the mixed fuel emissions profiles are similar to individual fuel emissions profiles, resulting in poorer separation with PCA. The results of the classification algorithm for mixed samples can be improved in future work through expanded testing and training on a broader range of fuel families and relevant mixtures.

The classification results for the FLAME-4 FL12 samples are shown in Fig. 4.9. This data set included six fuel species (ponderosa pine, black spruce, Indonesian peat, rice straw, wiregrass, and sawgrass); only one of which, ponderosa pine, was in the training set. Both ponderosa pine and black spruce samples were classified correctly (Fig. 4.9), with classification probabilities over 90%. The Indonesian peat, rice straw, wiregrass, and

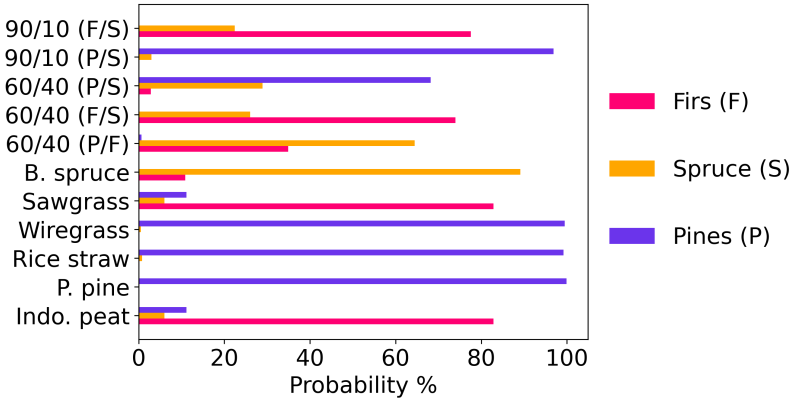


Figure 4.9: Classification results for synthetic mixtures and FLAME-4 samples.

sawgrass samples were classified as firs or pines (Fig. 4.9), with classification probabilities over 70%. The classification algorithm evaluated partial similarity against only three options (pine, firs or spruce), none of which represent the fuel families of the four fuel species. Figure 4.10 shows the average normalized emission ratio profiles for pines and firs, as well as Indonesian peat, rice straw, wiregrass, and sawgrass. It can be seen that only camphene appears in the sawgrass and Indonesian peat samples, and thus these fuels are classified as firs, which also have a high relative abundance of camphene. Wiregrass and rice straw samples also include camphene, but have higher relative abundances of b-pinene and 3-carene, and thus were classified as pines, which also have higher relative abundances of these two compounds (Fig. 4.10). As illustrated by the application to the synthetic fuel mixtures, the performance of the classification algorithm can be improved in future work by expanding the range of fuel families and mixtures included in the training and test sets.

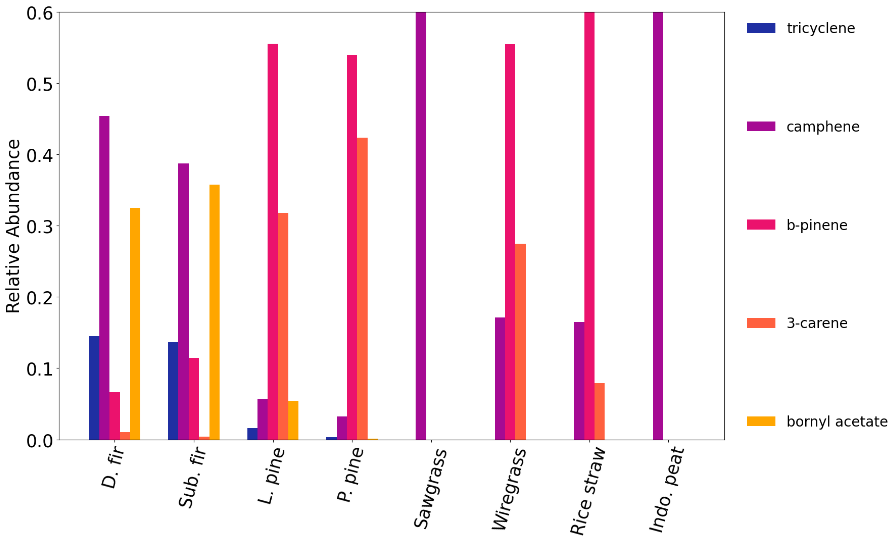


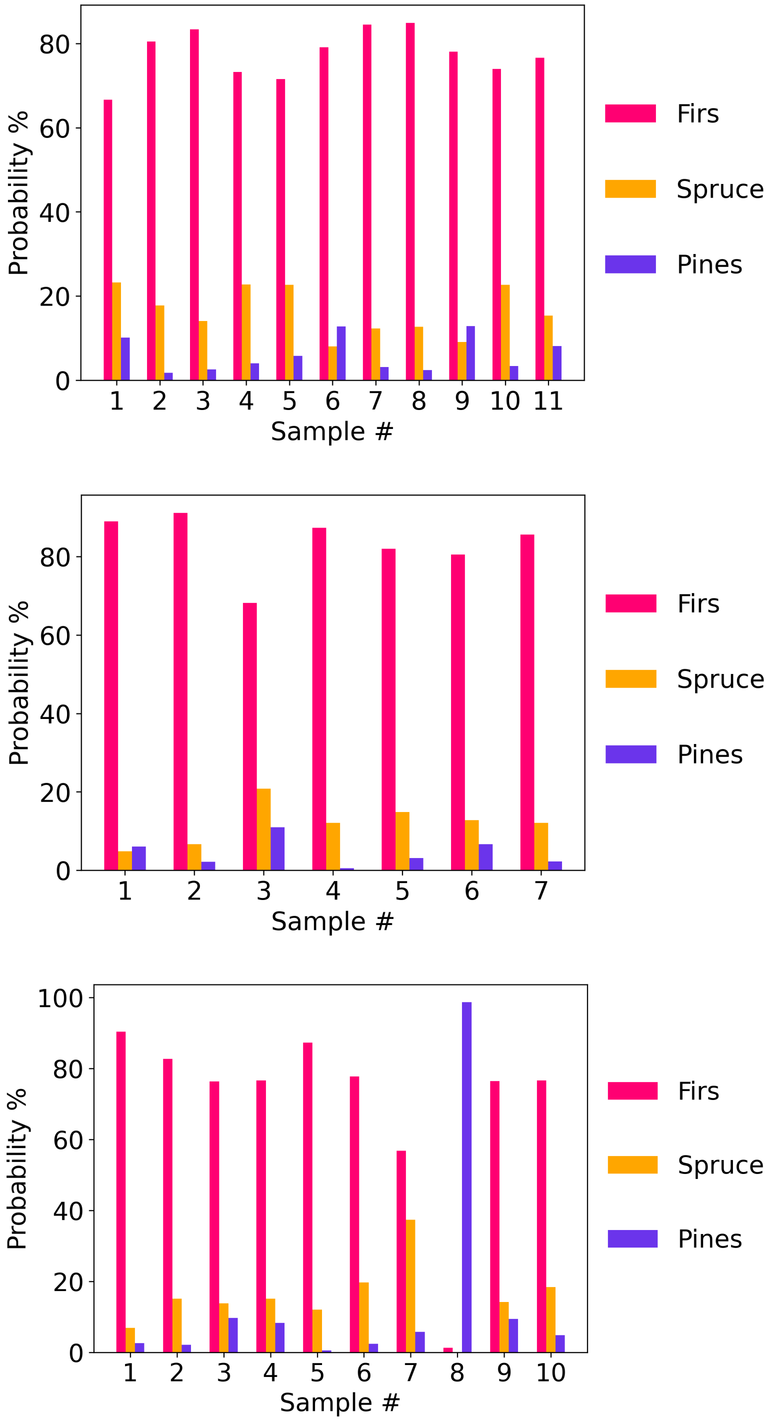
Figure 4.10: Normalized emission ratio profiles for FIREX FL16 samples: pines and firs; and FLAME-4 FL12 samples: sawgrass, wiregrass, rice straw, and Indonesian peat. The relative abundances of camphene and b-pinene in sawgrass, Indonesian peat, and rice straw were > 0.6, but for figure clarity, the axis limits were not changed.

4.4.2 Blodgett samples

Figures 4.11 [a]-[c] show the results of the classification algorithm applied to the BFRS samples from three different prescribed burn plots: 60, 340, and 400. Based on the fuel bed composition (Figs. 4.12-4.13) there was a total of seven different fuel species in the three burned plots: white fir, incense cedar, tanoak, sugar pine, ponderosa pine, Douglas fir, and California black oak. Due to the heterogeneity of the fuels, and the influence of meteorology and sampling location, it was not possible to determine the relative contribution of each fuel species to each sample. Instead the average overstory composition (Figs. 4.12 and 4.13) was used to determine likely influences from dominant sources close to each sampling location. For plot 60 (sites one and two) (Fig. 4.12 [a]) the main influence was from firs (47%) followed by similar amounts of pines and incense cedar (25% and 27%)

with no contribution from tanoak or California black oak. For site three in plot 60 (Fig. 4.12 [b]) the main influence was from incense cedar (43%) followed by firs (34%), pines (26%), and California black oak (10%). For plot 340 (Fig. 4.13 [a]) the main influence was from firs (63%) followed by pines (21%), incense cedar (12%), and (2%) from tanoak and California black oak. Finally for plot 400 (Fig. 4.12 [b]) the main influence was from firs (55%) followed by pines (26%) and incense cedar (18%). The classification algorithm classified all samples from plots 60 (Fig. 4.11 [a]) and 340 (Fig. 4.11 [b]) as fir dominant. Nine out of ten samples from plot 400 (Fig. 4.11 [c]) were classified as fir dominant and one as pine dominant. While spruce was absent in the burned plots, all samples (with the exception of the pine dominant sample in plot 400) had a higher classification probability for spruce than pines.

For plot 60 there were a total of 11 samples collected. Five samples in sites one and two, which is fir dominant (Fig. 4.12 [a]) and six samples in site three which is incense cedar dominant (Fig. 4.12 [b]). For sites one and two the classifier results are reasonable based on the overstory composition, but for site three the classification results are inconclusive since no emissions profiles were available for incense cedar which does not belong to the fir fuel family. It is likely that incense cedar or the mixture of incense cedar with firs most closely resembles the fir emissions profile of the selected compounds and thus was classified as firs. For plots 340 and 400 the classification results are reasonable, 16 out of 17 samples (Figs. 4.11 [b]-[c]) are fir dominant, since in both plots the influence of firs is above 50% with 63% contribution in 340 and 55% in 400. The one sample that was classified as pines in plot 400 was most likely affected strongly by ponderosa pine emissions during sampling. SI Fig. S4 (plot 400) in [1] shows that one of the inventoried land plots next to sites one and two had an average fractional overstory composition of ponderosa pine of more than 50%. Regarding the elevated probability for spruce, despite its absence from all burned plots, it is likely an artifact of mixed smoke between pines and firs. In section 3.3.1 the analysis showed that the synthetic mixture between pine and fir was more similar to spruce (Fig. 4.7 [a] and Fig. 4.8). Among the three plots, pines and firs together account for more than 70% on average of the overstory composition. Thus the contribution from both firs and pines could lead to smoke mixtures that resemble the spruce emissions profile. Tanoak and California black oak (Figs. 4.12 and 4.13) account for 2% - 10% of the total contributions among the three plots. Due to insufficient data regarding their emission profiles of the selected compounds their true contribution to the smoke samples could not be evaluated, but given their low overstory contribution it is likely that they did not influence the collected samples substantially. The results for the BFRS data showed that the lab-based emission profiles selected by the PR algorithm, can be applied to smoke samples collected in the field and can detect dominant fuel sources even in mixed smoke samples. While the algorithm has been tested and trained on only three fuel families, widespread application can be achieved with further training and testing using a more diverse set of compounds and broader range of fuel types.

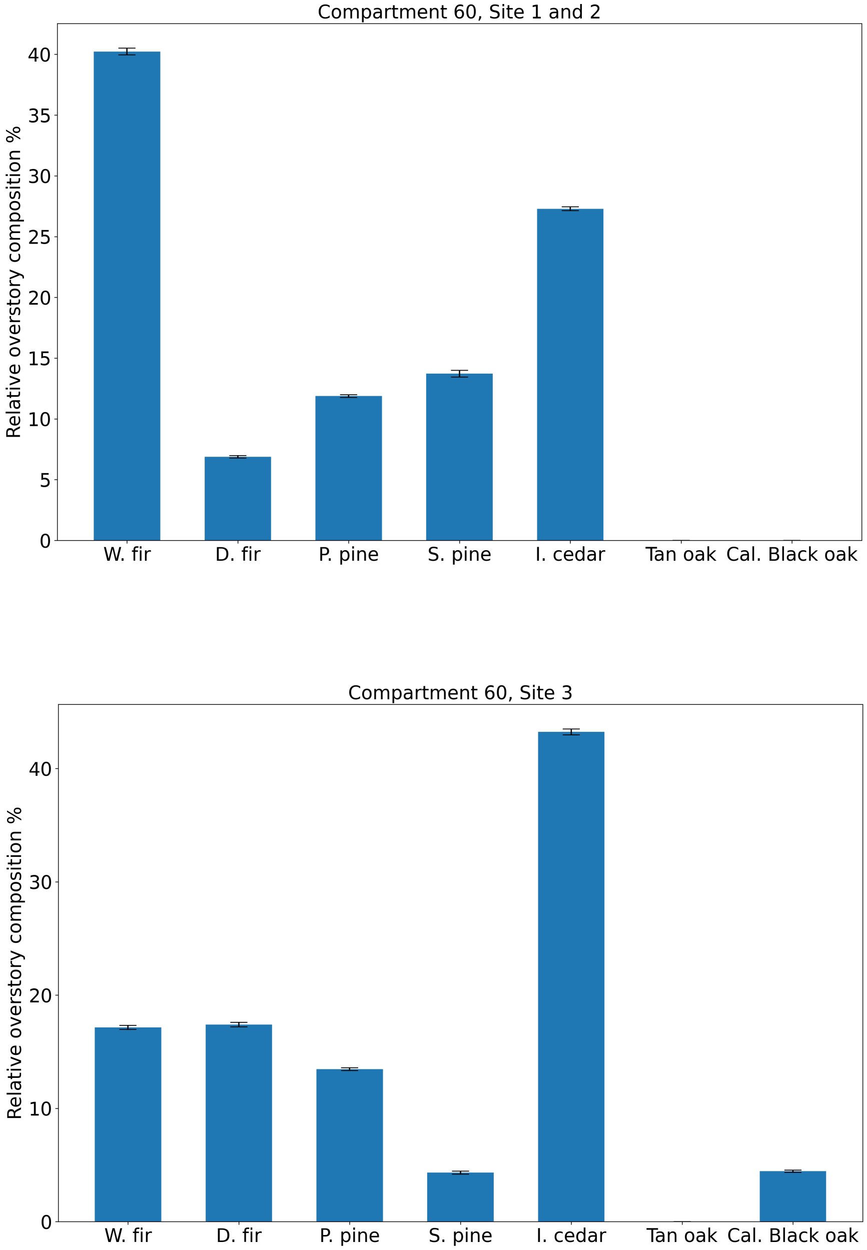


[a]

[b]

[c]

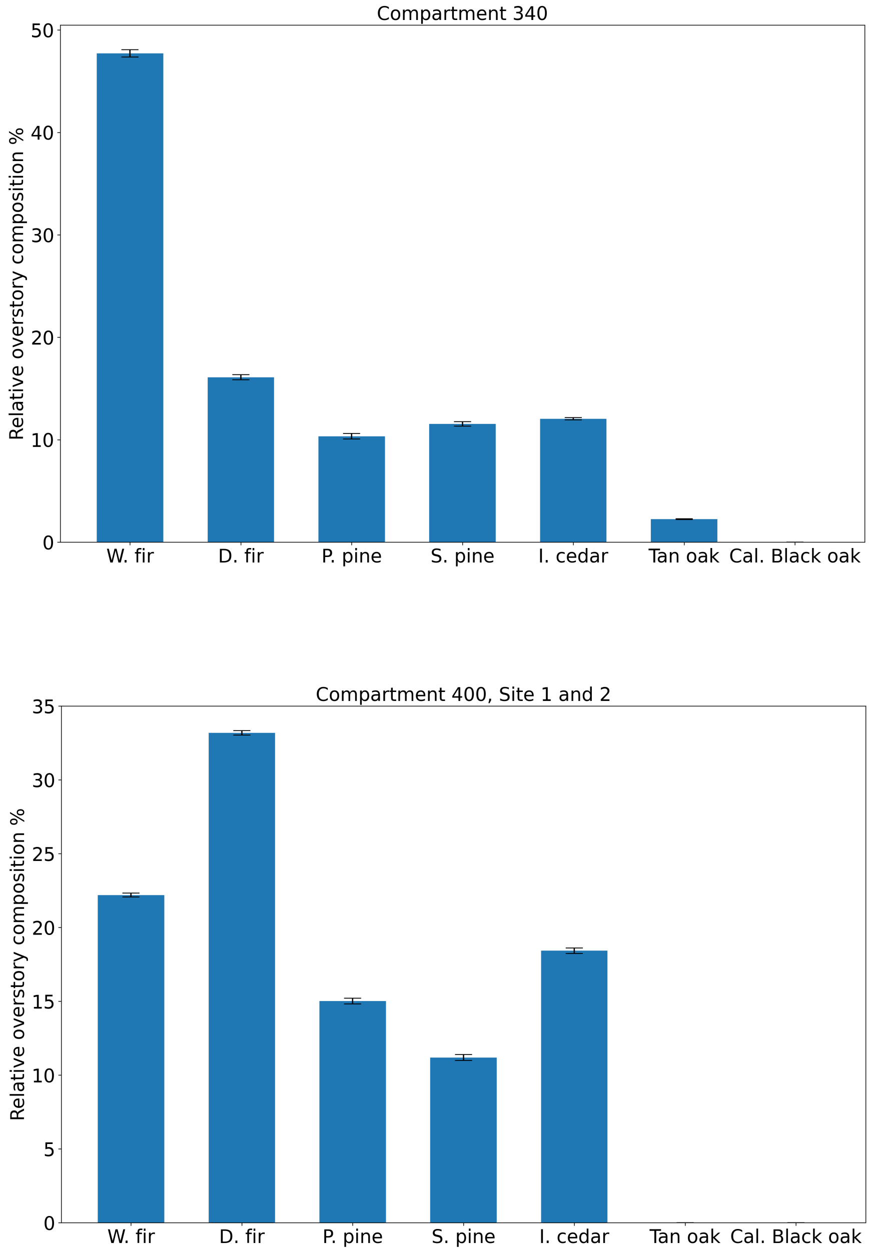
Figure 4.11: Plots [a]-[c] shows the classification probability by fuel class for plots 60, 340 and 400.



[a]

[b]

Figure 4.12: Plots [a]-[b] show the average overstory composition for plots 60 sites 1, 2, and 3. The fractional contribution of each species was determined by the proportion of basal area among all trees with diameter > 11.4 cm at breast height. Data for the overstory composition were taken from the supplemental material in [1].



[a]

[b]

Figure 4.13: Same as Fig. 4.12 but for plots 340 and 400.

4.5 Summary

A supervised PR algorithm was developed and applied in this study to: 1) differentiate sources/fuel types using NMOCs measured in smoke samples and selected with an ANOVA based feature selection method; and 2) train a classification algorithm to detect dominant sources/fuel types in smoke samples based on the unique speciation profiles identified by the PR algorithm. The PR algorithm was able to group five fuel species (Douglas and subalpine fir, ponderosa and loblolly pine, and Engelmann spruce) into three fuel families (pines, firs and spruce), with minimum overlap; only 5 of 39 total samples were grouped with families that were not representative of the fuel species. The separation was achieved using five monoterpenoids that the algorithm selected out of a pool of 93. The PR algorithm was tested with five synthetic fuel mixtures where it successfully separated three of five (60/40 fir/spruce, 90/10 pine/spruce, and 90/10 fir/spruce) and grouped them with their dominant fuel type. The same synthetic fuel mixtures were also tested using the classification algorithm, where four of five were classified correctly (60/40 pine/spruce, 60/40 fir/spruce, 90/10 pine/spruce, and 90/10 fir/spruce). The application of the classification algorithm to the synthetic mixtures demonstrated that dominant source contributions could be identified in fuel mixtures. For the FLAME-4 FL12 samples the classification algorithm correctly classified two of six samples (ponderosa pine and black spruce); these two samples were the only fuels represented by the three fuel families. For the BFRS field samples, based on the fractional overstory composition, the classification results were reasonable with 27 out 28 samples being classified as fir dominant and one sample as pine dominant. The incorrect classifications that occurred with the synthetic fuel mixture (60/40 pine/fir) and the FLAME-4 FL12 samples (Indonesian peat, rice straw, wiregrass, and sawgrass) were due to the similarity or partial similarity of their emissions profiles with the fuels used to train the classification model. This can be resolved in future applications in two ways; 1) by introducing a null case in the classification algorithm so that in cases such as with the grasses the algorithm will have the ability to not classify a specific fuel if there is partial similarity only or 2) by including more compounds and a broader range of fuel types, including mixtures. This will also facilitate the use of this approach for identifying contributing fuels outside of Western coniferous forests.

Chapter 5

Application of the Supervised Pattern Recognition Algorithm for Identification of SOA Precursors in Ambient Plume Samples

5.1 Introduction

Western wildland fires constitute a significant part of biomass burning (BB) in the US along with agricultural and prescribed burning. Overall, BB is globally the largest source of combustion-related source of NMOCs [20][21]. NMOCs have a significant influence on the atmosphere, which includes a major contribution to the formation of SOA. Given the complex chemistry and dynamics of SOA formation from fires[22] [23] there is a need for better understanding fire emissions and their contribution to SOA formation. Laboratory studies of controlled burns have shown that NMOC oxidation results in SOA production and an enhancement in OA mass but with increased variability. Many studies have tried to narrow down the variability in observed OA enhancement by focusing on specific SOA precursor groups: 1) monoterpenes, 2) oxygenated aromatics, 3) heterocyclic compounds, and 4) polyaromatic hydrocarbons [24][25][26]. Ahern et. al. [24] found that monoterpenes are important SOA precursors for black spruce and ponderosa pine while furans are important for grasses. Lim et. al. [27] showed that there is a strong correlation between SOA produced and the initial NMOC mass, with stronger correlations observed for NMOCs more volatile than monoterpenes. These previous studies suggest that there is significant variability among dominant SOA precursors based on fuel type and that there is need to be able to better understand SOA from biomass-burning sources.

In this work, a method is presented for identifying SOA precursors in wildland fire smoke plumes. To overcome some of the existing limitations in identifying the contribution of specific precursors to SOA formation the pattern recognition (PR) algorithm developed in Chapters 3 and 4 was adapted to be used for data obtained during two field campaigns in 2018 and 2019 (WE-CAN and FIREX-AQ Field).

5.2 Data

The data used in this work were collected from airborne platforms from two fields campaigns that took place in 2018 and 2019. More specifically the Western Experiments for Cloud Chemistry, Aerosol Absorption and Climate (WE-CAN) and the FIREX-AQ, already mentioned in Chapter 4. While the WE-CAN campaign took place only in the pacific northwest (PNW) and focused on wildland fires, FIREX-AQ covered both the PNW region as well as the midwest that included prescribed fires as well. For the purposes of this analysis only the PNW data were used. Both campaigns and the instrumentation have been described in more detail elsewhere [90][91][92][93] as well as their dedicated websites [94][95]. Briefly WE-CAN and FIREX-AQ were multiagency projects to intensively characterize the emissions and evolution of those emissions from biomass burning. WE-CAN took place during July - September of 2018 and used the NSF/NCAR C-130 aircraft to sample regional smoke and smoke emitted from specific wildfires in the Western United States during 19 flights. FIREX-AQ took place during the summer of 2019 in the United States to study both wildfire and agricultural burning smoke and used both airborne (NOAA Twin Otter and NASA DC8) and ground means to sample smoke plumes as part of this campaign.

Finally for the purposes of this analysis only the data from the following instruments were used. NMOC mixing ratios were obtained for both campaigns from two instruments; 1) the trace organic gas analyzer (TOGA) [96] and a proton transfer reaction-time of flight-mass spectrometer (PTR-ToF-MS) (WE-CAN: University of Montana, Ionicon Analytik PTR-TOF-MS 4000, and FIREX-AQ: NOAA ESRL Chemical Sciences Division NOAA PTR-ToF-MS). Organic aerosol (OA) concentrations using the aerosol mass spectrometer (AMS) (WE-CAN: Colorado State University, and FIREX-AQ: CIRES and Department of Chemistry, University of Colorado Boulder). For CO, data from two different instruments were used, the Picarro G2401-m WS-CRDS (National Center for Atmospheric Research) from WE-CAN and the Los Gatos Research (LGR) N2O/CO/H2O instrument (NOAA ESRL Chemical Sciences Division) from FIREX-AQ.

5.3 Algorithm Implementation, Results and Discussion

5.3.1 Data pre-processing

The data pre-processing methods described in Chapter 2 section 2.1 were all applied to the data sets prior to the PR analysis but there were also a few additional steps that were performed for this study. All the additional steps are described in the following paragraph.

The first step for both WE-CAN and FIREX-AQ data was to align the data from the different instruments. That step was necessary because the instruments were sampling at slightly different time resolutions. First for each campaign the instrument with the lowest time resolution was identified, in both campaigns it was TOGA with a time resolution of half a minute. All other instruments had an one second time resolution. Then the measurements from other instruments were averaged for each start and stop time stamps from TOGA. Following the data alignment the NMOC data from TOGA and the PTR-ToF-MS were merged. Specifically only the oxygenated aromatics from the PTR-ToF-MS. This was done because TOGA cannot measure oxygenated aromatics and this group has been shown to be important for SOA formation[97].

After the alignment and the merge between the PTR-ToF-MS and TOGA the normalized mixing ratios ($\frac{\mathrm{NMOC_i}}{\mathrm{CO}}$ and $\frac{\mathrm{OA}}{\mathrm{CO}}$) for all NMOCs and OA were calculated in order to account for plume dilution downwind. The next step in the pre-processing was to calculate the ΔOA and $\Delta\mathrm{NMOC_i}$ for every plume that was intercepted. This was done in two stages. First all samples were normalized using the min max approach [38] (Fig. 5.1) and filtered using a CO threshold of 0.1 in order to avoid interference with samples outside the plume. Then after reversing te normalization using an average time window of approximately an hour the ΔOA and $\Delta\mathrm{NMOC_i}$ were calculated using the first four points right after the first plume interception as $\mathrm{OA_{initial}}$ and $\mathrm{NMOC_{initial}}$ and the last four points as $\mathrm{OA_{final}}$ and $\mathrm{NMOC_{final}}$ (Fig. 5.1). The ΔOA and $\Delta\mathrm{NMOC_i}$ were calculated for two reasons; 1) the ΔOA shows in which cases there was a positive OA enhancement and which cases there was no enhancement, 2) the $\Delta\mathrm{NMOC_i}$ provide much more information based on the sign of the value. A positive or almost zero $\Delta\mathrm{NMOC_i}$ shows that the specific compound most likely did not contribute to SOA formation while a negative $\Delta\mathrm{NMOC_i}$ indicates that the compound reacted away and in some capacity might be informative about SOA formation. As for the one hour time window it was chosen; 1) so that the ΔOA and $\Delta\mathrm{NMOC_i}$ were calculated for plumes with similar aging and 2) after investigation of the OA time series the ΔOA values for a time window of more than an hour were all negative. It is likely that for longer than an hour time windows processes such as extensive dilution, which leads to SOA volatilization and therefore SOA losses affected the SOA levels.

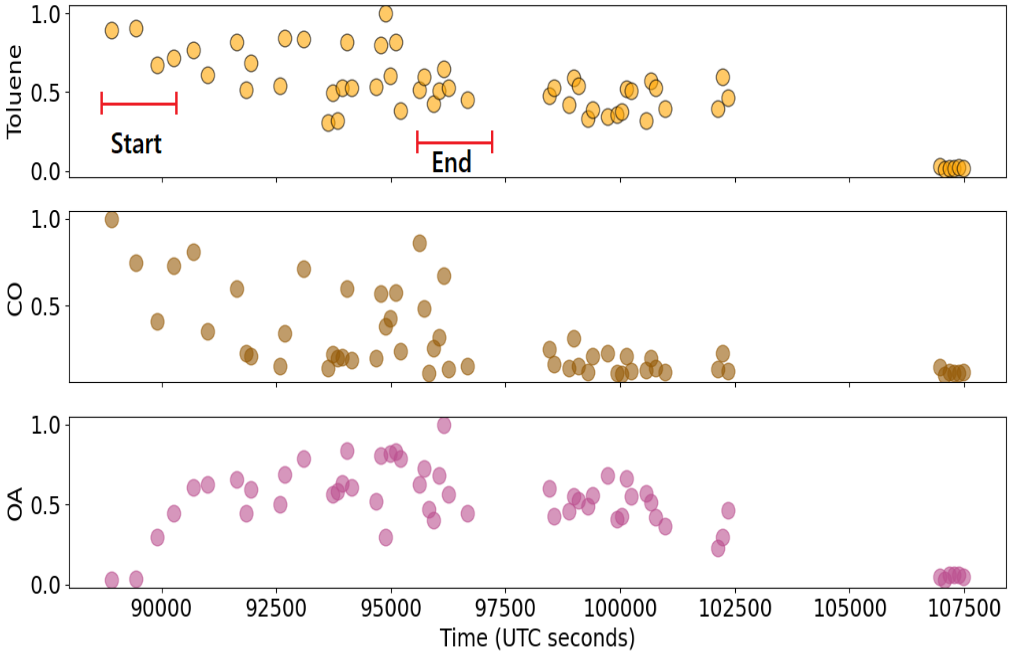


Figure 5.1: Min max normalized levels of toluene, CO and OA for one of the FIREX-AQ flights. The markers on the toluene plot show the start and end points for the calculation of ΔOA and $\Delta NMOC_i$.

Finally after the calculation of the ΔOA and $\Delta NMOC_i$ the data from WE-CAN and FIREX-AQ were merged into one data set and the basic pre-processing before the PR performed as described in Chapter 2, section 2.1. The final data set was consisted of 34 samples and 46 NMOCs.

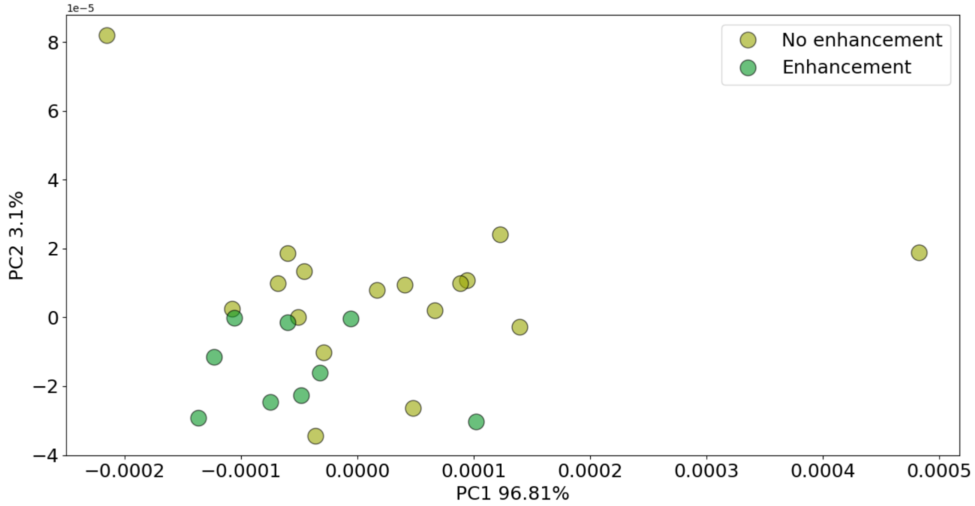
5.3.2 Pattern recognition results

In order to initiate the algorithm the classes that will be separated need to be defined. In this application the end goal was to target compounds that provide process level information about SOA formation. Therefore the classes that were created were based on

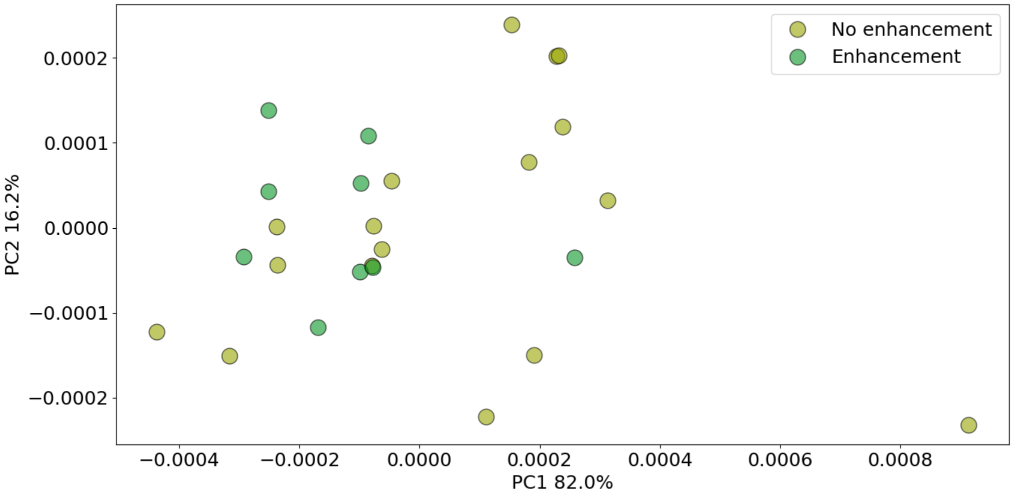
Table 5.1: Selected compounds for the two cases.

Case 1 (5 compounds)	Case 2 (10 compounds)
i-pentane	i-pentane
2-methyl, pentane	2-methyl, pentane
2,2,4-trimethyl, pentane	2,2,4-trimethyl, pentane
toluene	toluene
ethylbenzene	ethylbenzene
	m+p-xylenes
	benzene
	3-methyl, pentane
	phenol
	syringol

the ΔOA. Specifically two classes were created. Class one was for samples that had positive OA enhancement ($\Delta\text{OA} > 0$) and the second class was for samples that showed negative or no OA enhancement ($\Delta\text{OA} \leq 0$). Furthermore in this application of the PR algorithm the *k*-means component was not used. Instead the PCA plots and class separation was evaluated by using different colour coding for the samples in the PCA space between the two classes. The modified algorithm was tested for two cases; 1) with five selected compounds got out of 46 and 2) with ten (Table 5.1). Cases with more than ten compounds were investigated but did not yield better results. Figure. 5.2 shows the results for five and ten compounds. While in both cases the explained variance from the two retained PCs is over 80% there is no clear separation between the two classes of samples.



[a]



[b]

Figure 5.2: Plot [a] shows the PCA results for five compounds. Plot [b] shows the PCA results for ten compounds.

In Chapters 3 and 4 it was shown that the class separation is directly attributable to the selected compounds and their unique profiles between the two classes. In this case the average $\Delta\mathrm{NMOC_i}$ for each selected compound was plotted for the two classes (Figs. 5.3 [a]-[b]) along with their respective error bars. From Fig. 5.3 [a]-[b] it can be seen that while there are differences between the two classes there is substantial variability per compound for both cases (five and ten compounds).

While some variability might not pose a problem because the F-ratio is used for ranking purposes rather performing statistical inference in the algorithm, substantial variability will negatively impact the feature selection. Along with the calculation of the F-ratio there is a p-value that accompanies it and determines whether or not the difference between the classes reflected in the F-ratio is significant or not. As a rule of thumb a p-value ≤ 0.05 points to a significant different between the examined classes while a p-value > 0.05 means that any difference observed might be due to random variation. Figure 5.4 shows the F-ratio value for all the compounds in the data set along with respective p-values and the 0.05 threshold. From both Figs. 5.3 and 5.4 two things are evident. First the selected compounds have substantial variability and second the broader suit of compounds suffers from the same problems as it renders any difference between the classes statistically not significant.

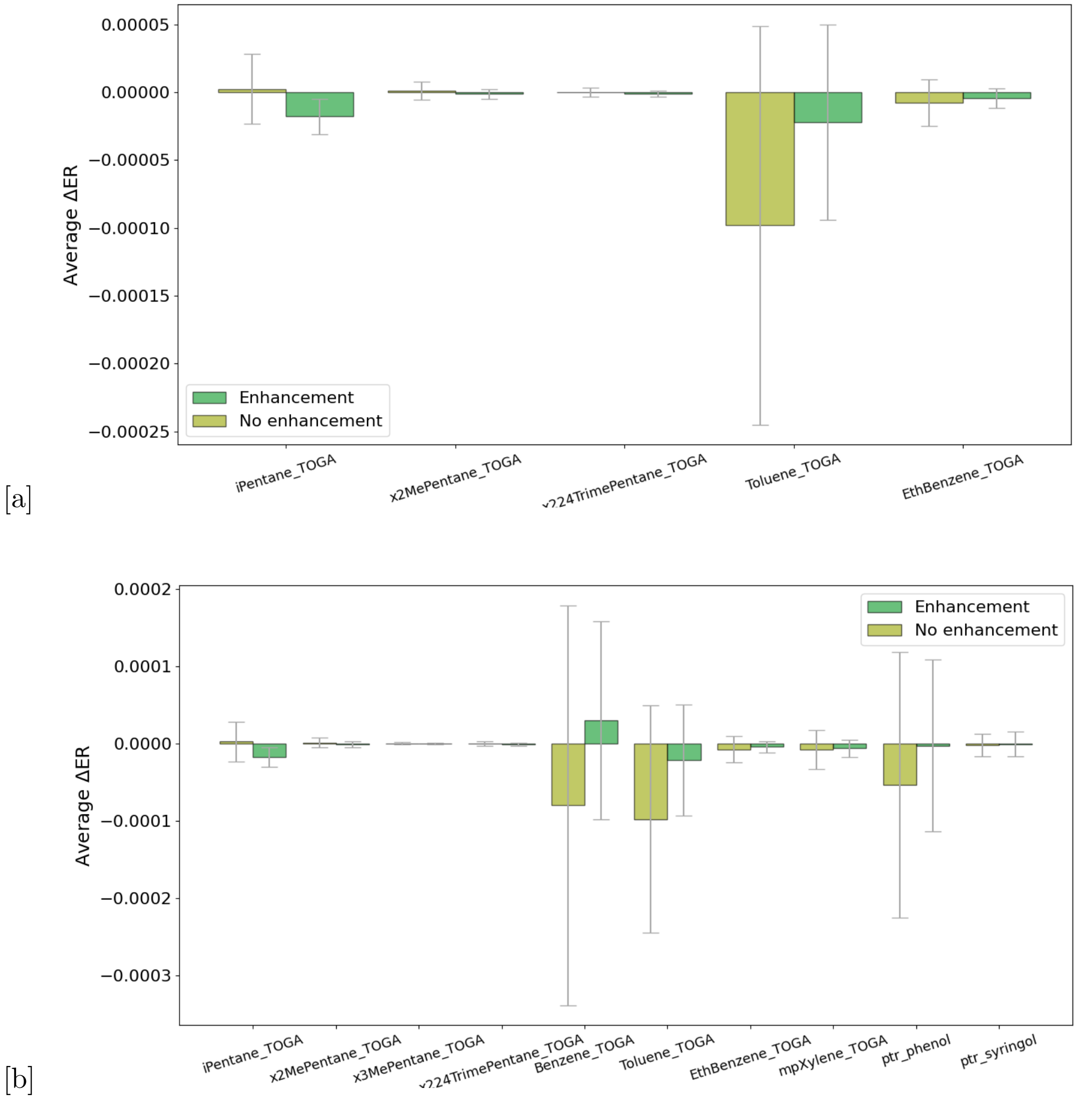


Figure 5.3: Plot [a] shows the average $\Delta\text{NMOC}_\text{i}$ for 5 compounds. Plot [b] shows the same as [a] but for 10 compounds.

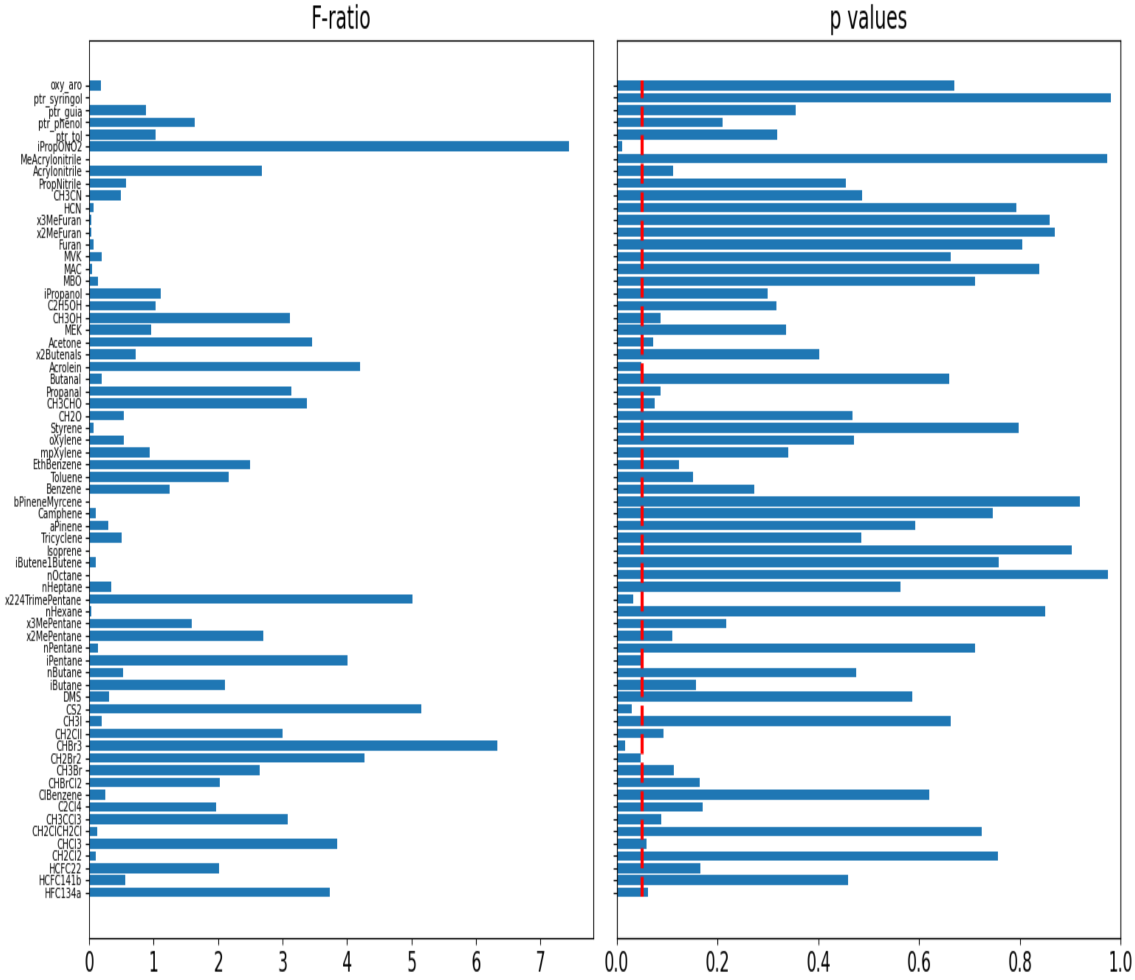


Figure 5.4: Left plot shows the F-ratio values for all compounds in the data set. Right plot shows the p-values for every compound. The red dashed line shows the threshold of statistical significance (0.05).

5.3.3 Dealing with feature dispersion

In order to work around the high variability in the data set there were two options. One is to try to collect more samples in order to constrain the variation. In this case though that is not feasible. The second option, which was followed, is to try to use a different set or variables that might not be as affected from the natural variation in the data while they provide similar information with the original variables ($\Delta\mathrm{NMOC_i}$).

One set of variables was tested with the PR algorithm. This time the $\Delta\mathrm{NMOC_i}$ were calculated but for entire groups of compounds. Specifically alkanes, terpenes, furans, aromatics and oxygenated aromatics. While individual compounds vary substantially as shown in Fig. 5.3 [a]-[b] the entire group can be more robust to natural variation. The algorithm results for the five groups of compounds are shown in Fig. 5.5 [a]-[b]. The algorithm with the five groups of compounds did not yield better results in terms of class separation (Fig. 5.5 [a]). At the same time the variability in the features was not reduced based on the average $\Delta\mathrm{NMOC_i}$ plot (Fig. 5.5 [b]).

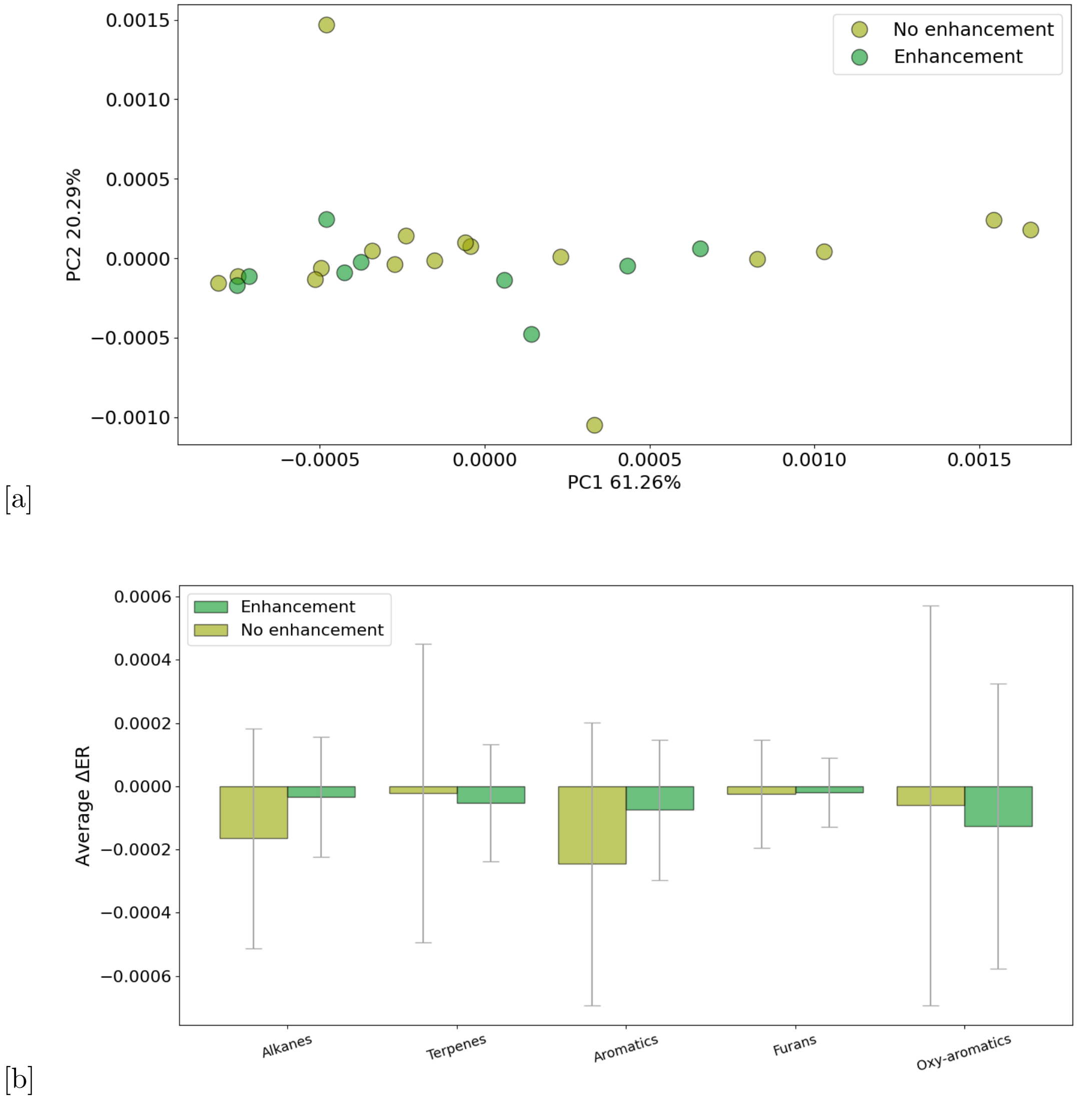


Figure 5.5: Plot [a] shows the PCA results for the 5 groups of compounds. Plot [b] shows the average $\Delta \mathrm{NMOC}_i$ for every group between the two classes.

5.4 Summary

In conclusion, the PR algorithm did not manage to separate the samples between the two specified classes of positive OA enhancment and no OA enhancement. The main issue was found to be the substantial variability of the NMOCs between the two specified classes. One alternative approach was tested in order to overcome this problem. Instead of using individual NMOCs the $\Delta\mathrm{NMOC_i}$ were calculated for entire groups of compounds. Specifically alkanes, terpenes, furans, aromatics and oxygenated aromatics. While this approach provided more distinct $\Delta\mathrm{NMOC_i}$s (Fig. 5.5 [b]) there was no separation between the two classes (Fig. 5.5 [a]) and the variation was not reduced.

The major issue in both cases was the excessive variability or else the signal to noise ratio of each variable between the two classes. This was mostly the result of plume to plume variability attributed to different factors such as different fire intensities, differences in the photochemical age of the plumes, plume edge-to-center variations for the NMOCs, and oxidant levels (OH, O_3 and NO_3) which are expected to be highly variable in the field compared to the lab. Future efforts for data driven SOA precursor detection should focus first on constraining the plume to plume variability by either targeting and analyzing plumes with similar properties (e.g. plume photochemical age, oxidant levels, etc.) to the extent possible or focus on a single plume analysis by increasing the sampling efforts on an individual plume.

Chapter 6

Conclusions

This thesis presents two tools for fingerprinting analyses of gaseous non-methane organic compounds (NMOCs) from emerging sources developed and tested in three applications. The first tool was a chromatographic alignment algorithm. The algorithm was successfully applied in the analysis of eight different tailpipe emissions samples (Chapter 3) and reduced the time required for manual alignment from several months to a few minutes. By minimizing time spent on chromatographic alignment there is more time to collect and analyze additional samples if needed, including for quality control analysis of samples. Further, automation of this task may help reduce errors associated with manual alignment.

The second tool was a supervised pattern recognition algorithm (Chapters 3 and 4) for the detection of unique emission profiles among different fuel types. The algorithm was successful in its application in two different projects, the eight different tailpipe emissions samples (Chapter 3) and the laboratory biomass burning smoke samples (Chapter 4). The algorithm; 1) found unique NMOC profiles among the tailpipe and smoke emissions samples and 2) using the the selected compounds in each case differntiated the samples. Furthermore using the unique NMOC profiles from the biomass burnng laboratory samples, a classification model was created that successfully classified prescribed burn smoke samples based on their dominant fuel source. The algorithm is particularly useful for detecting unique emission profiles for different fuels that can be used for smoke source identification in the field. This also allows selection of unique NMOC profile(s) that could be used to represent the emissions of the specific fuel types in air quality models.

The use of the second tool (pattern recognition algorithm) was explored in a third application. Using data from two field campaigns, WE-CAN and FIREX-AQ (Chapter 5) the algorithm was applied on selected plume transects in order to target compounds of interest that can potentially provide process level information regarding SOA formation. While the algorithm did select specific compounds there was no clear separation between samples with positive and neutral/negative enhancement. Further investigation showed that the plume to plume variability overwhelmed the NMOC concentrations which resulted in poor statistical power for proper NMOC selection. Future efforts should focus first on mitigating plume to plume variability by grouping plumes using metrics such as plume age and then applying data driven models.

The chromatographic alignment and pattern recognition algorithms developed and presented in this thesis are not application specific. The chromatographic alignment algorithm also has been used successfully in aligning urban NMOC emission samples collected in the Los Angeles Basin during the COVID-19 shutdown. The pattern recognition algorithm can also be used for the analysis of urban NMOCs. An interesting future application would be to couple the pattern recognition algorithm with a source apportionment method in order to detect unique decomposed emission profiles and help with the identification of emission sources based on algorithm-selected compounds. Documentation of the pattern recognition and classification algorithms, and example implementations can be found at the following GitHub repository: Pattern Recognition

Bibliography

[1] Lindsay E. Hatch, Coty N. Jen, Nathan M. Kreisberg, Vanessa Selimovic, Robert J. Yokelson, Christos Stamatis, Robert A. York, Daniel Foster, Scott L. Stephens, Allen H. Goldstein, and Kelley C. Barsanti. Highly speciated measurements of terpenoids emitted from laboratory and mixed-conifer forest prescribed fires. *Environmental Science & Technology*, 53(16):9418–9428, 2019. PMID: 31318536.

[2] Matthew James Alvarado and Ronald G. Prinn. Formation of ozone and growth of aerosols in young smoke plumes from biomass burning: 1. lagrangian parcel studies. *Journal of Geophysical Research: Atmospheres*, 114(D9), 2009.

[3] S. Zhou, S. Collier, D. A. Jaffe, N. L. Briggs, J. Hee, A. J. Sedlacek III, L. Kleinman, T. B. Onasch, and Q. Zhang. Regional influence of wildfires on aerosol chemistry in the western us and insights into atmospheric aging of biomass burning organic aerosol. *Atmospheric Chemistry and Physics*, 17(3):2477–2493, 2017.

[4] Sophie Tomaz, Tianqu Cui, Yuzhi Chen, Kenneth G. Sexton, James M. Roberts, Carsten Warneke, Robert J. Yokelson, Jason D. Surratt, and Barbara J. Turpin. Photochemical cloud processing of primary wildfire emissions as a potential source of secondary organic aerosol. *Environmental Science & Technology*, 52(19):11027–11037, 2018. PMID: 30153017.

[5] G. N. Theodoritsi and S. N. Pandis. Simulation of the chemical evolution of biomass burning organic aerosol. *Atmospheric Chemistry and Physics*, 19(8):5403–5415, 2019.

[6] G. R. McMeeking, S. M. Kreidenweis, C. M. Carrico, T. Lee, J. L. Collett Jr., and W. C. Malm. Observations of smoke-influenced aerosol during the yosemite aerosol characterization study: Size distributions and chemical composition. *Journal of Geophysical Research: Atmospheres*, 110(D9), 2005.

[7] Donald McKenzie, Susan M. O'Neill, Narasimhan K. Larkin, and Robert A. Norheim. Integrating models to predict regional haze from wildland fire. *Ecological Modelling*, 199(3):278–288, 2006. Ecological Models as Decision Tools in the 21st Century.

[8] Rokjin J. Park, Daniel J. Jacob, Naresh Kumar, and Robert M. Yantosca. Regional visibility statistics in the united states: Natural and transboundary pollution influences,

and implications for the regional haze rule. *Atmospheric Environment*, 40(28):5405–5423, Sep 2006.

[9] Y. Q. Hu, N. Fernandez-Anez, T. E. L. Smith, and G. Rein. Review of emissions from smouldering peat fires and their contribution to regional haze episodes. *INTERNATIONAL JOURNAL OF WILDLAND FIRE*, 27(5):293–312, 2018.

[10] Daniel A. Jaffe, Susan M. O'Neill, Narasimhan K. Larkin, Amara L. Holder, David L. Peterson, Jessica E. Halofsky, and Ana G. Rappold. Wildfire and prescribed burning impacts on air quality in the united states. *Journal of the Air & Waste Management Association*, 70(6):583–615, 2020. PMID: 32240055.

[11] J. D. Miller, H. D. Safford, M. Crimmins, and A. E. Thode. Quantitative evidence for increasing forest fire severity in the sierra nevada and southern cascade mountains, california and nevada, usa. *Ecosystems*, 12(1):16–32, Feb 2009.

[12] Philip E. Dennison, Simon C. Brewer, James D. Arnold, and Max A. Moritz. Large wildfire trends in the western united states, 1984–2011. *Geophysical Research Letters*, 41(8):2928–2933, 2014.

[13] McDonald Brian C., de Gouw Joost A., Gilman Jessica B., Jathar Shantanu H., Akherati Ali, Cappa Christopher D., Jimenez Jose L., Lee-Taylor Julia, Hayes Patrick L., McKeen Stuart A., Cui Yu Yan, Kim Si-Wan, Gentner Drew R., Isaacman-VanWertz Gabriel, Goldstein Allen H., Harley Robert A., Frost Gregory J., Roberts James M., Ryerson Thomas B., and Trainer Michael. Volatile chemical products emerging as largest petrochemical source of urban organic emissions. *Science*, 359(6377):760–764, Feb 2018.

[14] Patrick Roth, Jiacheng Yang, Christos Stamatis, Kelley C. Barsanti, David R. Cocker, Thomas D. Durbin, Akua Asa-Awuku, and Georgios Karavalakis. Evaluating the relationships between aromatic and ethanol levels in gasoline on secondary aerosol formation from a gasoline direct injection vehicle. *Science of The Total Environment*, 737:140333, Oct 2020.

[15] Akihiro Fushimi, Yoshinori Kondo, Shinji Kobayashi, Yuji Fujitani, Katsumi Saitoh, Akinori Takami, and Kiyoshi Tanabe. Chemical composition and source of fine and nanoparticles from recent direct injection gasoline passenger cars: Effects of fuel and ambient temperature. *Atmospheric Environment*, 124:77–84, Jan 2016.

[16] Jiacheng Yang, Patrick Roth, Thomas D. Durbin, Kent C. Johnson, Akua Asa-Awuku, David R. Cocker, and Georgios Karavalakis. Investigation of the effect of mid- and high-level ethanol blends on the particulate and the mobile source air toxic emissions from a gasoline direct injection flex fuel vehicle. *Energy & Fuels*, 33(1):429–440, Jan 2019.

[17] Tak W. Chan, David Lax, Garry C. Gunter, Jill Hendren, Joseph Kubsh, and Rasto Brezny. Assessment of the fuel composition impact on black carbon mass, particle number size distributions, solid particle number, organic materials, and regulated gaseous

emissions from a light-duty gasoline direct injection truck and passenger car. *Energy & Fuels*, 31(10):10452–10466, Oct 2017.

[18] J. Peng, M. Hu, Z. Du, Y. Wang, J. Zheng, W. Zhang, Y. Yang, Y. Qin, R. Zheng, Y. Xiao, Y. Wu, S. Lu, Z. Wu, S. Guo, H. Mao, and S. Shuai. Gasoline aromatics: a critical determinant of urban secondary organic aerosol formation. *Atmos. Chem. Phys.*, 17(17):10743–10752, Sep 2017.

[19] A. L. Westerling, H. G. Hidalgo, D. R. Cayan, and T. W. Swetnam. Warming and earlier spring increase western u.s. forest wildfire activity. *Science*, 313(5789):940–943, 2006.

[20] J.-F. Lamarque, T. C. Bond, V. Eyring, C. Granier, A. Heil, Z. Klimont, D. Lee, C. Liousse, A. Mieville, B. Owen, M. G. Schultz, D. Shindell, S. J. Smith, E. Stehfest, J. Van Aardenne, O. R. Cooper, M. Kainuma, N. Mahowald, J. R. McConnell, V. Naik, K. Riahi, and D. P. van Vuuren. Historical (1850–2000) gridded anthropogenic and biomass burning emissions of reactive gases and aerosols: methodology and application. *Atmos. Chem. Phys.*, 10(15):7017–7039, Aug 2010.

[21] Claire Granier, Bertrand Bessagnet, Tami Bond, Ariela D'Angiola, Hugo Denier van der Gon, Gregory J. Frost, Angelika Heil, Johannes W. Kaiser, Stefan Kinne, Zbigniew Klimont, Silvia Kloster, Jean-François Lamarque, Catherine Liousse, Toshihiko Masui, Frederik Meleux, Aude Mieville, Toshimasa Ohara, Jean-Christophe Raut, Keywan Riahi, Martin G. Schultz, Steven J. Smith, Allison Thompson, John van Aardenne, Guido R. van der Werf, and Detlef P. van Vuuren. Evolution of anthropogenic and biomass burning emissions of air pollutants at global and regional scales during the 1980–2010 period. *Climatic Change*, 109(1):163, Aug 2011.

[22] Manish Shrivastava, Christopher D. Cappa, Jiwen Fan, Allen H. Goldstein, Alex B. Guenther, Jose L. Jimenez, Chongai Kuang, Alexander Laskin, Scot T. Martin, Nga Lee Ng, Tuukka Petaja, Jeffrey R. Pierce, Philip J. Rasch, Pontus Roldin, John H. Seinfeld, John Shilling, James N. Smith, Joel A. Thornton, Rainer Volkamer, Jian Wang, Douglas R. Worsnop, Rahul A. Zaveri, Alla Zelenyuk, and Qi Zhang. Recent advances in understanding secondary organic aerosol: Implications for global climate forcing. *Reviews of Geophysics*, 55(2):509–559, Jun 2017.

[23] A. L. Hodshire, E. Ramnarine, A. Akherati, M. L. Alvarado, D. K. Farmer, S. H. Jathar, S. M. Kreidenweis, C. R. Lonsdale, T. B. Onasch, S. R. Springston, J. Wang, Y. Wang, L. I. Kleinman, A. J. Sedlacek III, and J. R. Pierce. Dilution impacts on smoke aging: evidence in biomass burning observation project (bbop) data. *Atmos. Chem. Phys.*, 21(9):6839–6855, May 2021.

[24] A. T. Ahern, E. S. Robinson, D. S. Tkacik, R. Saleh, L. E. Hatch, K. C. Barsanti, C. E. Stockwell, R. J. Yokelson, A. A. Presto, A. L. Robinson, R. C. Sullivan, and N. M. Donahue. Production of secondary organic aerosol during aging of biomass burning smoke from fresh fuels and its relationship to voc precursors. *Journal of Geophysical Research: Atmospheres*, 124(6):3583–3606, 2019.

[25] Emily A. Bruns, Imad El Haddad, Jay G. Slowik, Dogushan Kilic, Felix Klein, Urs Baltensperger, and André S. H. Prévôt. Identification of significant precursor gases of secondary organic aerosols from residential wood combustion. *Scientific Reports*, 6(1):27881, Jun 2016.

[26] G. Stefenelli, J. Jiang, A. Bertrand, E. A. Bruns, S. M. Pieber, U. Baltensperger, N. Marchand, S. Aksoyoglu, A. S. H. Prévôt, J. G. Slowik, and I. El Haddad. Secondary organic aerosol formation from smoldering and flaming combustion of biomass: a box model parametrization based on volatility basis set. *Atmos. Chem. Phys.*, 19(17):11461–11484, Sep 2019.

[27] C. Y. Lim, D. H. Hagan, M. M. Coggon, A. R. Koss, K. Sekimoto, J. de Gouw, C. Warneke, C. D. Cappa, and J. H. Kroll. Secondary organic aerosol formation from the laboratory oxidation of biomass burning emissions. *Atmospheric Chemistry and Physics*, 19(19):12797–12809, 2019.

[28] Hervé Abdi and Lynne J. Williams. Principal component analysis. *WIREs Computational Statistics*, 2(4):433–459, 2010.

[29] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, Nov 2009.

[30] Jake Lever, Martin Krzywinski, and Naomi Altman. Principal component analysis. *Nature Methods*, 14(7):641–642, Jul 2017.

[31] J.L. Schafer. *Analysis of Incomplete Multivariate Data.* Chapman and Hall/CRC, 1997.

[32] Paul T. Von Hippel. How to impute interactions, squares, and other transformed variables. *Sociological Methodology*, 39(1):265–291, Aug 2009.

[33] Paul T. von Hippel. Should a normal imputation model be modified to impute skewed variables? *Sociological Methods & Research*, 42(1):105–138, Feb 2013.

[34] Ian R. White, Patrick Royston, and Angela M. Wood. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, 30(4):377–399, Feb 2011.

[35] Leigh Metcalf and William Casey. *Chapter 4 - Introduction to data analysis*, pages 43–65. Syngress, Boston, Jan 2016.

[36] Dalwinder Singh and Birmohan Singh. Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, 97:105524, Dec 2020.

[37] B. M. Bolstad, R. A. Irizarry, M. Åstrand, and T. P. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2):185–193, Jan 2003.

[38] Trevor Hastie, Tibshirani, and Jerome Robert Friedman. *The Elements of Statistical Learning*. Springer, 2009.

[39] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, Jan 2014.

[40] B. M. King. *Analysis of Variance*, pages 32–36. Elsevier, Oxford, Jan 2010.

[41] Inderjit S. Dhillon and Suvrit Sra. Generalized nonnegative matrix approximations with bregman divergences. In *Neural Information Processing Systems (NIPS)*, dec 2005.

[42] Tenenbaum Joshua B., Silva Vin de, and Langford John C. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, Dec 2000.

[43] Herman O. A. Wold. Path models with latent variables: The nipals approach. 1975.

[44] Gene H. Golub and Charles F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, USA, 1996.

[45] M. Martens Harald Martens. *Multivariate Analysis of Quality: An Introduction*. John Wiley Sons, USA, 2001.

[46] I.T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

[47] Anil K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010. Award winning papers from the 19th International Conference on Pattern Recognition (ICPR).

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[49] Katty X Wan, Ilan Vidavsky, and Michael L Gross. Comparing similar spectra: from similarity index to spectral contrast angle. *J Am Soc Mass Spectrom*, 13(1):85–88, January 2002.

[50] Koichiro Aikawa, Takayuki Sakurai, and Jeff J. Jetter. Development of a predictive model for gasoline vehicle particulate matter emissions. *SAE International Journal of Fuels and Lubricants*, 3(2):610–622, oct 2010.

[51] A. K. Kochanski, E. R. Pardyjak, R. Stoll, A. Gowardhan, M. J. Brown, and W. J. Steenburgh. One-way coupling of the wrf–quic urban dispersion modeling system. *Journal of Applied Meteorology and Climatology*, 54(10):2119 – 2139, 2015.

[52] Radenko Pavlovic, Jack Chen, Kerry Anderson, Michael D. Moran, Paul-André Beaulieu, Didier Davignon, and Sophie Cousineau. The firework air quality forecast system with near-real-time biomass burning emissions: Recent developments and evaluation of performance for the 2015 north american wildfire season. *Journal of the Air & Waste Management Association*, 66(9):819–841, 2016. PMID: 26934496.

[53] J. Chen, K. Anderson, R. Pavlovic, M. D. Moran, P. Englefield, D. K. Thompson, R. Munoz-Alpizar, and H. Landry. The firework v2.0 air quality forecast system with biomass burning emissions from the canadian forest fire emissions prediction system v2.03. *Geoscientific Model Development*, 12(7):3283–3310, 2019.

[54] Susan Prichard, N. Sim Larkin, Roger Ottmar, Nancy H.F. French, Kirk Baker, Tim Brown, Craig Clements, Matt Dickinson, Andrew Hudak, Adam Kochanski, Rod Linn, Yongqiang Liu, Brian Potter, William Mell, Danielle Tanzer, Shawn Urbanski, and Adam Watts. The fire and smoke model evaluation experiment—a plan for integrated, large fire–atmosphere field campaigns. *Atmosphere*, 10(2), 2019.

[55] Jon G. Goode, Robert J. Yokelson, Darold E. Ward, Ronald A. Susott, Ronald E. Babbitt, Mary Ann Davies, and Wei Min Hao. Measurements of excess o3, co2, co, ch4, c2h4, c2h2, hcn, no, nh3, hcooh, ch3cooh, hcho, and ch3oh in 1997 alaskan biomass burning plumes by airborne fourier transform infrared spectroscopy (aftir). *Journal of Geophysical Research: Atmospheres*, 105(D17):22147–22166, Sep 2000.

[56] S. P. Urbanski. Combustion efficiency and emission factors for wildfire-season fires in mixed conifer forests of the northern rocky mountains, us. *Atmos. Chem. Phys.*, 13(14):7241–7262, Jul 2013.

[57] Xiaoxi Liu, L. Gregory Huey, Robert J. Yokelson, Vanessa Selimovic, Isobel J. Simpson, Markus Müller, Jose L. Jimenez, Pedro Campuzano-Jost, Andreas J. Beyersdorf, Donald R. Blake, Zachary Butterfield, Yonghoon Choi, John D. Crounse, Douglas A. Day, Glenn S. Diskin, Manvendra K. Dubey, Edward Fortner, Thomas F. Hanisco, Weiwei Hu, Laura E. King, Lawrence Kleinman, Simone Meinardi, Tomas Mikoviny, Timothy B. Onasch, Brett B. Palm, Jeff Peischl, Ilana B. Pollack, Thomas B. Ryerson, Glen W. Sachse, Arthur J. Sedlacek, John E. Shilling, Stephen Springston, Jason M. St. Clair, David J. Tanner, Alexander P. Teng, Paul O. Wennberg, Armin Wisthaler, and Glenn M. Wolfe. Airborne measurements of western u.s. wildfire emissions: Comparison with prescribed burning and air quality implications. *Journal of Geophysical Research: Atmospheres*, 122(11):6108–6129, Jun 2017.

[58] C. E. Stockwell, R. J. Yokelson, S. M. Kreidenweis, A. L. Robinson, P. J. DeMott, R. C. Sullivan, J. Reardon, K. C. Ryan, D. W. T. Griffith, and L. Stevens. Trace gas emissions from combustion of peat, crop residue, domestic biofuels, grasses, and other fuels: configuration and fourier transform infrared (ftir) component of the fourth fire lab at missoula experiment (flame-4). *Atmospheric Chemistry and Physics*, 14(18):9727–9754, 2014.

[59] C. E. Stockwell, P. R. Veres, J. Williams, and R. J. Yokelson. Characterization of biomass burning emissions from cooking fires, peat, crop residue, and other fuels with high-resolution proton-transfer-reaction time-of-flight mass spectrometry. *Atmospheric Chemistry and Physics*, 15(2):845–865, 2015.

[60] A. R. Koss, K. Sekimoto, J. B. Gilman, V. Selimovic, M. M. Coggon, K. J. Zarzana, B. Yuan, B. M. Lerner, S. S. Brown, J. L. Jimenez, J. Krechmer, J. M. Roberts,

C. Warneke, R. J. Yokelson, and J. de Gouw. Non-methane organic gas emissions from biomass burning: identification, quantification, and emission factors from ptr-tof during the firex 2016 laboratory experiment. *Atmos. Chem. Phys.*, 18(5):3299–3319, Mar 2018.

[61] K. Sekimoto, A. R. Koss, J. B. Gilman, V. Selimovic, M. M. Coggon, K. J. Zarzana, B. Yuan, B. M. Lerner, S. S. Brown, C. Warneke, R. J. Yokelson, J. M. Roberts, and J. de Gouw. High- and low-temperature pyrolysis profiles describe volatile organic compound emissions from western us wildfire fuels. *Atmos. Chem. Phys.*, 18(13):9263–9281, Jul 2018.

[62] Susan J. Prichard, Susan M. O'Neill, Paige Eagle, Anne G. Andreu, Brian Drye, Joel Dubowy, Shawn Urbanski, and Tara M. Strand. Wildland fire emission factors in north america: synthesis of existing data, measurement needs and management applications. *International Journal of Wildland Fire*, 29(2):132–147, 2020.

[63] Wade Permar, Qian Wang, Vanessa Selimovic, Catherine Wielgasz, Robert J. Yokelson, Rebecca S. Hornbrook, Alan J. Hills, Eric C. Apel, I-Ting Ku, Yong Zhou, Barkley C. Sive, Amy P. Sullivan, Jeffrey L. Collett Jr, Teresa L. Campos, Brett B. Palm, Qiaoyun Peng, Joel A. Thornton, Lauren A. Garofalo, Delphine K. Farmer, Sonia M. Kreidenweis, Ezra J. T. Levin, Paul J. DeMott, Frank Flocke, Emily V. Fischer, and Lu Hu. Emissions of trace organic gases from western u.s. wildfires based on we-can aircraft measurements. *Journal of Geophysical Research: Atmospheres*, 126(11):e2020JD033838, Jun 2021.

[64] Shawn P. Urbanski, Wei Min Hao, and Stephen Baker. *Chapter 4 Chemical Composition of Wildland Fire Emissions*, volume 8 of *Wildland Fires and Air Pollution*, pages 79–107. Elsevier, Jan 2008.

[65] I. J. Simpson, S. K. Akagi, B. Barletta, N. J. Blake, Y. Choi, G. S. Diskin, A. Fried, H. E. Fuelberg, S. Meinardi, F. S. Rowland, S. A. Vay, A. J. Weinheimer, P. O. Wennberg, P. Wiebring, A. Wisthaler, M. Yang, R. J. Yokelson, and D. R. Blake. Boreal forest fire emissions in fresh canadian smoke plumes: C-1-c-10 volatile organic compounds (vocs), co2, co, no2, no, hcn and ch3cn. *ATMOSPHERIC CHEMISTRY AND PHYSICS*, 11(13):6445–6463, 2011.

[66] Shawn Urbanski. Wildland fire emissions, carbon, and climate: Emission factors. *Forest Ecology and Management*, 317:51–60, Apr 2014.

[67] A. L. Holder, B. K. Gullett, S. P. Urbanski, R. Elleman, S. O'Neill, D. Tabor, W. Mitchell, and K. R. Baker. Emissions from prescribed burning of agricultural fields in the pacific northwest. *Atmospheric Environment*, 166:22–33, Oct 2017.

[68] M. O. Andreae. Emission of trace gases and aerosols from biomass burning - an updated assessment. *ATMOSPHERIC CHEMISTRY AND PHYSICS*, 19(13):8523–8546, July 2019.

[69] R. J. Yokelson, I. R. Burling, J. B. Gilman, C. Warneke, C. E. Stockwell, J. de Gouw, S. K. Akagi, S. P. Urbanski, P. Veres, J. M. Roberts, W. C. Kuster, J. Reardon, D. W. T. Griffith, T. J. Johnson, S. Hosseini, J. W. Miller, D. R. Cocker III, H. Jung, and D. R. Weise. Coupling field and laboratory measurements to estimate the emission factors of identified and unidentified trace gases for prescribed fires. *Atmos. Chem. Phys.*, 13(1):89–116, Jan 2013.

[70] L. E. Hatch, R. J. Yokelson, C. E. Stockwell, P. R. Veres, I. J. Simpson, D. R. Blake, J. J. Orlando, and K. C. Barsanti. Multi-instrument comparison and compilation of non-methane organic gas emissions from biomass burning and implications for smoke-derived secondary organic aerosol precursors. *Atmospheric Chemistry and Physics*, 17(2):1471–1489, 2017.

[71] Robert E. Keane and Duncan Lutes. *First-Order Fire Effects Model (FOFEM)*, pages 1–5. Springer International Publishing, Cham, 2018.

[72] Roger Ottmar. Consume 3.0—a software tool for computing fuel consumption. *Fire Science Brief*, page 6, June 2009.

[73] Matthew C. Reeves, Kevin C. Ryan, Matthew G. Rollins, and Thomas G. Thompson. Spatial fuel data products of the landfire project. *International Journal of Wildland Fire*, 18(3):250–267, 2009.

[74] J. E. Vogelmann, J. R. Kost, B. Tolk, S. Howard, K. Short, X. Chen, C. Huang, K. Pabst, and M. G. Rollins. Monitoring landscape change for landfire using multi-temporal satellite imagery and ancillary data. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 4(2):252–264, 2011.

[75] Kurtis J. Nelson, Joel Connot, Birgit Peterson, and Charley Martin. The landfire refresh strategy: Updating the national dataset. *Fire Ecology*, 9(2):80–101, Aug 2013.

[76] Jakob Lindaas, Ilana B. Pollack, Lauren A. Garofalo, Matson A. Pothier, Delphine K. Farmer, Sonia M. Kreidenweis, Teresa L. Campos, Frank Flocke, Andrew J. Weinheimer, Denise D. Montzka, Geoffrey S. Tyndall, Brett B. Palm, Qiaoyun Peng, Joel A. Thornton, Wade Permar, Catherine Wielgasz, Lu Hu, Roger D. Ottmar, Joseph C. Restaino, Andrew T. Hudak, I-Ting Ku, Yong Zhou, Barkley C. Sive, Amy Sullivan, Jeffrey L. Collett Jr, and Emily V. Fischer. Emissions of reactive nitrogen from western u.s. wildfires during summer 2018. *Journal of Geophysical Research: Atmospheres*, 126(2):e2020JD032657, Jan 2021.

[77] Xin Wan, Kimitaka Kawamura, Kirpa Ram, Shichang Kang, Mark Loewen, Shaopeng Gao, Guangming Wu, Pingqing Fu, Yanlin Zhang, Hemraj Bhattarai, and Zhiyuan Cong. Aromatic acids as biomass-burning tracers in atmospheric aerosols and ice cores: A review. *Environmental Pollution*, 247:216–228, Apr 2019.

[78] Bernd R.T Simoneit. Biomass burning — a review of organic tracers for smoke from incomplete combustion. *Applied Geochemistry*, 17(3):129–162, Mar 2002.

[79] Roberta Zangrando, Elena Barbaro, Piero Zennaro, Silvia Rossi, Natalie M. Kehrwald, Jacopo Gabrieli, Carlo Barbante, and Andrea Gambaro. Molecular markers of biomass burning in arctic aerosols. *Environmental Science & Technology*, 47(15):8565–8574, Aug 2013.

[80] Pingqing Fu, Kimitaka Kawamura, and Leonard A. Barrie. Photochemical and other sources of organic compounds in the canadian high arctic aerosol pollution during winterspring. *Environmental Science & Technology*, 43(2):286–292, Jan 2009.

[81] Ying Zhang, Shaofei Kong, Jiujiang Sheng, Delong Zhao, Deping Ding, Liquan Yao, Huang Zheng, Jian Wu, Yi Cheng, Qin Yan, Zhenzhen Niu, Shurui Zheng, Fangqi Wu, Yingying Yan, Dantong Liu, and Shihua Qi. Real-time emission and stage-dependent emission factors/ratios of specific volatile organic compounds from residential biomass combustion in china. *Atmospheric Research*, 248:105189, Jan 2021.

[82] Coty N. Jen, Yutong Liang, Lindsay E. Hatch, Nathan M. Kreisberg, Christos Stamatis, Kasper Kristensen, John J. Battles, Scott L. Stephens, Robert A. York, Kelley C. Barsanti, and Allen H. Goldstein. High hydroquinone emissions from burning manzanita. *Environmental Science & Technology Letters*, 5(6):309–314, 2018.

[83] Juliane Elisa Welke, Vitor Manfroi, Mauro Zanus, Marcelo Lazzarotto, and Cláudia Alcaraz Zini. Differentiation of wines according to grape variety using multivariate analysis of comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometric detection data. *Food Chemistry*, 141(4):3897–3905, 2013.

[84] Angelika Ziółkowska, Erwin Wasowicz, and Henryk H. Jeleń. Differentiation of wines according to grape variety and geographical origin based on volatiles profiling using spme-ms and spme-gc/ms methods. *Food Chemistry*, 213:714–720, 2016.

[85] Kevin J Johnson and Robert E Synovec. Pattern recognition of jet fuels: comprehensive gc×gc with anova-based feature selection and principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 60(1):225–237, 2002. Fourth International Conference on Environ metrics and Chemometrics held in Las Vegas, NV, USA, 18-20 September 2000.

[86] L. E. Hatch, W. Luo, J. F. Pankow, R. J. Yokelson, C. E. Stockwell, and K. C. Barsanti. Identification and quantification of gaseous organic compounds emitted from biomass burning using two-dimensional gas chromatography–time-of-flight mass spectrometry. *Atmos. Chem. Phys.*, 15(4):1865–1899, Feb 2015.

[87] V. Selimovic, R. J. Yokelson, C. Warneke, J. M. Roberts, J. de Gouw, J. Reardon, and D. W. T. Griffith. Aerosol optical properties and trace gas emissions by pax and op-ftir for laboratory-simulated western us wildfires during firex. *Atmos. Chem. Phys.*, 18(4):2929–2948, Mar 2018.

[88] R. J. Yokelson, J. G. Goode, D. E. Ward, R. A. Susott, R. E. Babbitt, D. D. Wade, I. Bertschi, D. W. T. Griffith, and W. M. Hao. Emissions of formaldehyde, acetic acid, methanol, and other trace gases from biomass fires in north carolina measured

by airborne fourier transform infrared spectroscopy. *Journal of Geophysical Research: Atmospheres*, 104(D23):30109–30125, 1999.

[89] L. E. Hatch, A. Rivas-Ubach, C. N. Jen, M. Lipton, A. H. Goldstein, and K. C. Barsanti. Measurements of i/svocs in biomass-burning smoke using solid-phase extraction disks and two-dimensional gas chromatography. *Atmos. Chem. Phys.*, 18(24):17801–17817, Dec 2018.

[90] Julieta F. Juncosa Calahorrano, Jakob Lindaas, Katelyn O'Dell, Brett B. Palm, Qiaoyun Peng, Frank Flocke, Ilana B. Pollack, Lauren A. Garofalo, Delphine K. Farmer, Jeffrey R. Pierce, Jeffrey L. Collett Jr., Andrew Weinheimer, Teresa Campos, Rebecca S. Hornbrook, Samuel R. Hall, Kirk Ullmann, Matson A. Pothier, Eric C. Apel, Wade Permar, Lu Hu, Alan J. Hills, Deedee Montzka, Geoff Tyndall, Joel A. Thornton, and Emily V. Fischer. Daytime oxidized reactive nitrogen partitioning in western u.s. wildfire smoke plumes. *Journal of Geophysical Research: Atmospheres*, 126(4):e2020JD033484, Feb 2021.

[91] Lauren A. Garofalo, Matson A. Pothier, Ezra J. T. Levin, Teresa Campos, Sonia M. Kreidenweis, and Delphine K. Farmer. Emission and evolution of submicron organic aerosol in smoke from wildfires in the western united states. *ACS Earth and Space Chemistry*, 3(7):1237–1247, Jul 2019.

[92] Qiaoyun Peng, Brett B. Palm, Kira E. Melander, Ben H. Lee, Samuel R. Hall, Kirk Ullmann, Teresa Campos, Andrew J. Weinheimer, Eric C. Apel, Rebecca S. Hornbrook, Alan J. Hills, Denise D. Montzka, Frank Flocke, Lu Hu, Wade Permar, Catherine Wielgasz, Jakob Lindaas, Ilana B. Pollack, Emily V. Fischer, Timothy H. Bertram, and Joel A. Thornton. Hono emissions from western u.s. wildfires provide dominant radical source in fresh wildfire smoke. *Environmental Science & Technology*, 54(10):5954–5963, May 2020.

[93] Z. C. J. Decker, M. A. Robinson, K. C. Barsanti, I. Bourgeois, M. M. Coggon, J. P. DiGangi, G. S. Diskin, F. M. Flocke, A. Franchin, C. D. Fredrickson, G. I. Gkatzelis, S. R. Hall, H. Halliday, C. D. Holmes, L. G. Huey, Y. R. Lee, J. Lindaas, A. M. Middlebrook, D. D. Montzka, R. Moore, J. A. Neuman, J. B. Nowak, B. B. Palm, J. Peischl, F. Piel, P. S. Rickly, A. W. Rollins, T. B. Ryerson, R. H. Schwantes, K. Sekimoto, L. Thornhill, J. A. Thornton, G. S. Tyndall, K. Ullmann, P. Van Rooy, P. R. Veres, C. Warneke, R. A. Washenfelder, A. J. Weinheimer, E. Wiggins, E. Winstead, A. Wisthaler, C. Womack, and S. S. Brown. Nighttime and daytime dark oxidation chemistry in wildfire plumes: an observation and model analysis of firex-aq aircraft data. *Atmos. Chem. Phys.*, 21(21):16293–16317, Nov 2021.

[94] WE-CAN. Western wildfire experiment for cloud chemistry, aerosol absorption and nitrogen.

[95] FIREX-AQ. Fire influence on regional to global environments and air quality.

[96] Eric Apel. Trace organic gas analyzer.

[97] Ali Akherati, Yicong He, Matthew M. Coggon, Abigail R. Koss, Anna L. Hodshire, Kanako Sekimoto, Carsten Warneke, Joost de Gouw, Lindsay Yee, John H. Seinfeld, Timothy B. Onasch, Scott C. Herndon, Walter B. Knighton, Christopher D. Cappa, Michael J. Kleeman, Christopher Y. Lim, Jesse H. Kroll, Jeffrey R. Pierce, and Shantanu H. Jathar. Oxygenated aromatic compounds are important precursors of secondary organic aerosol in biomass-burning emissions. *Environmental Science & Technology*, 54(14):8568–8579, Jul 2020.