Sensing Spaces: Machine Learning Methods for Characterising Airborne Particulate Pollution in Indoor-Outdoor Micro-Environments

John Palmer

Master of Science

Data Science

School of Informatics

University of Edinburgh

2020

Abstract

Air pollution levels in urban centres are a worldwide environmental and health concern. This work has investigated machine learning techniques to analyse airborne particulate data in a selection of indoor-outdoor micro-environments. A novel source apportionment method has been developed based on the particle size distributions data produced by the Airspeck monitors for classifying dominant airborne particulate sources in a mixture of selected pollution sources. A series of classifiers were trained on various data sets which consisted of both individual pollution source measurements as well as data sets created through taking a linear combination of either the individual pollution source bin counts or particle size distributions to form mixtures of pollution sources. Data sets generated by approximating the individual pollution sources as MVG random variables were also investigated. It was found that by creating a data set consisting of mixtures of the individual pollution sources using either the particle counts or particles size distributions that the overall classification performance could be improved. The second major contribution is an indoor-outdoor classifier based on the personal Airspeck data which learns and adapts to unseen micro-environments. At its core is a variation of the co-training algorithm which exploits the rich variety of features in the location, environment and air quality measurements. The indoor-outdoor classifier was integrated into the source apportionment system, which improved its F1 score in nearly all the cases by using the indoor-outdoor label as an additional feature. The third contribution was an investigation into the prediction of elemental composition of the airborne pollution from the Airspeck particle size distribution information for a selected micro-environment (road-side traffic junctions) in two locations in London. Given the limited number of leaf samples available for examining traffic related deposits in a Scanning Electron Microscope (SEM), machine learning methods were no better than baseline methods of using the averages of training elemental composition measurements.

Acknowledgements

I would like to give my thanks to my supervisor Professor D. K. Arvind for his amazing guidance and support during this project, as well as the other members of the The Centre of Speckled Computing for their support whenever it was needed.

I would also like to thank Professor Alexandra Porter, Dr. Katharina Marquardt, Dr Michal Klosowski, and Marta Chiapasco. Imperial College London, for providing SEM data on the air pollution deposits on the leaf samples.

Table of Contents

Chapter 1

Introduction

Airborne pollutants have been linked to environmental effects such as acid rain [10], ozone depletion [32] and Eutrophication [24] as well as health conditions in humans such as cardiovascular diseases [29][20], various cancers [17] and even skin diseases [3]. It is apparent that being able to monitor and understand air pollution levels is of vital importance. Airspeck [4][5] is a set of devices which can be used to monitor various features relating to the local air quality in the vicinity of the sensor. Features such as the Particulate Matter (PM) under a certain size, particle size fractions, temperature and Relative Humidity (RH) and GPS can all be measured. The Airspeck-P measures all of the above while being portable and the Airspeck-S is a static sensor which can additionally measure other gas information such as NO_2 and O_3 concentrations.

The micro environments that we interface with on a daily basis heavily influence the pollution sources that the average person is exposed to. The pollution sources that a person may be exposed to while at home will likely be very different to those that a person may be exposed to while standing at a road junction with a heavy traffic flow. Understanding the pollution sources that might be found in various micro-environments is an ongoing open area of research [14][33][21]. This work aims to build machine learning tools which can be used to better understand the pollution sources and levels within various micro environments. The main contributions of this dissertation are:

1. An indoor/outdoor classification system which using the measurements from the Airspeck-P can predict whether a new measurement was taken either indoors or outdoors.

2. A source apportionment system using machine learning methods which can classify the dominant pollution source from a mixture of known pollution sources.

3. An evaluation of the changes in the concentrations of airborne particulates less than 2.5 microns in diameter (PM2.5), nitrogen dioxide and ozone based on data gathered from a network of four Airspeck monitors located in South Kensington as London emerged from the lockdown restrictions due to the COVID-19 pandemic.

4. An investigation of the elemental composition of roadside particulate deposits on leaf samples in the vicinity of the Airspeck monitors and relating them to Particle Size Distribution produced by the Airspeck monitor.

The novel results presented in this dissertation are:

1. Three methods to mix individual pollution source data and the resulting source apportionment classifiers were evaluated using data sets created by mixing the PSDs, raw particle counts and by modelling the PSD as random variables to enable enable classifiers to learn more robust decision boundaries.

2. The second novel contribution of this work was to develop and transfer state of the art semi-supervised learning techniques from the mobile phone domain to the Airspeck-P to evaluate whether the air quality measurements can be used to determine if a device is either indoors or outdoors and more importantly learn from new measurements in previously unseen micro environments.

3. The indoor/outdoor classification system was integrated into the source apportionment system and shown to increase the overall performance over using only the particle size information.

This report is split into four main sections. Chapter 2 will summarise any related work, as well as the data and machine learning algorithms used. Chapter 3 summarises the methodologies for how the data was pre-processed, while chapter 4 summarises the experiments conducted and evaluation metric used. Chapter 5 presents the results with a detailed analysis. Finally, Chapter 6 summarises the findings and presents potentially fruitful avenues of further research.

Chapter 2

Background

2.1 Related Work

Indoor/outdoor detection is an open research task with relevance in many areas. In particular with mobile phones due to the multitude of available measurements [34][28]. Previous work investigated using the GPS signal such as Hansen et al. [16] who showed the GPS signal strength is an important feature in determining whether a mobile phone is indoors or outdoors, however, later work has shown that the GPS signal variability and signal to noise ratio [22] is a stronger feature for classification. There are other measurements on mobile devices which can aid in outdoor/indoor classification. Many of these features are too weak on their own to make an accurate classification, but when additional features such as temperature [25][36] are coupled with the GPS features this can increase the classification accuracy.

All of the methods discussed so far have used supervised learning techniques to predict whether a device is outdoors or indoors. These techniques require large quantities of labelled data from numerous sources which cover different locations, times of day and season in order to build classifiers which are robust for use. Another method is to use semi-supervised learning techniques which can use new unlabelled data to improve classification performance and adapt to new environments that the classifier may not have been previously exposed to.

An important piece of work by Radu et al [36] investigated using both supervised and semi-supervised methods for outdoor/indoor classification of a mobile phone. It was shown that based using the available features from the mobile phone it was possible to accurately build a classifier which could predict whether the phone was either indoors or outdoors. It was also shown that certain features such as the GPS strength

and variability, and the cell signal were very strong features for classification. A number of semi-supervised learning algorithms were also tested by training a set of initial classifiers using data from a specific data and then using semi-supervised classification techniques such as cluster-then-label, self training and co-training [8]. It was found that co-training gave the best results and performance could be improved by learning from new unlabelled data. Co-training uses two classifiers to label the unlabelled data, with the final label being selected from the classifier with the highest confidence. Co-training has been shown to be effective in a number of different domains, however it is vital that the two classifiers are trained using features which are conditionally independent [8].

Buzatu [9] built an indoor/outdoor classification system using an Airspeck-P and mobile phone. It was shown that by using GPS features from the phone, as well as particle size information and temperature, relative humidity and light intensity that an accurate indoor classification algorithm could be built. However, it was also observed that for new environments that the generalisation performance of the trained classifiers also degraded indicating that a semi-supervised approach could improve the generalisation performance to new unseen environments.

Source apportionment is the process of determining where a specific pollution source has originated from. Traditionally this is done on a macro scale in order to determine the specific source of the pollution from for example industrial air [26][18][15] and water [7], or from the effects of weather systems [30][42]. Traditionally airborne particulate pollution source apportionment methods involve expensive instrumentation used to measure specific pollutants to enable common techniques like chemical element balances [12][23] or receptor models [6].

In recent years machine learning methods have been employed in a wide variety of source apportionment applications. For example Requia et al [37] used methods such as ordinary kriging and random forests to spatially predict the specific concentrations of $PM_{2.5}$ components. Other work has used methods such as positive matrix factorisation with support vector machines to develops source apportionment techniques for trace elements in rivers [11]. Buzatu [9] used machine learning techniques in conjunction with measurements from the Airspeck-P in order to classify a number of indoor pollution sources with a good level of accuracy. Sanatani [38] extended this further in order to classify multiple pollution sources. In particular binary mixed pollution data was generated by mixing the count data of one pollution source with the mean of another. Although, in both cases there was poor generalisation to new micro environ-

ments with no capacity to adapt.

Traffic particulates from traffic emissions contain many harmful elements which have been shown to be toxic towards humans [27][19][13][31] with the levels of toxicity being dependent on the elemental composition of the particulate matter [35]. However, being able to measure and quantify the toxicity to humans of particulate matter from traffic emissions is still very challenging. Recent work shows that it is possible to measure the compositions of traffic related particulates which have been deposited on nearby vegetation [2].

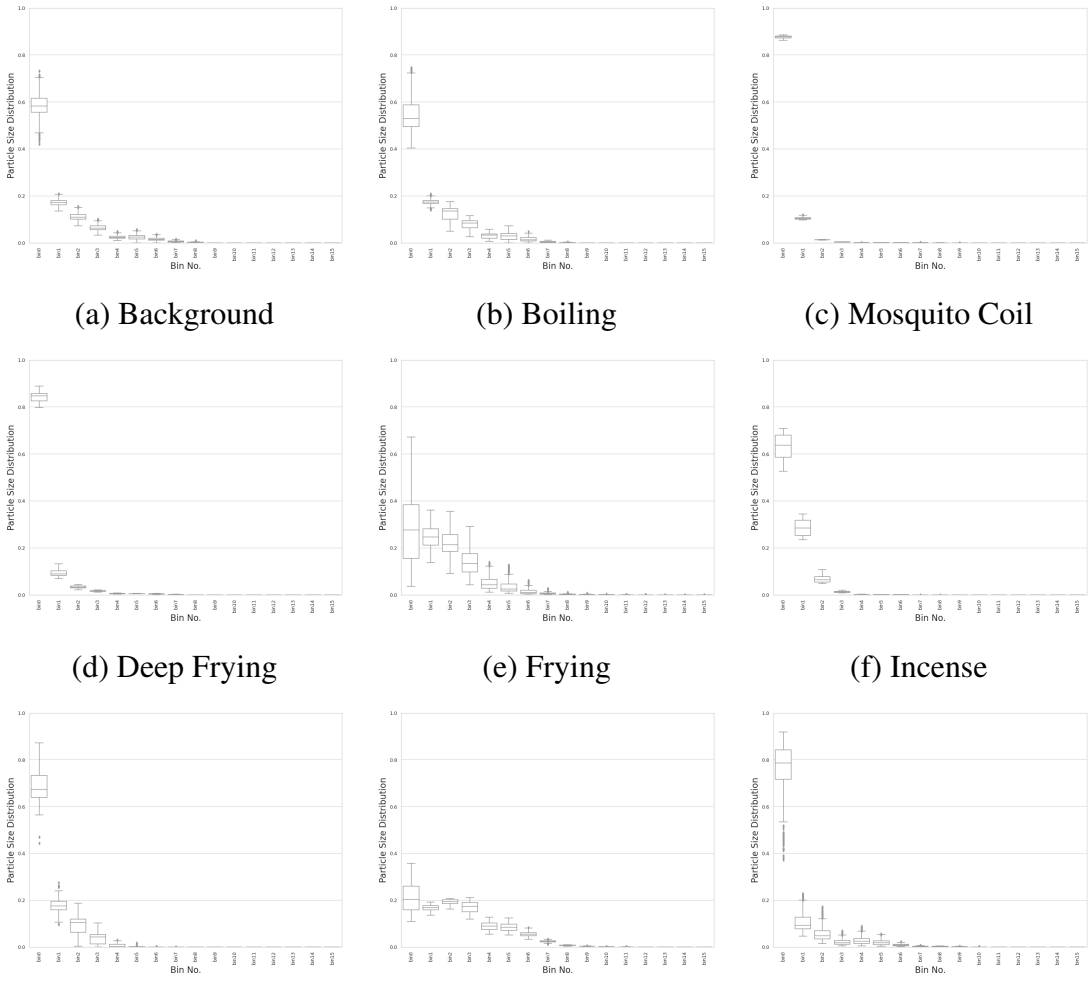
2.2 Data Sources and Exploratory Data Analysis

2.2.1 Source Apportionment

This work classifies nine different pollution sources by fingerprinting the PSDs of the individual pollution sources. The Airspeck devices contain an optical particle size measurement which counts the number of airborne particles in given size ranges and automatically counts them into different size fractions called bins (For the exact bin ranges see Appendix A), which can be converted to the PSD (Sections 3.1.1). A data set consisting of frying, deep-frying, boiling, burning incense, mosquito coil, cigarette smoke and traffic pollution sources and a background class (no major sources of pollution present) was compiled. Everything except the smoking and traffic data used the data sets as specified by Sanatani [38]. The traffic pollution source data was collected from the Airspeck-S sensor D849BF7848210A4A positioned in a high traffic area between the period of the 22^{nd} June to the 4^{th} July 2020 between the hours of 7am and 7pm for which there was expected to be heaviest amounts of traffic. The cigarette smoke data was the data generated by Buzatu [9].

The PSDs of the individual pollution sources can be seen in Figure 2.1. The pollution sources all have distinct PSDs however there are a number of pollution sources with similar PSDs with some overlap observed in the variability of the results. For example, the mosquito coil and deep frying are very similar, while incense and smoking also appear to have similar PSDs. This could cause issues when attempting to classify these pollution sources as they may not be fully separable when attempting to form a decision boundary.

The PSDs of the multi source data can be seen in Figure 2.2. These were previously measured as reported by Sanatani [38]. It is observable that the PSDs from the



(a) Background (b) Boiling (c) Mosquito Coil

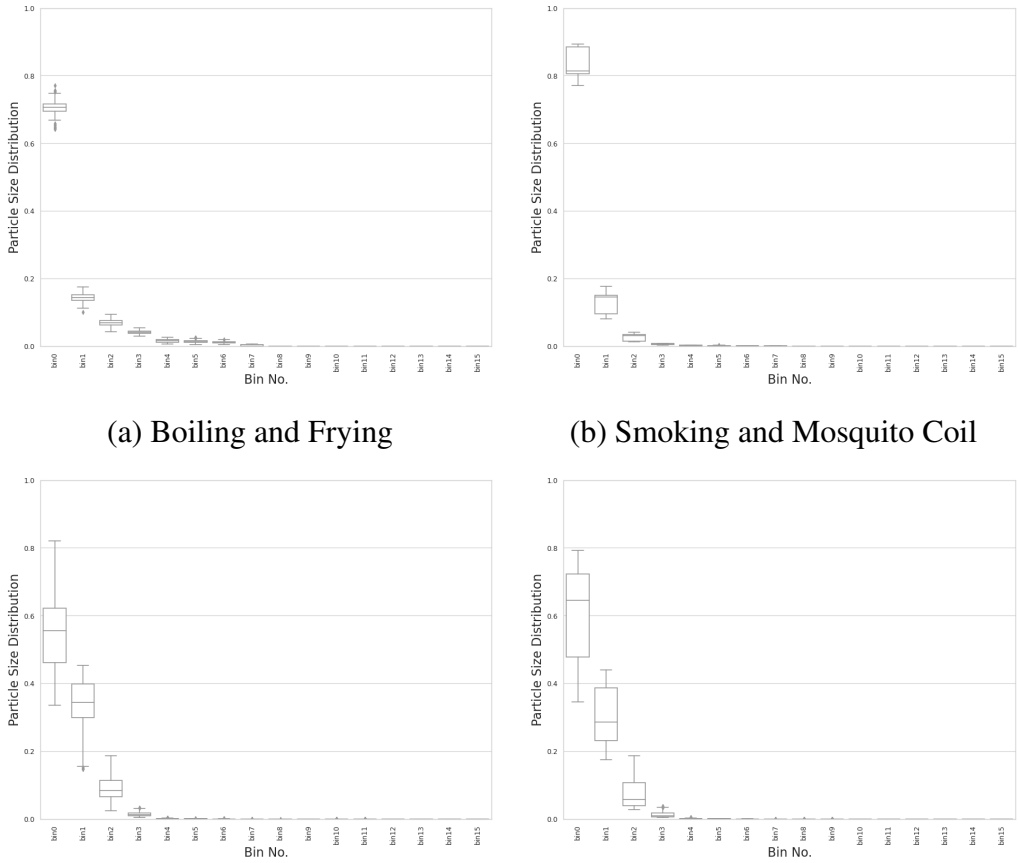
(d) Deep Frying (e) Frying (f) Incense

(g) Smoking (h) Spraying (i) Traffic

Figure 2.1: Box plots showing the measured individual pollution source PSDs

mixed pollution sources are not equal mixtures of the individual sources. Many of the pollution sources produce many more particles than others during the measurement. For example, the measurements from burning incense typically has multiple orders of magnitude more particles in the smaller size fractions than pollution sources such as the mosquito coil.

In order to understand the difference between the individual and multiple pollution sources, an Isomap dimensionality reduction transformation was applied to the PSDs of all the data. The first and second scores of the decomposition can be seen in Figure 2.3a. This decomposition again highlights some of the pollution sources are of a high degree of similarity. The mixed pollution sources which contain incense smoke are very similar to one another and also to the individual incense PSD. Again, this is not surprising as the raw incense pollution sources contains many more particles than other



(a) Boiling and Frying

(b) Smoking and Mosquito Coil

(c) Mosquito Coil and Incense

(d) Smoking and Incense

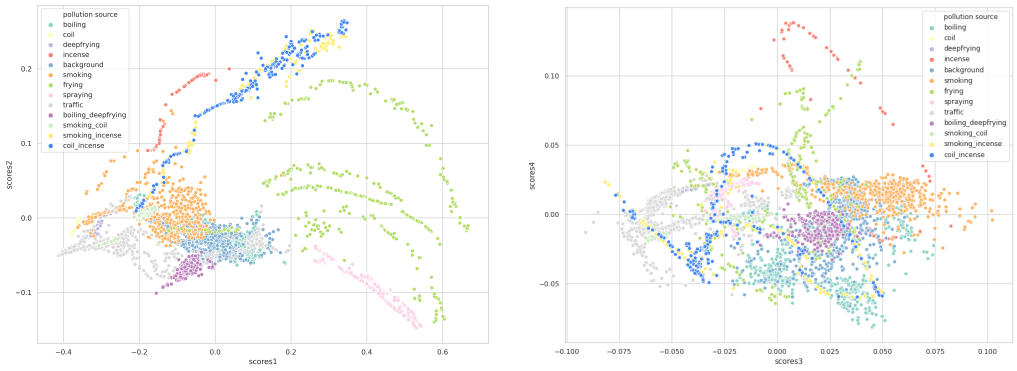
Figure 2.2: Box plots showing the measured mixed pollution source PSDs

pollution sources and therefore will heavily skew the mixed probability distribution towards the individual incense distribution.

Finally, the degree of class imbalance was evaluated. It was found that some of the classes had many more samples than others. The maximum number of samples within a class was then set at 670 by subsampling the data from a particular pollution source with more than the maximum allowable samples. The classes were still imbalanced (between 90 and 670 measurements per class) which was overcome by using class weighting during training and the F1 score for assessment.

2.2.2 Indoor/Outdoor classification

The data used in this section was primarily the same as the data used for the source apportionment work (Section 2.2.1). The main difference in this section, was that the measurements were only used if they were collected from the portable Airspeck-P,



(a) (b)

Figure 2.3: Isomap decomposition of the source apportionment data

as the movement of the device as determined by the GPS was thought to be key to understanding whether the device is indoors or outdoors. The measurements consisted of three sets of features: total bin counts and particle size information, temperature and humidity, and GPS related measurements. The indoor data sets consisted of all of the source apportionment data, including the cooking and burning data as well as general indoor data which has been specifically labelled for this purpose. Other outdoor data which was added to the data set included data collected as part of other projects during walks around specific parts of Edinburgh. After combining the data sets the number of samples in each of the indoor and outdoor classes were balanced, consisting of a total of 10266 samples.

2.2.3 Traffic Toxicity Prediction

Air quality data was collected from four static Airspeck-S sensors installed at four junctions in South Kensington, London. Compare to the Airspeck-P, the Airspeck-S additionally measure the concentration gases such as nitrogen dioxide and ozone. These sensors were installed on 15^{th} May and have been used to monitor the pollution levels in London as the economic activity in London ramps back up after the recent lock down events.

Additionally, samples of leaves in the vicinity to the sensors were collected on three specific dates. These leaves were then analysed by the Department of Materials, Imperial College London, using Scanning Electron Microscopy (SEM) by sampling a number of individual particles at random across the leaves' surface. This analysis resulted in a composition of elements for each of the measured particles on the leaves

surface allowing for an average elemental composition for each sample to be determined as shown in Figure 2.4. A number of the dominant elements can be attributed to traffic related pollution. For example, calcium is one of the main elements measurable in the particulate matter from the emissions of vehicular pollution as it is used as a fuel additive [41]. Iron particles are probably as a result of wear and tear of the pads during braking, and the engine parts during combustion, although the particle sizes in the case of the latter would be smaller than those detected in this exercise [41]. The composition of iron in the particles from the Kensington Gore samples are much higher than those at the Christ Church site and this is thought to be because the Kensington Gore site is on a busy junction where a lot of traffic will need to stop.

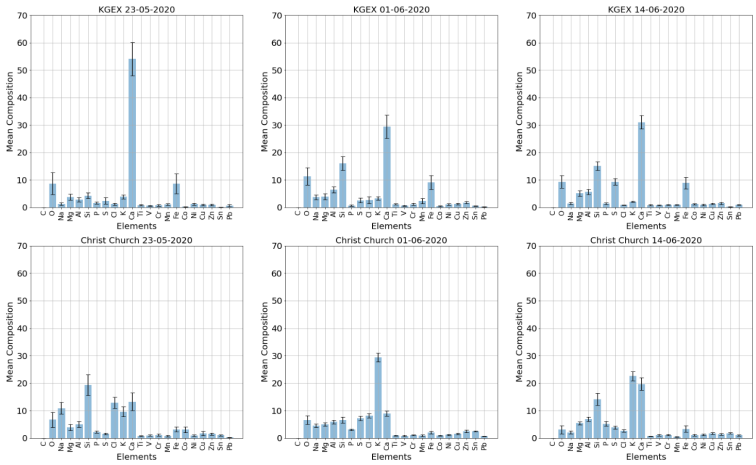
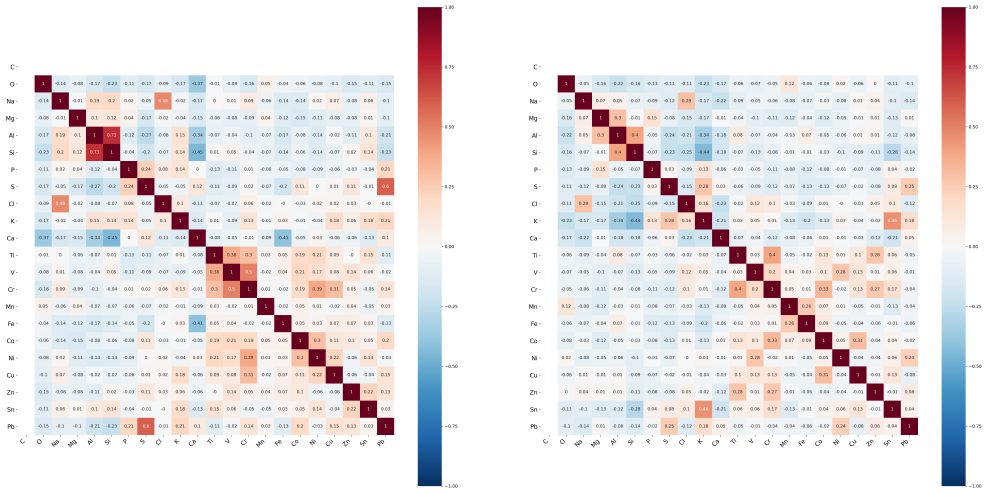


Figure 2.4: Elemental Composition of the Leaf Samples

Typically the elements measured are not normally found in their raw state, but as some kind of compound. A correlation analysis (Figure 2.5) was conducted for all of the data found across the two sites in order to determine if the elemental compositions follow the expected correlations and where the elements found in traffic based particulate emissions form observable correlations. The main correlation which appears in both sites is between aluminium and silicon, which are commonly used in automotive alloys such as Silumin and Alusil in high-wear applications in pistons, and in the linings in cylinders and engine blocks [41]. The other main correlation is between sulphur and lead which are constituents of emission from diesel and petrol engines. The other heavy metal correlations are a bi-product of being such small compositions of the overall mixture which will likely cause them to move in the same direction together when one of the other larger components changes.

The SEM analysis also gives information on particle sizes and their morphology



(a) (b)

Figure 2.5: Correlation Matrix for the measured elemental compositions from the different sites, (a) Kensington Gore, (b) Christ Church

in terms of aspect ratio and surface roughness. The average particle size of the two longest dimensions was used to convert the SEM particle sizes into the PSDs of all the particles within a sample with the same bin limits as the Airspeck devices (See Appendix A for bin ranges). This allowed a direct comparison of the PSDs between the two measurement source as shown in Figure 2.6 and Figure 2.7. It was observed that the PSDs between the two measurements are significantly different. The SEM measurements only considers particles of size greater than $1.5\mu m$ with some particles falling into bin 5; particles of sizes below 95 pixels in the SEM imaging software were removed due to the low signal-to-noise ratio making their identification difficult. The Kensington Gore site PSD is more heavily biased towards finer particles than the Christ Church site data. In comparison the Airspeck-S particle size measurements are much smaller with the majority of the distribution in smaller particle size range.

The final source of data considered was from an application written to scrape traffic data using the 'HERE' traffic API [1], and parsed at 5-minute interval (to align with the frequency of the Airspeck-S data collection). Two features of interest are: the jam factor - a value between zero (free-flowing) and ten (standstill) giving a relative level of traffic on that road; and, the average speed of traffic in Km/hr.

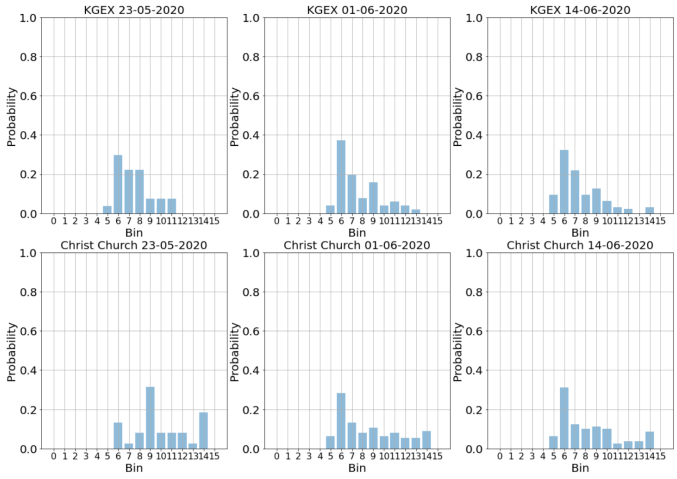


Figure 2.6: PSD for the SEM data

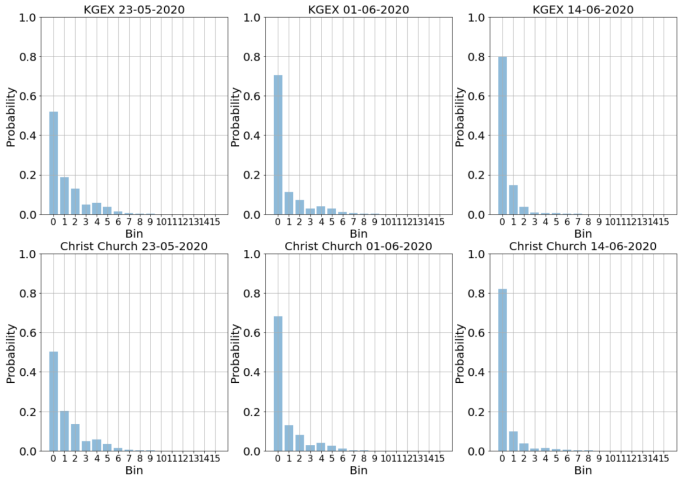


Figure 2.7: PSD for the Airspeck-S data

2.3 Machine Learning Methods

2.3.1 Supervised Machine Learning Methods

A number of supervised classification algorithms are used in this work in order to predict a label from a set of labels that the classifier is trained against. Both the source apportionment and indoor/outdoor classifiers use these algorithms. Logistic regression is a binary classification which uses linear combinations between learnt weights and input features in conjunction with a sigmoid function in order to form a probability distribution used to determine the class. The multiclass version is sometimes also called softmax regression and uses the softmax function to form the probability distribution

across the classes. Naive Bayes classifiers are also used in this work due to their need for only small amounts of training data and their speed. Naive Bayes use the assumption of conditional independence between features and the associated class variable in order to estimate the probability of a class given an associated feature vector. The k-Nearest Neighbours (kNN) algorithm is a non parametric method whereby a sample is classified by the most common class from it's k nearest neighbours. These methods are useful algorithms to try in order to ascertain a baseline level of performance of a classifier, especially methods like k-Nearest neighbours which do not require any parameterisation.

Other more complex and powerful classification algorithms are also tested for use in both source apportionment and indoor/outdoor classification. Methods such as random forest, Support Vector Machines (SVM) and Artificial Neural Networks (ANN) are all used. Random forests use an ensemble of n decision trees with the highest number of a given class from all of the trees forming the final class. Random forests work on on the principle that as the number of individual uncorrelated trees increases the error will be reduced. For this reason random forests are a very powerful classification algorithm which can produce strong results and are less likely to over-fit the data than other methods. The SVM model attempts to classify the data by fitting a decision boundary with the maximum margin between classes. It is also powerful as it can be used with different kernels allowing for both linear and non-linear classification. In this work the linear kernel is used for linear classification while the RBF kernel is used for non-linear classification. The regularisation constant C inversely affects the mount of regularisation applied to prevent overfitting. Finally ANN's are a very powerful and popular type of model. ANN's consist of a number of layers (L), with each layer consisting of a number of affine transformations with non linear activation functions (U). This work uses the common ReLU non-linear activation function due to it's robustness to the vanishing gradient problem which are often encountered with other activation functions such as the tanh and sigmoid functions. This structure allows for very flexible models with the capacity to fit incredibly complex functions which can lead to problems with the models overfitting to the data. In order to stop overfitting an early stopping approach is used, where once the loss on an independent validation data set starts to increase during the training then training is terminated.

Linear regression is used to regress the airborne PSD to the measured elemental composition distribution of the particle deposited on leaf samples. Linear regression uses a linear combination between input features and a learned weight matrix in order

to predict a single or set of continuous output variables. Although linear regression is by its nature a linear method the data can be transformed using methods such as RBF's in order to allow for fitting of non-linear data. In this work RBF's are used by centering an RBF over each data point and the weights are penalised to stop overfitting by using L2 regularisation (also known as ridge regression).

2.3.2 Semi-Supervised Machine Learning Methods

Unlike supervised learning techniques, semi-supervised machine learning methods are typically able to learn from un-labelled data in order to improve the overall performance system and to adapt to new situations. In this work an adaption of the co-training [8] algorithm is used for indoor/outdoors detection based on the recent work by Radu et al. [36] which showed the application of this algorithm to indoor/outdoor detection of mobile phones.

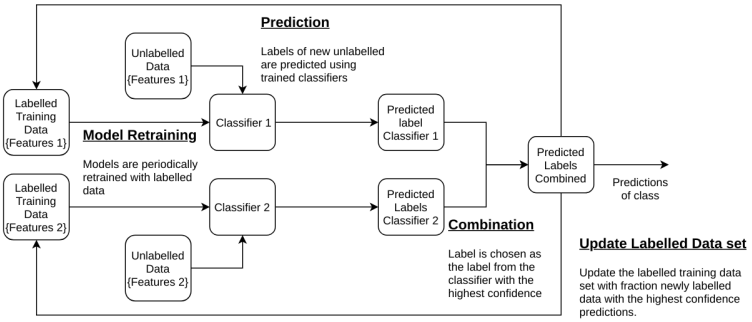


Figure 2.8: Flow diagram showing the co-training algorithm

The flow diagram for the co-training algorithm can be seen in Figure 2.8. It uses two classifiers which are trained on a labelled data set with conditionally independent feature sets being used for each classifier. In this case the classifiers learn to classify the label with different views on the data. As new unlabelled data is collected both of the classifiers within the co-training algorithm are used to predict the class label. The label that is assigned is given by the classifier with the highest confidence for a prediction. A fraction of the initial labelled training set with the lowest confidence is then replaced with the newly labelled data with the highest confidence and both classifiers are retrained periodically. This enables the co-training algorithm to learn and adapt to new environments. Importantly though for this to work, both classifiers need to be able to make accurate predictions from the features available to them.

Chapter 3

Methodology

3.1 Data Preparation

The data used in this project was split into training, validation and test sets as described in the specific methodology sections with the training set being used to train the models, the validation set being used for hyper-parameter tuning and the test set being used to test the final generalisation performance of the final model. In each case it was necessary to apply a number of transformations to the data as described in the following sections.

3.1.1 Bin Count Normalisation

The Airspeck-P measures the counts of particles within an equivalent diameter range over the measurement period resulting in sixteen bins of particle counts (see Appendix A for exact particle size ranges). The values in the sixteen bins were normalised by dividing the counts in each bin (c_k) by the total counts across all the bins using Equation 3.1 in order to remove measurement effect such as distance. This forms the PSD. It is hypothesised that each major pollution source will have a unique PSD.

$$p_k = \frac{c_k}{\sum_{b=1}^{B} c_b} \tag{3.1}$$

3.1.2 Outlier Detection and Removal

Outliers were detected and removed from the training set using Tukey's Fences (k=1.5) as seen in Equation 3.2. The Tukey's Fences outlier method is applied to the required

feature independently for each pollution source. During the deployment of the classifier the pollution source will not be known and therefore outlier detection cannot be used as detailed here. Therefore outlier removal is only applied to any training data in order to remove any measurement artefacts before training any models.

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)] \tag{3.2}$$

3.1.3 Median Filter

A median filter is used to smooth the data and to remove noise. An eleven point sliding window is used for each feature. A median filter replaces the the window with the median value from the values within the window.

3.1.4 Feature Engineering

Two further features are required for this work. The first is the euclidean distance which is calculate from the change in longitude ($p_x^{longitude}$) and latitude ($p_x^{latitude}$) of each measurement (Equation 3.3), and the second is the standard deviation of the euclidean distance using a ten point sliding window. Any measurement included in the buffer before the window is filled with a NaN value, meaning it will be removed from the data set.

$$d(p_1, p_2) = \sqrt{(p_1^{longitude} - p_1^{latitude})^2 + (p_2^{longitude} - p_2^{latitude})^2} \tag{3.3}$$

3.2 Multi-source Apportionment

3.2.1 Data Pre-Processing

Any data points with data with missing or null values in the bin values were are removed as well as any outliers are removed from the data set using Tukeys' Fences (Section 3.1.2). The data was then filtered using a median filter as specified in Section 3.1.3 to smooth the data. Finally, the count data was normalised to a probability distribution as specified by Section 3.1.1. It is important that the normalisation process is applied last in order to avoid invalid probability distributions, i.e where the cumulative area is not equal to unity.

3.2.2 Data Generation and Augmentation

One of the novel contributions of this work is investigating how augmenting the individual pollution source data and generating new data sets which consist of mixtures of the individual components can improve both the accuracy and the robustness of the source apportionment system. Two assumptions have been made about the mixtures formed. The first was that the maximum number of pollution sources in a mixture would be three and the second was that there would be one dominant pollution source which comprises of at least 50% of the overall mixture. Three main methods were investigated:

3.2.2.1 Mixing Pollution Sources using Linear Combinations of the PSD

The first method for producing the PSDs of mixtures of pollution sources used linear combinations of the individual pollution source PSDs. The pseudo code for how the multi-source data was generated can be seen in Appendix D.

3.2.2.2 Mixing Pollution Sources using Linear Combinations of the Raw Count Measurements

This method of generating a data set of mixed pollution source data uses linear combinations of the raw count values in the bins from the individual pollution source measurements before normalisation to the PSD. This method doesn't assume that each individual pollution source will have an equal weighting towards the overall PSD. For example, the measurements from the incense pollution source have orders of magnitude more particles in a given measurement than other pollution sources and therefore will dominate the probability distribution when mixed with other components. This method could be more sensitive to the differences in how the measurements were taken. The pseudo code used to generate this data set can be seen in Appendix D.

3.2.2.3 Modelling the Pollution Sources as Multivariate Gaussian Random Variables

The bins were normalised across all the particles resulting in a probability distribution of the likelihood of a measured particle falling into one of the bins. The PSD for each of the pollution sources was approximated using a MVG (MVG) distribution (Equation 3.4) with both the mean and covariance calculated empirically from the pre-processed data using Equation 3.5 and 3.6. This method assumes that the variability

is centered evenly around the mean. In some cases there is some time based variation which violates this assumption. It was not appropriate to model this variance as a different type of distribution and it is assumed to be negligible to the overall measured distributions.

$$f_{\mathbf{p}}(p_1,...,p_k) = \frac{1}{\sqrt{(2\pi)^k|\Sigma|}} e^{(-\frac{1}{2}(\mathbf{p}-\bar{\mathbf{p}})^T\Sigma^{-1}(\mathbf{p}-\bar{\mathbf{p}}))} \tag{3.4}$$

$$\bar{\mathbf{p}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{p}^{(n)} \tag{3.5}$$

$$cov(p_i,p_j) = \frac{1}{N}\sum_{n=1}^{N}(p_i^{(n)}-\bar{p}_i)(p_j^{(n)}-\bar{p}_j) \tag{3.6}$$

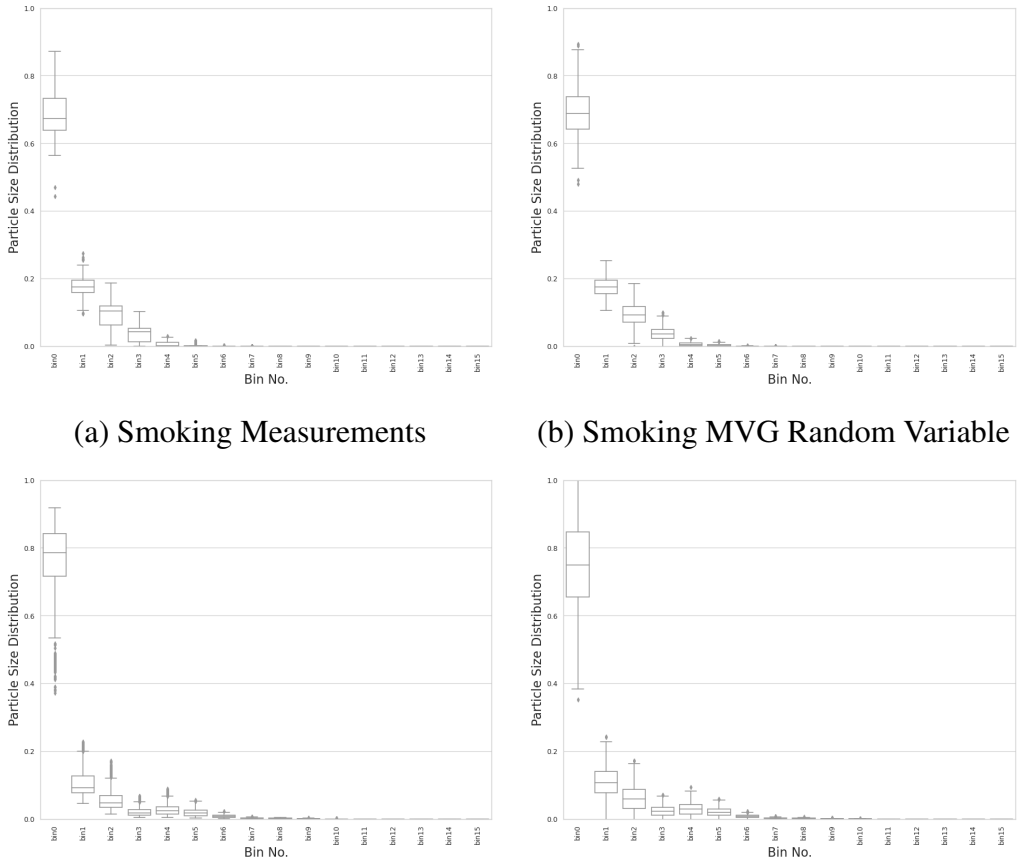
In order to form the mixture, it was assumed that the PSD of a mixture made of N independent single pollution sources is another MVG distribution consisting of linear combinations of the individual distributions in the mixtures. This method provides a simple method for approximating mixed distributions and has the main advantage that if the data can accurately be represented by a random variable with a small number of fully descriptive parameters then storing large amounts of data is not required. The pseudo code used to generate this data set can be seen in Appendix D.

Examples of the single source data distributions for both cigarette smoke and traffic can be seen in Figure 3.1 by randomly sampling 10,000 data points from the distribution. These pollution sources demonstrate both examples where the MVG gives a good representation of the measured data with smoking and then a poor representation with traffic pollution.

3.3 Indoor/Outdoor Classification

3.3.1 Data Pre-Processing

Firstly, the euclidean distance and euclidean distance variability features were calculated as described in Section 3.1.4. All data points with missing or null values in the required features were removed. The count data was normalised to a probability distribution as specified by Section 3.1.1 and any other features which were required to be scaled to zero mean and unit variance were also scaled at this step (based on the training data). This included the total particle counts across all bins, temperature,



(a) Smoking Measurements

(b) Smoking MVG Random Variable

(c) Traffic Measurements

(d) Traffic MVG Random Variable

Figure 3.1: Box plots showing the individual pollution source measurement and MVG random variable PSDs

relative humidity, GPS Accuracy, euclidean distance and euclidean distance standard deviation.

Chapter 4

Experiments and Evaluation Metrics

Experiments were conducted to test the various pollution source classifiers proposed in this work. Details of the experiments are given in the following sections. There are a number of experimental parameters which are common to all the experiments. Experiments were repeated five times with a different random seed. The random seed for each repeat in every experiment were 10, 100, 1000, 10000 and 100000. The random seed is used to determine the data which is split into the training and validation data set as well as the initial states of the model if applicable (test data is segmented in the same way each time to prevent leakage into the training and validation data sets). The experimental results reported are the average of the five results with standard error calculated using Equation 4.1.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \tag{4.1}$$

4.1 Indoor/Outdoor Classification Experiments

Due to the semi-supervised classification algorithm that were tested as part of this work the data was split into labelled and an unlabelled data sets (Section 3.3) based on whole data sets. In these experiments, the generalisation performance of the classifiers are evaluated by calculating the F1 score (Section 4.3.1) from both the labelled validation data set and the unlabelled validation data set. The performance of the classifiers against the labelled validation data set gives information about how the classifiers generalise to unseen data from the same environment that the classifier was trained against. The performance of the classifiers against the unlabelled validation data set gives information on how the classifiers will generalise against unseen data from different

environments. For all experiments in this section the data was pre-processed using the methodology described in Section 3.3.

4.1.1 Baseline Experiments

Baseline experiments were conducted in order to understand how standard classification algorithms would perform at indoor and outdoor classification. The baseline classifiers tested were logistic regression, SVM with RBF kernel, Random Forest and Naive Bayes classifiers.

4.1.2 Feature Importance Experiments

The co-training method described in Section 2.3.2 requires multiple classifiers trained on independent feature sets. The aim of these experiments was to evaluate the suitability of different classifiers which have been trained on independent subsets of features. Previous work [8][36] has shown the classifiers have to be trained using feature sets which are conditionally independent. There are certain sets of features which are not conditionally independent which are shown in Table 4.1. The features must be kept as subsets when splitting the features and therefore cannot be split across the different classifiers.

Feature Set	**Dependent Features**
1	Temperature, Relative Humidity
2	p_x, Total Particle Counts
3	GPS Accuracy, GPS Euclidean Distance, GPS Euclidean Distance Std.

Table 4.1: Table showing the conditionally dependent features

Experiments were conducted to determine the feature set split which gives the best two classifier performance on the labelled validation data. The experimental matrix can be seen in Table 4.2. Note that in each case, a hyperparameter search had been conducted in order to find the best performing classifier for each feature set.

Classifier 1	Classifier 2	Classifier 3	Classifier 4	Classifier 5
Temperature, Relative Humidity, p_x, Total particle counts	p_x, Total particle counts	GPS Accuracy, GPS Euclidean Distance, GPS Euclidean Distance Std.	Temperature, Relative Humidity, GPS Accuracy, GPS Euclidean Distance, GPS Euclidean Distance Std.	GPS Accuracy, GPS Euclidean Distance, GPS Euclidean Distance Std., p_x

Table 4.2: Table showing the feature sets tested for the different classifiers

4.1.3 Semi-Supervised Learning Experiments

Finally, the semi-supervised method is assessed to evaluate whether it is possible to use unlabelled data to improve indoor/outdoor classification performance in new environments. The co-training algorithm as defined in Section 2.3.2 was used. There are two additional hyperparameters which can be tuned in addition to those from the individual classifiers. The first is the number of samples which are collected before the classifiers are retrained ($n_{retrain}$). The larger $n_{retrain}$, the more diverse the data that will likely be collected. The second is the proportion of the new unlabelled data which has been labelled with the predictions from the classifiers which is incorporated into the stored labelled data set within the co-training algorithm (n_{frac}). The higher the proportion of samples incorporated into the labelled data set the quicker the algorithm will learn and adapt. However, it can also include predictions which may be of low confidence causing a degradation in classification performance over time.

A grid search was conducted using values of $n_{retrain} = 50, 100, 200, 500$ and $n_{frac} = 0.01, 0.05, 0.1$. The best performing model was chosen as the model with the highest average F1 score on the unlabelled validation set. Finally, the performance of the co-training system was evaluated against the previously unseen test set.

Classification Algorithm	Hyperparameters
Random Class	-
k-Nearest Neighbours	k = 1, 3, 5, 7, 9, 11
Logistic Regression	-
Naive Bayes	-
Support Vector Machine	kernel=linear, RBF C=10, 100, 250, 500, 1000, 2500, 10000, 50000, 100000
Random Forest	n = 1, 3, 5, 7, 9, 11, 15, 21, 31, 51
Neural Network	No. layers (L) = 1, 2, 3 No Hidden Units (U) = 50, 100, 250, 500 Learning Rate = 0.001 Activation = ReLU Optimiser = ADAM

Table 4.3: Classifiers with hyperparameters used for source apportionment

4.2 Source Apportionment Experiments

4.2.1 Classification Algorithms

For each data set a kNN baseline classifier was used, which compared the similarity of the measurement to be classified against the training set. This was then compared against the more complex model based approaches against. Further classification algorithms were then trained using each of the training data sets as specified in Sections 3.2.2.1. The classification algorithms tested, with hyperparameter ranges are shown in Table 4.3.

4.2.2 Training Data Sets

Each of the classifiers were then fitted on the different data sets available to train the source apportionment classifiers. The details of the data sets used can be seen in Table 4.4. In each case the classification algorithms were trained on the specified data set and then assessed against the validation data sets for the measured individual and multi-pollution sources. The best hyperparameters for each of the classification algorithms were selected using the highest F1 score on the measured individual pollution source data set. The aim of these experiments is to assess if training the classifiers using

the data sets consisting of mixtures of pollution sources can improve the classification performance against both the individual and mixed measure pollution source data.

4.3 Evaluation Metrics

4.3.1 Precision, Recall and F1 Score

The precision is the ratio of correctly classified examples to total true classified examples (Equation 4.2). This is the fraction of true predicted positive examples to total predicted positive examples.

$$Precision = \frac{true\ positives}{true\ positives + false\ positives} \tag{4.2}$$

The recall is the ratio of the examples that are successfully predicted (Equation 4.3) and can be described as the number of true positive examples to total positive examples.

$$Recall = \frac{true\ positives}{true\ positives + true\ negatives} \tag{4.3}$$

The F1 score is the harmonic mean between the recall and the precision (Equation 4.4). F1 score is less sensitive to class imbalance than the accuracy [40] and is therefore used in this work to assess classification performance when the classes are imbalanced. F1 score as described here is used for binary classification tasks. When multisource classification is conducted in this work, the F1 score is calculated by micro-averaging the F1 scores for each class. In this case the averaging is biased by the class frequency and therefore is not sensitive to class imbalances. In the case of the measured binary multisource data, the F1 score is calculated based on the number of correctly classified instances if the predicted class is either of the main pollution sources.

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{4.4}$$

4.3.2 Confusion Matrix

A confusion matrix is used to assess the performance of the classification methods used in this work. Each column in the matrix represents the predicted class with the rows representing the actual class and each cell represents the precision as described in 4.3.1.

Data Set No.	**Data Set Name**	**Description**
1	Individual Source (Measured)	Data set consisting of PSD measurements from individual pollution sources Section 2.2.1
2	Individual Source (MVG)	Data set consisting of PSDs sampled from the MVG random variables used to approximate the individual pollution source measurements Section 3.2.2.3
3	Mixed Source (Linear combinations of PSDs)	Data set consisting of linear combinations of PSD measurements from individual pollution sources Section 3.2.2.1
4	Mixed Source (Linear combination of Counts)	Data set consisting of PSDs from linear combinations of bin count measurements from individual pollution sources Section 3.2.2.2
5	Mixed Source (Linear combination of MVG's)	Data set consisting of samples from random variables of linear combinations of individual pollution source random variables measurements from individual pollution sources Section 3.2.2.3
6	Mixed Source (Measured)	Data set consisting of PSD measurements from multiple pollution sources Section 2.2.1

Table 4.4: Data sets used to train and validate the source apportionment classifiers

Chapter 5

Results and Discussion

5.1 Indoor/Outdoor Classification

5.1.1 Baseline Experimental Results

Baseline experiments were initially conducted in order to confirm that it was possible to accurately classify whether the Airspeck-P was indoors or outdoors from the available measured micro-environments. The results can be seen in Figure 5.1 as described in Section 4.1.1. All the classifiers gave a near perfect F1 score on the labelled validation data set using all the available features as the distributions of the data within the labelled data is from very limited environments and therefore generalisation performance is very strong. It can be seen that the generalisation performance against the unlabelled data which is from different environments is worse in all cases. These baseline results indicate that classification of whether the Airspeck-P is indoor or outdoors using the different measurements is feasible.

5.1.2 Feature Importance Experiments Results

The experiments as described in 4.1.2 were conducted and the results can be seen in Figure 5.2. It can be seen that the random forest models and the SVM's perform consistently better across all of the feature sets on the labelled validation data set (Figure 5.2). It can also be seen that the strongest features for classification was the particle information with the temperature and relative humidity and the GPS features with the temperature and humidity. As the two classifiers required for the co-training algorithm can't use the same features, the best feature sets to give the best overall performance for both classifiers is given in Table 5.1 with both models being random forests.

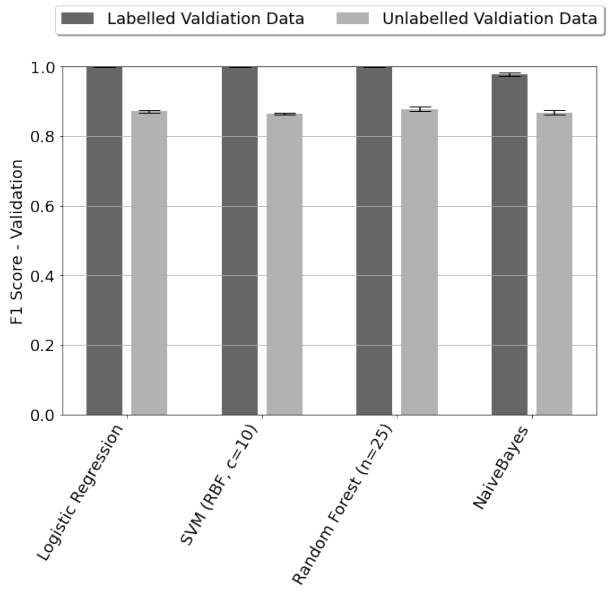
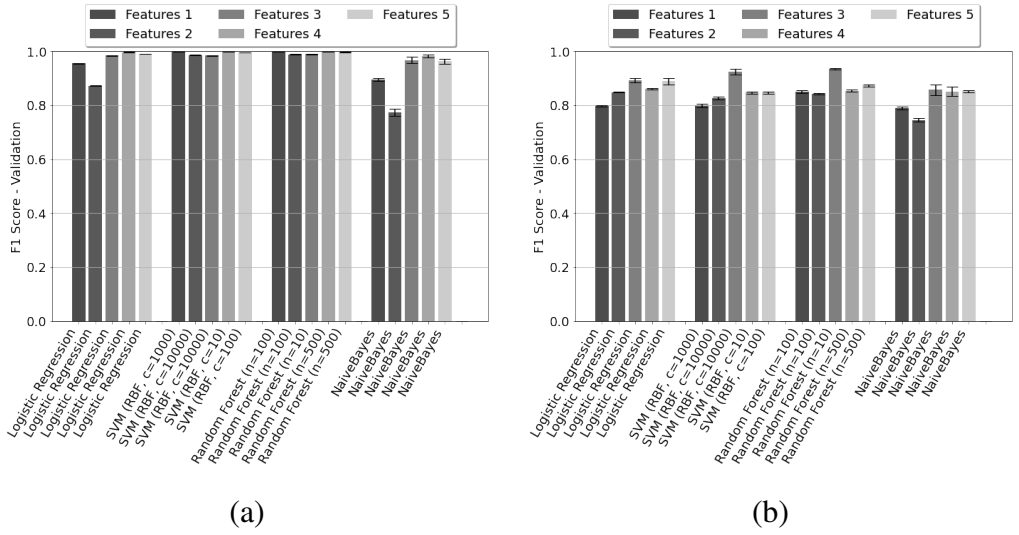


Figure 5.1: Baseline results of the indoor outdoor classifiers on the labelled and unlabelled validation data



(a) (b)

Figure 5.2: F1 scores for the classifiers trained with the different feature sets as detailed in Table 4.2 against (a) the labelled validation data, (b) the unlabelled validation data

The co-training algorithm requires that both classifiers should be able to correctly predict the class label with a high enough confidence that they are correct in the majority of cases. It is also important that the different classifiers make different classification mistakes so when one of the classifiers is wrong and gives a prediction with a low confidence, this can be corrected by the other classifier with a higher confidence.

Classifier 1	Classifier 2
p_x, Total particle counts	GPS Accuracy, GPS Euclidean Distance, GPS Euclidean Distance Std., Temperature, Relative Humidity

Table 5.1: Table showing the best feature sets tested for the two classifiers with conditionally independent features to be used with the co-training algorithm

Although there is no F1 score and feature set which will guarantee this in this case, it is hoped that with the features selected here which will predict the label based on very different information that this will be the case.

5.1.3 Co-Training Indoor/Outdoor Classification Results

The co-training algorithm as described in Section 2.3.2 was then optimised by conducting the hyperparameter tuning experiments outlined in Section 4.1.3. The algorithm was shown data sequentially from different environments and the classifiers were retrained periodically based on $n_{retrain}$.

The best performing model with an $n_{retrain} = 100$ and $n_{frac} = 0.1$ can be seen in Figure 5.3. If the co-training algorithm has indeed learned useful representations from the unlabelled data, then it would be expected that the generalisation performance would increase as the classifiers are retrained on the data from the new environments. The first observation is that using ensembling of the two models to choose the label with the highest confidence increases the overall classification performance significantly over either of the individual classification algorithms. The primary reason for this is that as both classifiers are trained using independent feature sets, the classifiers typically make mistakes on different examples.

Secondly, it is possible to observe that initially the classifier trained on the particle size information has a lower F1 score overall than the classifier using the GPS information and temperature/relative humidity. As training progresses the classifier performance increases steadily over all the unlabelled training examples, increasing the overall classification performance and the performance of each of the individual classifiers. At the end of training the classification performance is an F1 score of above 0.95 which is much higher than the performance of the baseline classifiers on the unlabelled validation data set (F1 score of 0.88).

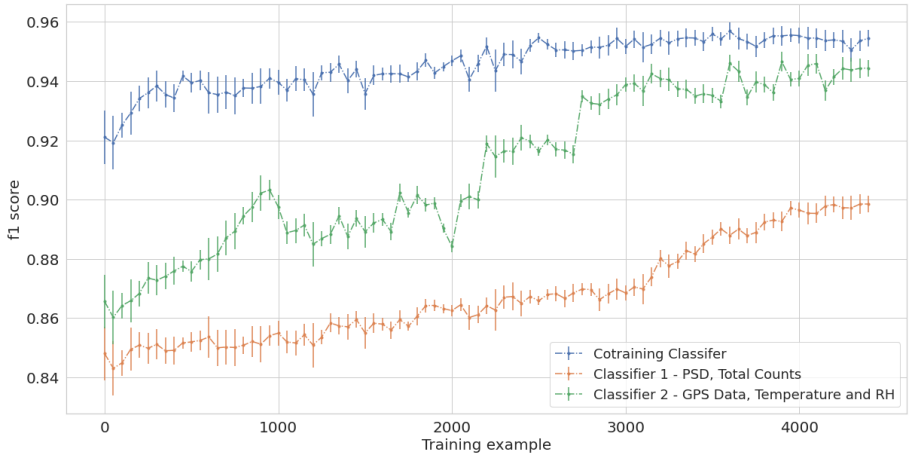


Figure 5.3: Performance of the Co-training algorithm on the unlabelled validation data set with $n_{retrain} = 100$ and $n_{frac} = 0.1$

One of the important factors in making sure the co-training algorithm works robustly when learning from un-labelled data is that both classification algorithms can accurately predict the class in all environments. When completing this work it was found that there was one data set which contained a micro environment in which neither classifier could accurately predict the label. In this case the labelled data set will become corrupted with data with incorrect labels as it is updated and this will degrade performance. This can be seen in Figure 5.4 where the overall classification performance drops after around 1000 unlabelled samples. Interestingly, the performance of the two underlying classifiers continues to improve but the overall performance does not. The reason for this is not know, however if the classifiers start to make confident mistakes on the same examples, then this may cause an overall decrease in performance. It would be expected that the system could potentially recover in time. Even in this case the final classification performance was still greater than the baseline.

The indoor/outdoor classifier presented here has shown to be able to learn in the Edinburgh micro environment from unlabelled examples. The best model was tested against the held out unlabelled test data and it achieved an F1 score of 0.949 indicating that the model generalises well. It is important to note that initial classifiers are still trained on a limited amount of data from a limited number of micro environments. In order to robustly deploy this type of classifier it would be necessary to train the initial classifiers on data collected in different seasons, households and times of day, which would reduce the risk of newly encountered environments being detrimental to overall classification performance.

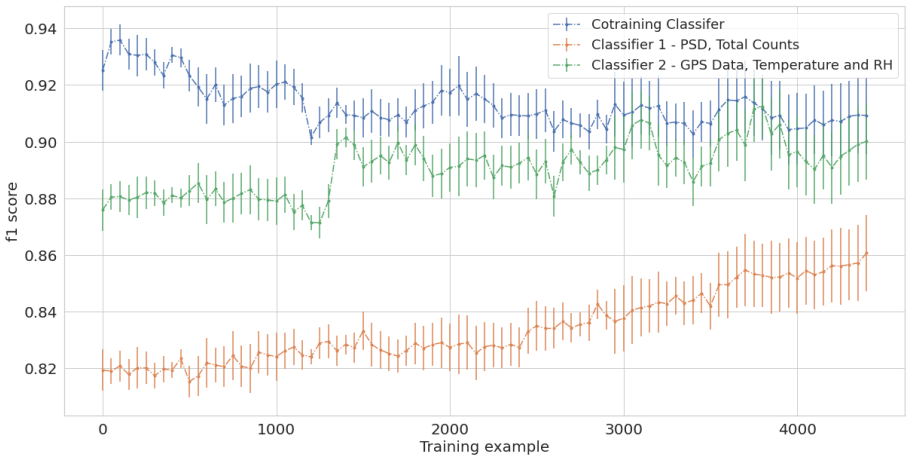


Figure 5.4: Performance of the Co-training algorithm on the unlabelled validation data set with $n_{retrain} = 100$ and $n_{frac} = 0.1$ when data is present that neither classifier can label accurately

5.2 Source Apportionment

5.2.1 Models Trained on the Individual Pollution Source Data

Baseline experiments were completed as detailed in Section 4.2.1. The kNN baseline method doesn't use learnable parameters in order to fit models instead it is based on the similarity of a data point to the training data. The importance of these simple baseline methods is to understand whether the more complex and computationally intensive methods tested later bring any additional benefits. The results for these baseline methods can be seen in Table 5.2. The baselines are fitted only against the single source measured data set and validated using the validation data from this data set and the measured binary pollution source data set.

Baseline Algorithm	**Single Source Measured**	**Binary Multi-Source Measured**
K-Nearest Neighbour (k=1)	0.814+/-0.024	0.791+/-0.006

Table 5.2: Baseline Classifier F1 Scores

Unsurprisingly the models which are based on randomly choosing a class give very poor results. The K-Nearest Neighbours methods give very strong results on the data set that it is trained on, giving a good F1 score for both the single source and multi source measured data. The confusion matrix can be seen in Figure 5.6a. There are

two main types of error. The first is the smoking class which primarily has errors by incorrectly predicting the background class. In reality, there will be some smoking data where the samples will not be recording smoking as the cigarette source is not near to the detector. However, there is no way to validate this. The other type of error that the classifier typically makes is where the PSDs of the classes overlap significantly and are relatively similar. For example, by examining the PSDs shown in Section 2.2.1, it can be seen that the background and boiling have similar PSDs as do the deep frying and traffic.

The experimental results from training on the individual pollution source data can be seen in Figure 5.5. The results indicate the performance of using a classifier to be able to predict the pollution source from the PSD when trained using the measured particle size data. The F1 scores from the validation set shows that all the classifiers tested can generalise well to new data. The classifiers in this section have all had relevant hyper-parameters tuned using a grid search methodology. The results show that none of the linear classification methods can classify the pollution source accurately with lower F1 score than the other non-linear classification algorithms. The best performing model was the SVM with RBF kernel. This performed well on both the individual and multiple pollution source data sets. However, when identifying the individual pollution sources the baseline kNN algorithm gives a higher F1 score than any of the other models. This is not surprising as the K-nearest neighbours algorithm designates the class based on the similarity of the measurements to the classes in the training data. Importantly though the kNN algorithm does not generalise as well to the measured multi source data, giving a lower F1 score than some of the other algorithms.

The confusion matrix for the best performing model can be seen in Figure 5.6b. From the confusion matrix it can be seen that the model makes mistakes between the classes which again have similar PSDs as shown in Section 2.2.1. These are typically from similar pollution source classes, for example, from boiling water and the background, as well as the mosquito coil and deep frying.

The second set of experiments investigated the feasibility of training the classification algorithms using the data sampled from the MVG random variables which approximate the measured PSDs. The results from training the different classification algorithms can be seen in Figure 5.5b. It is observable that there is a degradation in performance in all of the classifiers when validating against the measured particle size data as opposed to the generated validation data. The SVM using the RBF Kernel gives the highest F1 score on the individual pollution sources while the kNN (k=1) and the

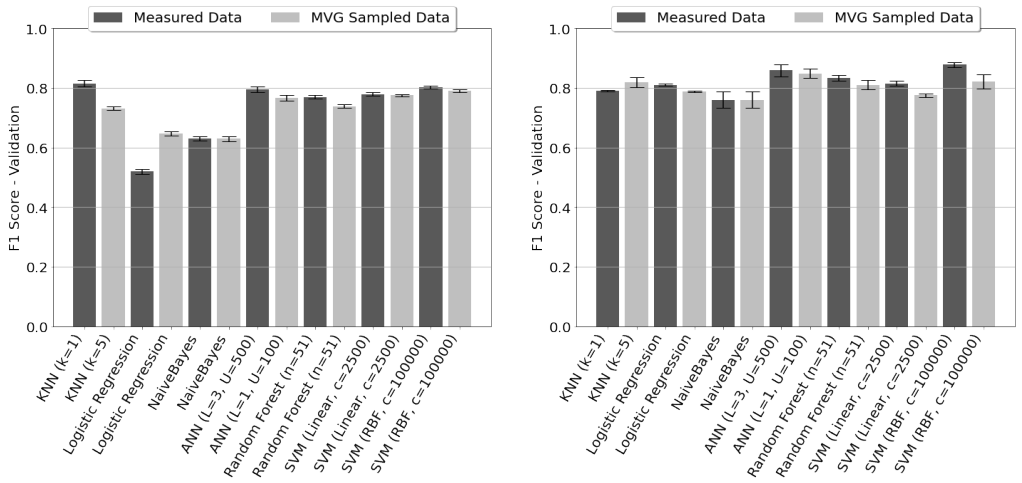


Figure 5.5: F1 scores for the models trained on the individual pollution source data (a) Measured Data, (b) MVG Random Variables

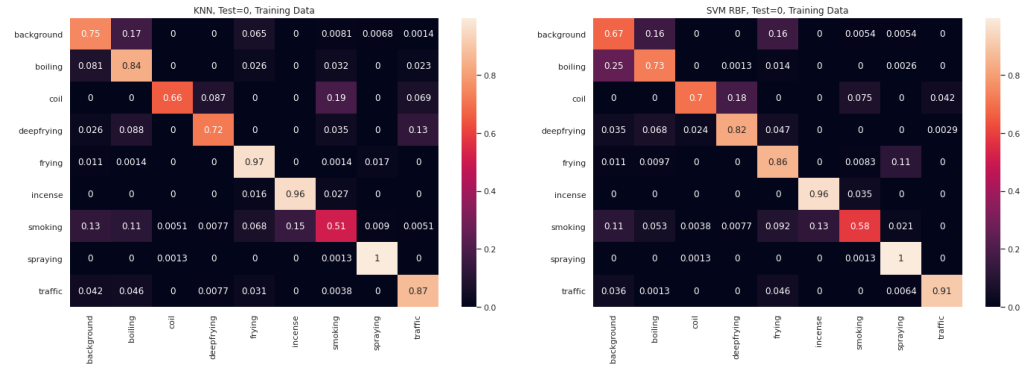
SVM with RBF kernel both perform best on the measured multiple pollution source data, but this is still worse than the results shown in Figure 5.5a where the classifiers were trained on the raw measurements.

The degradation in performance was due to the MVG random variables which are fitted using the measured mean and covariance not representing the distribution of the data exactly. Although the variability of the measurements were expected to be random and Gaussian around the mean, in reality they are not. There is time-based variability in the data which is not fully captured by the distribution fitted to the data.

5.2.2 Multi-source Apportionment Experimental Results

Three methods were assessed to produce a robust classification system which can be used to determine the dominant pollution source from a set of known pollution sources. The aim of mixing the pollution source measurements was to produce more robust decision boundaries when multiple pollution sources are present in the measured samples than the decision boundaries which are formed when individual pollution source measurements are used.

The first method investigated training the classifiers on the the mixed multi pollution source data which was created by taking linear combinations of the PSDs from the individual pollution sources. The summary of the results from the different classification algorithms can be seen in Figure 5.7. Overall, the performance of the classifiers

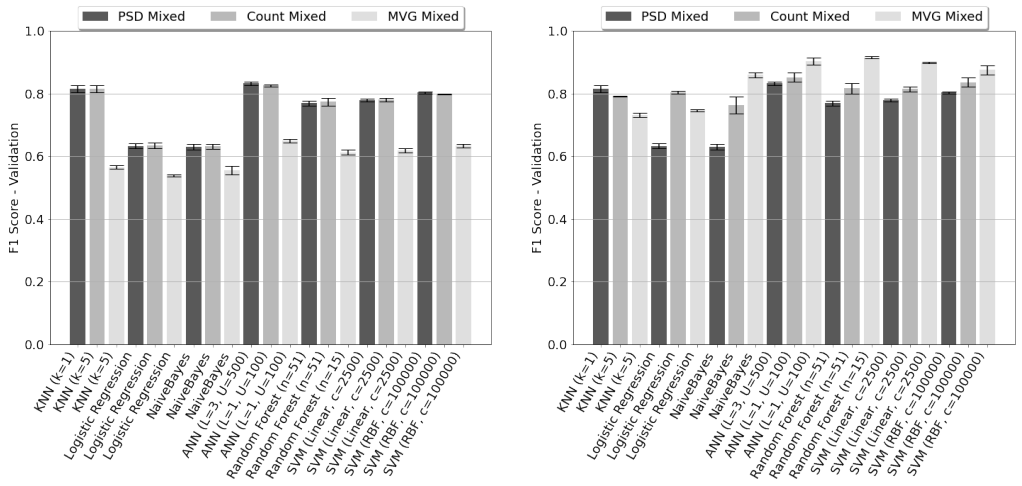


(a) (b)

Figure 5.6: Confusion matrices showing (a) The kNN (k=1) baseline model and (b) the SVM with RBF kernel and regularisation constant c=100000 both trained on the measured individual pollution source data.

against the measured individual pollution source data set increased. This was especially true for the ANN model which produced the best validation results against this data set and unlike when trained on the individual pollution source measurements beat the kNN classifier on both the validation data sets. This model also increased the F1 score compared with the ANN trained only on the individual pollution source measurements (Figure 5.5). ANN can form very complex decision boundaries due to the flexibility of the models, which can lead to over fitting. Even though early stopping was used in order to stop the ANN overfitting the data this could still be occurring. Taking linear combinations of the PSD measurements from the individual pollution source data will have a regularising effect on the classifier as there will be more data structure around where the decision boundaries will be formed.

The second method investigated the use of multi-source data which was mixed through taking linear combinations of the raw bin count measurements from the individual pollution sources before normalising to the PSD. This mixed data set was created in this way in order to account for the different total number of particles in the different pollution sources. However, this may also introduce sensitivities in the classification of the pollution source based on how far the measurement device is from the pollution source. The summary of the results from the different classification algorithms can be seen in Figure 5.7. A similar trend as the previously discussed method of mixing the pollution sources is observed with a general improvement of the F1 score against the individual measured pollution source data set. However, overall the performance of the classifiers trained against this data set is marginally worse than when



(a) (b)

Figure 5.7: F1 scores for the models trained on the mixed pollution source data validated against (a) Individual measured pollution source data, (b) multi-source measured data

simply using the linear combinations of the individual PSDs. Again the ANN was the best performing classifier and the only model to outperform the kNN baseline.

The last method discussed here used the data sampled from the mixed MVG Random Variables as described in Section 3.2.2.3. The results can be seen in Figure 5.7. The performance on the individual measured validation data was poor with low F1 scores from all of the algorithms including the kNN and ANN models which have previously all performed well. This indicates that the assumption of approximating the mixed pollution source distributions as MVG random variables is not valid. One very interesting point though is the performance against the measured multi-pollution source data is actually superior with an increase in F1 score. This is an artefact of the classifiers having a very high accuracy at correctly labelling the coil and incense pollution source which are two of the pollution sources which are present in most of the multiple pollution source validation set.

This section has shown that mixing the pollution source gives a more robust classification performance against the multi-source pollution data than when the classifier is trained solely against the individual source data. The confusion matrix for the best model can be seen in Figure 5.8 where it can be observed that the model has improved performance in nearly all classes except for smoking where performance is still relatively poor. It is worth noting however that the classifier trained using only the individ-

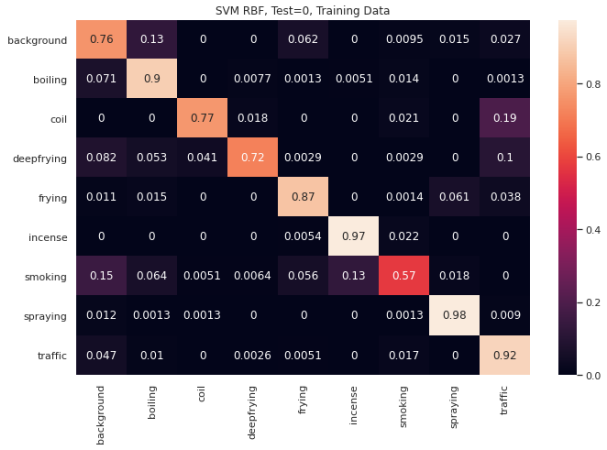


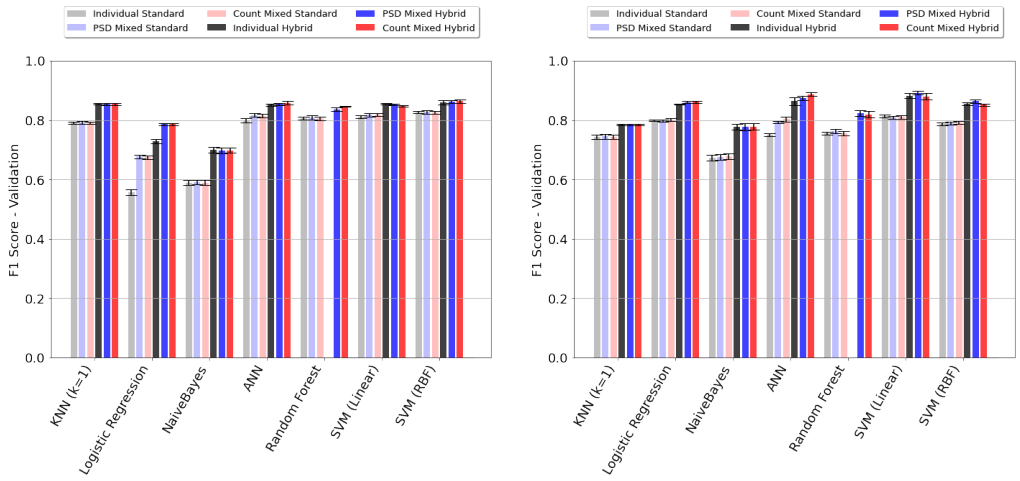
Figure 5.8: Confusion matrix demonstrating the best performing model ANN L=3, U=500 trained on the multi source data generated using linear combinations of the individual PSDs, validated on the measure individual pollution source validation data set

ual pollution sources still performs well and with enough training data could produce robust decision boundaries which could robustly deal with multi source pollution data.

5.2.3 A Hybrid Source Apportionment Classifier

The indoor and outdoor micro-environments are characterised by dominant pollution sources: for example, at certain times of the day cooking sources dominate indoor air pollution; similarly, traffic-related pollution sources will dominate around busy traffic junctions. This observation is used to improve the source apportionment classifiers developed in Section 5.2.1 and 5.2.2, by adding an indoor/outdoor label – the output of the indoor/outdoor classifier was added to the features during training using the classification system developed in Section 4.2.2 (in a minority of cases this had to done manually as the data was gathered with a previous generation of Airspeck device which did not have the necessary features). The indoor-outdoor label enhanced data set was then used to train the source apportionment classifiers in Sections 5.2.1 and 5.2.2. Note that these results are not directly comparable; therefore, the hybrid model and the classifiers in Sections 5.2.1 and 5.2.2 were trained on the same data set, with and without the indoor-outdoor labels, using the same model architectures and hyper-parameters as used previously to see if the hybrid system yielded benefits.

The results of the hybrid source apportionment model on the individual pollution



(a) (b)

Figure 5.9: F1 scores for the models trained on the individual pollution source data validated against (a) Individual measured pollution source data, (b) multisource measured data

source measurement can be seen in Figure 5.9a and on the multi-pollution source data in Figure 5.9b. The hybrid classifier out performs the standard classifier in all cases. Including the indoor/outdoor label adds another degree of freedom to the model and forms a decision boundary when the PSDs of the individual pollution sources were not separable.

5.2.4 Testing the Source Apportionment System

The source apportionment systems were tested on a held-out data set not used in any of the training or validation processes. It was also tested on a subset of the Dublin data set previously used by Sanatani [38] which is a data set consisting of 900 measurements specifically labelled for the presence of cigarette smoke.

The results can be seen in Table 5.3 for the best standard (ANN, L=3 and U=500) and hybrid (ANN, L=3 and U=500) models. The models perform well against the held out test set which is a subset of the data that the source apportionment system was trained against with F1 scores which were similar to the F1 scores from the validation data sets indicating that over-fitting has not occurred. The F1 scores against the Dublin test set, however, are significantly lower. This is probably due to a number of reasons: firstly, the Airspeck device used was a previous model with different characteristics compared to the current version of the Airspeck devices on which the models were

Data Set	Standard Model (ANN, L=3 and U=500) Trained on the Mixed PSD Data	Hybrid model (ANN L=3, U=500) Trained on the Mixed Count Data
Test Data	0.826	0.862
Dublin Data	0.191	0.211

Table 5.3: F1 score of the best performing source apportionment models against the Held out test sets

trained; and, secondly, the Dublin data was collected from shelters outside pubs and cannot be accurately classified as either fully indoors or outdoors. This to some extent explains why there is no improvement when using the hybrid source apportionment model as it is unable to distinguish between cigarette smoke and other indoor/outdoor pollutants with similar PSDs. If the labels are manually adjusted to be outside (as is the correct label) the F1 score of the hybrid system increases to 0.884 indicating how import the classification of the indoor/outdoor label could be with this hybrid system.

5.3 Monitoring Air Quality During the Easing of COVID-19 Lock-down Restrictions in London

A network of four Airspeck-S monitors were mounted on lamp posts in the Royal Borough of Kensington and Chelsea to monitor at 5-minute intervals the concentrations of airborne particulates (PM), nitrogen dioxide and ozone. Figure 5.10 shows the cumulative PM for Sensor 905801CA0E1F1D11, with data from the other sensors shown in Appendix E (exact locations of the sensors can be seen in Appendix E). It was expected that as the economic activity ramped up after the lock-down, that the air quality would worsen with an increase in the different PM fractions associated with vehicular traffic. Although the monitoring period has not been sufficiently long to discern major trends as yet, some significant spikes can be observed. For example, during the long weekend in May starting in the afternoon of Friday 22 May, and, reduction during the week commencing 13^{th} July, 2020 when the schools in the area closed for summer vacation. There was also large increases in PM_1 and $PM_{2.5}$ for some of the sensors. Traffic data was also collected, as described in Section 2.2.3 using the HERE traffic API. Figure 5.11a illustrates an example of its time-series output for sensor 905801CA0E1F1D11

aligned with the raw PM data (other sensor data is included in Appendix E). The traffic data shows increases in the jam factor (which is a proxy for the quantity of traffic) along this road which could account for the rise in PM. Finally, the nitrogen dioxide gas concentrations Figure 5.11b show a similar trend to the PM values as they are both derived mainly from vehicular traffic [39] and the concentrations of ozone, a secondary pollutant of traffic, which is formed in the presence of sunlight and precursors such as oxides of nitrogen and volatile organic compounds, increase when the temperature rises.

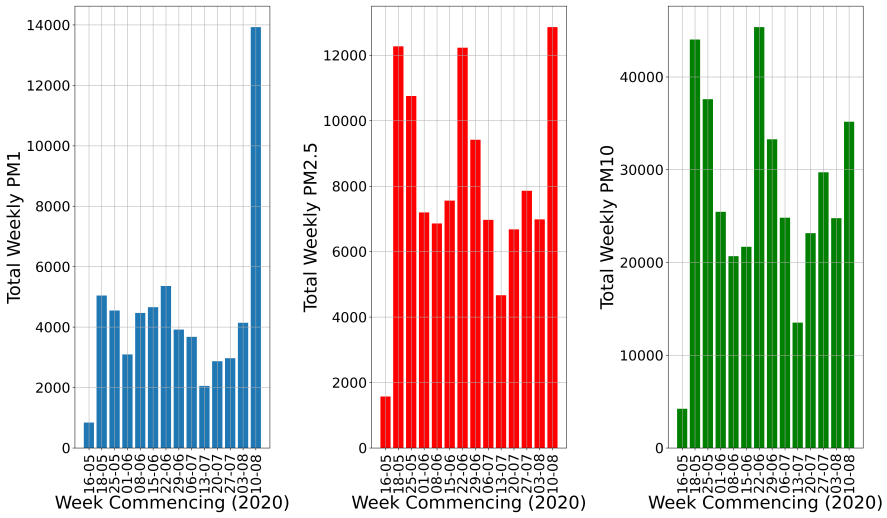
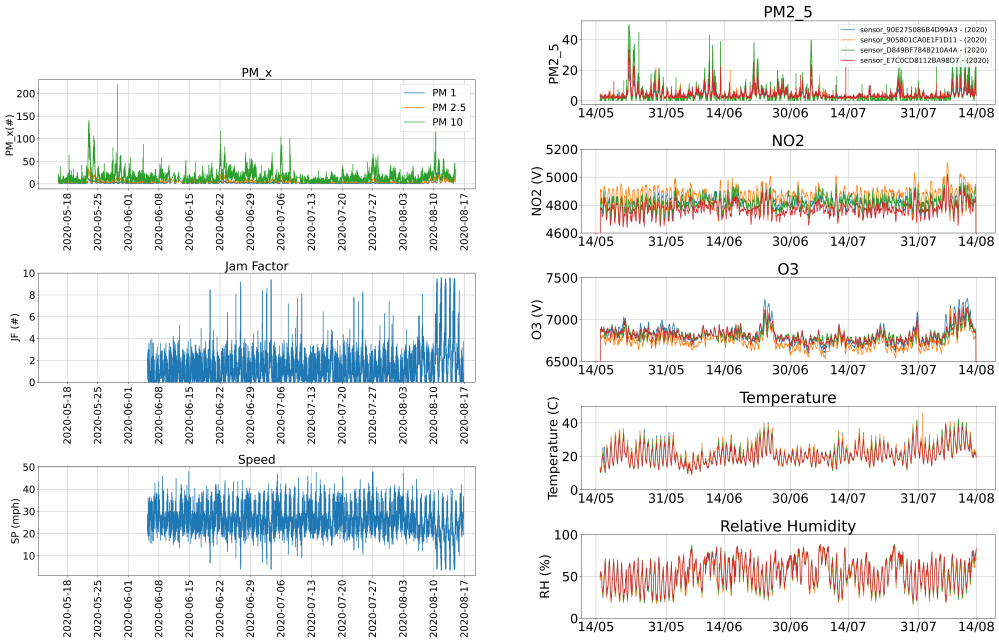


Figure 5.10: Weekly cumulative PM_1, $PM_{2.5}$ and PM_{10} for Sensor 905801CA0E1F1D11

5.4 Prediction of the Elemental Composition from Airborne Particle Size Information

Regression methods were used to predict the elemental composition from the Airspeck-S PSD measurements. Experiments were conducted which involved fitting both the baseline and linear regression models with associated hyper-parameter searches. As the data set was so small (six leaf samples) a leave-one-out cross validation approach was used. The results of the regression can be seen in Table 5.4. Importantly it can be seen that even the best results can only match the baseline method of simply making the prediction using the average of the training elemental compositions. This indicates



(a) (b)

Figure 5.11: (a) Time-series of the PM fractions and traffic data for sensor 905801CA0E1F1D11, (b) Airspeck-S measurements for all sensors

that there is no correlation between the measured airborne PSD and the composition. Different time periods for assessing the PSD were chosen, however, this had no effect on the performance of the models. In order to improve and assess the feasibility of the project, it is recommend that more composition based samples are taken and assessed.

Model	**Average Validation RMSE (%)**
Baseline (mean of training compositions)	4.920
Ridge Regression	4.923
Ridge Regression with RBF basis functions	4.920

Table 5.4: RMSE of the elemental composition prediction from the measured Airspeck-S PSD

Chapter 6

Conclusions and Future Work

This work has shown how machine learning techniques can be used to develop understanding of airborne pollution. An indoor/outdoor classification system was developed which could be trained on limited amounts of training data and then learn and improve the classification performance from new unlabelled measurements. This was shown to allow the system to learn and improve the overall classification system as it was introduced to new and unseen micro-environments. It was also observed however that if the system encountered a micro-environment which the system couldn't accurately classify, then this could have a detrimental impact on the overall system performance. In reality though, the system was trained with less data than would normally be used in order to show the adapted co-training algorithm could improve performance. In order to deploy this system for real world use, the initial supervised classification algorithms would be trained on all the available training data available which would increase the robustness of the system to this type of problem.

An airborne pollution apportionment system was developed using classical classification techniques by training the models on data sets augmented by mixing measurements of the individual pollution source PSDs through either directly mixing the PSDs using linear combinations or through linear combinations of the particle counts in each bin before normalisation to the PSD. Lastly, each of the individual pollution source measurements was approximated as a MVG random variable and then mixed by taking linear combinations of the random variable parameters. In general the models which were trained on the data sets from the mixed and individual pollution sources showed an increase in performance in comparison to the models trained only on the individual pollution source data. However, the technique for approximating the pollution source as MVG random variables performed poorly. This was due to the distributions

formed by mixing the random variables not being a good representation of the actual data and therefore forcing the classifiers to learn decision boundaries that degraded performance more than any of the other methods. A hybrid source apportionment system which used the label from the indoor classifier as an additional feature was developed. It was shown that in all cases the hybrid source apportionment model increased the performance of the source apportionment system as the systems trained only on the PSD. The only time this was not the case was during testing against the Dublin data sets where performance was poor for both the standard and hybrid source apportionment systems.

The classification architectures here have all treated the data as independent measurements. The data however is of course a time series and therefore some form model architecture which takes into account previous measurements would be worth investigating. However, current data sets are not suitable for this purpose as they are all disjoint events in controlled conditions, instead labelled data sets with natural transitions for different pollution sources in different micro environments would be needed, although potentially difficult to obtain. Another potential avenue of interest would be to use an approach for which the source apportionment system could learn from newly measured data. As was seen when testing the system, measurements from unobserved micro-environments may be very different to those that the initial classifier was trained on. Having a system which can learn from newly taken unlabelled data could help improve the performance and be a cost effective way to adapt to new micro-environments.

Finally machine learning methods were evaluated in order to predict the elemental composition of road side particulate pollution on leaf samples from Airspeck-S air quality measurements. It was found that with the limited number of samples available, it was not possible to accurately predict the elemental composition any more accurately than just taking the average of the training samples. In order to assess the feasibility of this work further a large sample size needs to be investigated. If found that it is possible to relate the Airspeck air quality measurements to the elemental composition when more samples are available, then the next logical step is to attempt to predict the toxicity to humans as some form of toxicity index [35] from the predicted elemental compositions.

Bibliography

[1] Here traffic api. www.here.com. Accessed: 2020-08-19.

[2] KV Abhijith and Prashant Kumar. Quantifying particulate matter reduction and their deposition on the leaves of green infrastructure. *Environmental Pollution*, page 114884, 2020.

[3] E Araviiskaia, E Berardesca, T Bieber, G Gontijo, M Sanchez Viera, L Marrot, B Chuberre, and B Dreno. The impact of airborne pollution on skin. *Journal of the European Academy of Dermatology and Venereology*, 33(8):1496–1505, 2019.

[4] DK Arvind, CA Bates, DJ Fischer, and Janek Mann. A sensor data collection environment for clinical trials investigating health effects of airborne pollution. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 88–91. IEEE, 2018.

[5] DK Arvind, CA Bates, DJ Fischer, and Janek Mann. Spatially-resolved estimation of personal dosage of airborne particulates for ambulatory subjects using wearable sensors. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 29–32. IEEE, 2018.

[6] Claudio A Belis, Bo R Larsen, Fulvio Amato, Imad El Haddad, Olivier Favez, Roy M Harrison, Philip K Hopke, Silvia Nava, Pentti Paatero, André Prevot, et al. European guide on air pollution source apportionment with receptor models, 2014.

[7] Rakesh Bhutiani, Dipali Bhaskar Kulkarni, Dev Raj Khanna, and Ashutosh Gautam. Water quality, pollution source apportionment and health risk assessment of heavy metals in groundwater of an industrial area in north india. *Exposure and Health*, 8(1):3–18, 2016.

[8] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, 1998.

[9] Vlad-Andrei Buzatu. *Sensing Spaces Indoor-Outdoor Detection Classification of Pollution Sources*. 4th year project report, School of Informatics, University of Edinburgh, 2018.

[10] Dario Camuffo. Acid rain and deterioration of monuments: how old is the phenomenon? *Atmospheric Environment. Part B. Urban Atmosphere*, 26(2):241–247, 1992.

[11] Haiyang Chen, Yanguo Teng, Jinsheng Wang, Liuting Song, and Rui Zuo. Source apportionment of trace element pollution in surface sediments using positive matrix factorization combined support vector machines: application to the jinjiang river, china. *Biological trace element research*, 151(3):462–470, 2013.

[12] Sheldon K Friedlander. Chemical element balances and identification of air pollution sources. *Environmental science & technology*, 7(3):235–240, 1973.

[13] Miriam E Gerlofs-Nijland, Jan AMA Dormans, Henk JT Bloemen, Daan LAC Leseman, A John F Boere, Frank J Kelly, Ian S Mudway, Al A Jimenez, Ken Donaldson, Cecilia Guastadisegni, et al. Toxicity of coarse and fine particulate matter from sites with contrasting traffic profiles. *Inhalation toxicology*, 19(13):1055–1069, 2007.

[14] Sharad Gokhale, Tibor Kohajda, and Uwe Schlink. Source apportionment of human personal exposure to volatile organic compounds in homes, offices and outdoors by chemical mass balance and genetic algorithm receptor models. *Science of the Total Environment*, 407(1):122–138, 2008.

[15] AK Gupta, Kakoli Karar, and Anjali Srivastava. Chemical mass balance source apportionment of pm10 and tsp in residential and industrial sites of an urban region of kolkata, india. *Journal of hazardous materials*, 142(1-2):279–287, 2007.

[16] Rene Hansen, Rico Wind, Christian S Jensen, and Bent Thomsen. Seamless indoor/outdoor positioning handover for location-based services in streamspin. In *2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware*, pages 267–272. IEEE, 2009.

[17] A Izzotti, S Parodi, A Quaglia, C Fare, and M Vercelli. The relationship between urban airborne pollution and short-term mortality: quantitative and qualitative aspects. *European journal of Epidemiology*, 16(11):1027–1034, 2000.

[18] Kakoli Karar and AK Gupta. Source apportionment of pm10 at residential and industrial sites of an urban region of kolkata, india. *Atmospheric Research*, 84(1):30–41, 2007.

[19] Frank J Kelly and Julia C Fussell. Size, source and chemical composition as determinants of toxicity attributable to ambient particulate matter. *Atmospheric environment*, 60:504–526, 2012.

[20] Ki-Hyun Kim, Shamin Ara Jahan, and Ehsanul Kabir. A review on human health perspective of air pollution with respect to allergies and asthma. *Environment international*, 59:41–52, 2013.

[21] Young Min Kim, Stuart Harrad, and Roy M Harrison. Concentrations and sources of vocs in urban domestic and public microenvironments. *Environmental science & technology*, 35(6):997–1004, 2001.

[22] Yungeun Kim, Songhee Lee, Seokjoon Lee, and Hojung Cha. A gps sensing strategy for accurate and energy-efficient outdoor-to-indoor handover in seamless localization systems. *Mobile Information Systems*, 8(4):315–332, 2012.

[23] Gregory S Kowalczyk, Carl E Choquette, and Glen E Gordon. Chemical element balances and identification of air pollution sources in washington, dc. *Atmospheric Environment (1967)*, 12(5):1143–1153, 1978.

[24] Leon PM Lamers, Sarah-J Falla, Edyta M Samborska, Ivo AR van Dulken, Gijs van Hengstum, and Jan GM Roelofs. Factors controlling the extent of eutrophication and toxicity in sulfate-polluted freshwater wetlands. *Limnology and oceanography*, 47(2):585–593, 2002.

[25] B Lee, C Lim, and K Lee. Classification of indoor-outdoor location using combined global positioning system (gps) and temperature data for personal exposure assessment. *Environmental health and preventive medicine*, 22(1):29, 2017.

[26] Jong-Myoung Lim, Jin-Hong Lee, Jong-Hwa Moon, Yong-Sam Chung, and Ki-Hyun Kim. Source apportionment of pm10 at a small industrial area using positive matrix factorization. *Atmospheric Research*, 95(1):88–100, 2010.

[27] Chih-Chung Lin, Shui-Jen Chen, Kuo-Lin Huang, Wen-Ing Hwang, Guo-Ping Chang-Chien, and Wen-Yinn Lin. Characteristics of metals in nano/ultrafine/fine/coarse particles collected beside a heavily trafficked road. *Environmental Science & Technology*, 39(21):8113–8122, 2005.

[28] Uri Lipowezky. Indoor-outdoor detector for mobile phone cameras using gentle boosting. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 31–38. IEEE, 2010.

[29] PM Mannucci. Airborne pollution and cardiovascular disease: burden and causes of an epidemic, 2013.

[30] Mauro Masiol, Stefania Squizzato, Giancarlo Rampazzo, and Bruno Pavoni. Source apportionment of pm2. 5 at multiple sites in venice (italy): spatial variability and the role of weather. *Atmospheric Environment*, 98:78–88, 2014.

[31] Flavia Mazzoli-Rocha, Clarissa Bichara Magalhaes, Olaf Malm, Paulo Hilário Nascimento Saldiva, Walter Araujo Zin, and Débora Souza Faffe. Comparative respiratory toxicity of particles produced by traffic and sugar cane burning. *Environmental research*, 108(1):35–41, 2008.

[32] Richard L McKenzie, Pieter J Aucamp, Alkiviades F Bais, Lars Olof Björn, Mohamad Ilyas, and Sasha Madronich. Ozone depletion and climate change: impacts on uv radiation. *Photochemical & Photobiological Sciences*, 10(2):182–198, 2011.

[33] David A Olson, John Turlington, Rachelle M Duvall, Stephen R McDow, Carvin D Stevens, and Ron Williams. Indoor and outdoor concentrations of organic and inorganic molecular markers: Source apportionment of pm2. 5 using low-volume samples. *Atmospheric Environment*, 42(8):1742–1751, 2008.

[34] Kazushige Ouchi and Miwako Doi. Indoor-outdoor activity recognition by a smartphone. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 600–601, 2012.

[35] Minhan Park, Hung Soo Joo, Kwangyul Lee, Myoseon Jang, Sang Don Kim, Injeong Kim, Lucille Joanna S Borlaza, Heungbin Lim, Hanjae Shin, Kyu Hyuck Chung, et al. Differential toxicities of fine particulate matters from various sources. *Scientific reports*, 8(1):1–11, 2018.

[36] Valentin Radu, Panagiota Katsikouli, Rik Sarkar, and Mahesh K Marina. A semi-supervised learning approach for robust indoor-outdoor detection with smartphones. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*, pages 280–294, 2014.

[37] Weeberb J Requia, Brent A Coull, and Petros Koutrakis. Evaluation of predictive capabilities of ordinary geostatistical interpolation, hybrid interpolation, and machine learning methods for estimating pm2. 5 constituents over space. *Environmental research*, 175:421–433, 2019.

[38] Tanya Sanatani. *Multiple pollution source appointment based on airborne particle size fractions using the Airspeck monitor*. Masters thesis, School of Informatics, University of Edinburgh, 2018.

[39] Zang-Ho Shon, Ki-Hyun Kim, and Sang-Keun Song. Long-term trend in no2 and nox levels and their emission ratio in relation to road traffic activities in east asia. *Atmospheric Environment*, 45(18):3120–3131, 2011.

[40] Alaa Tharwat. Classification assessment methods. *Applied Computing and Informatics*, 2018.

[41] Peter Wåhlin, Ruwim Berkowicz, and Finn Palmgren. Characterisation of traffic-generated particulate matter in copenhagen. *Atmospheric Environment*, 40(12):2151–2159, 2006.

[42] Jin Zhang, Ruifei Li, Xiaoying Zhang, Changfeng Ding, and Pei Hua. Traffic contribution to polycyclic aromatic hydrocarbons in road dust: A source apportionment analysis under different antecedent dry-weather periods. *Science of the Total Environment*, 658:996–1005, 2019.

Appendix A

General Information

Bin no.	**Particle Size Range (μm)**
0	0.38-0.52
1	0.52-0.75
2	0.75-1.0
3	1.0-1.3
4	1.3-1.5
5	1.5-2.0
6	2.0-3.0
7	3.0-4.0
8	4.0-5.0
9	5.0-6.5
10	6.5-8.0
11	8.0-10.0
12	10.0-12.0
13	12.0-14.0
14	14.0-16.0
15	16.0-Max

Table A.1: Bin Value Ranges

Appendix B

Indoor/Outdoor Classification

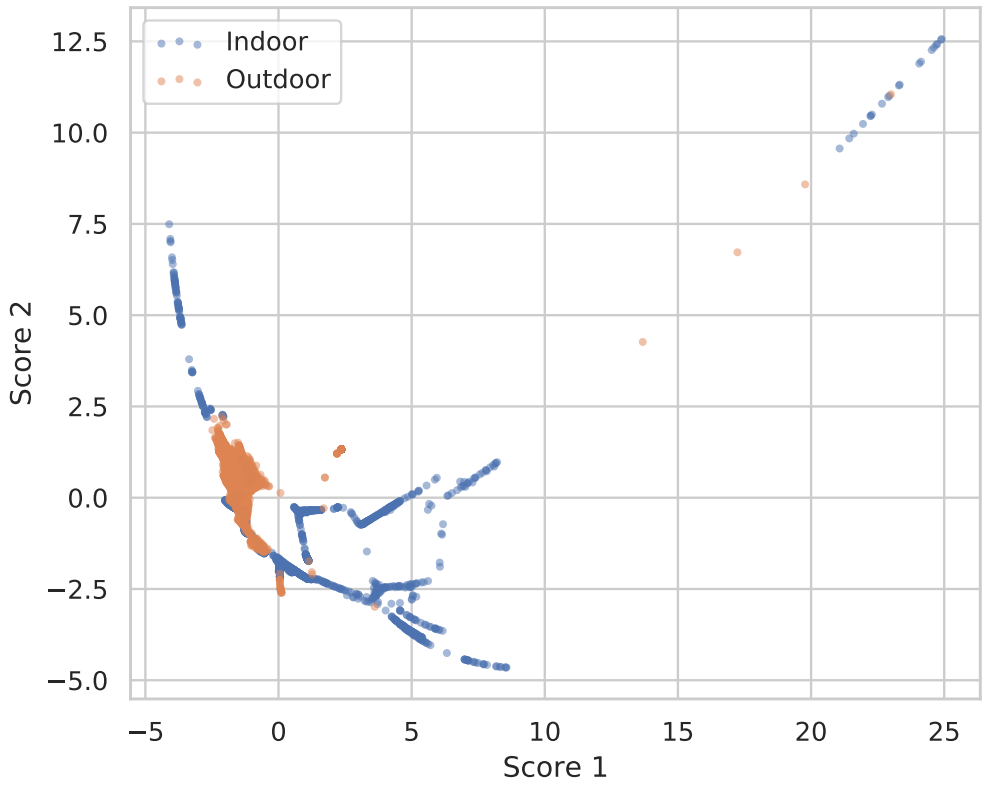


Figure B.1: Isomap decomposition of the indoor/outdoor classification data set

Classification Algorithm	**Validation F1 Score (labelled)**	**Validation F1 Score (unlabelled)**
Logistic Regression	0.998+/-0.001	0.873+/-0.004
SVM (RBF Kernel) (c=10000)	0.998+/-0.001	0.862+/-0.004
Random Forest (500)	0.999+/-0.000	0.880+/-0.005
Naive Bayes	0.977+/-0.005	0.866+/-0.008

Table B.1: Table showing the baseline classification results using all available features

Classification Algorithm	**Validation F1 Score (labelled)**	**Validation F1 Score (unlabelled)**
Logistic Regression	0.955+/-0.002	0.797+/-0.003
SVM (RBF Kernel) (c=1000)	0.999+/-0.000	0.799+/-0.008
Random Forest (n=100)	0.998+/-0.001	0.851+/-0.005
Naive Bayes	0.895+/-0.005	0.790+/-0.004

Table B.2: F1 score of the classifier trained with Feature set 1 from Table 5.1

Classification Algorithm	**Validation F1 Score (labelled)**	**Validation F1 Score (unlabelled)**
Logistic Regression	0.873+/-0.001	0.849+/-0.000
SVM (RBF Kernel) (c=10000)	0.986+/-0.001	0.826+/-0.004
Random Forest (n=100)	0.988+/-0.001	0.843+/-0.003
Naive Bayes	0.773+/-0.013	0.745+/-0.007

Table B.3: F1 score of the classifier trained with Feature set 2 from Table 5.1

Classification Algorithm	**Validation F1 Score (labelled)**	**Validation F1 Score (unlabelled)**
Logistic Regression	0.984+/-0.001	0.894+/-0.008
SVM (RBF Kernel) (c=10000)	0.985+/-0.001	0.924+/-0.010
Random Forest (n=10)	0.990+/-0.001	0.935+/-0.002
Naive Bayes	0.968+/-0.011	0.857+/-0.019

Table B.4: F1 score of the classifier trained with Feature set 3 from Table 5.1

Classification Algorithm	**Validation F1 Score (labelled)**	**Validation F1 Score (unlabelled)**
Logistic Regression	0.997+/-0.001	0.860+/-0.003
SVM (RBF Kernel) (c=10)	0.999+/-0.001	0.847+/-0.003
Random Forest (n=500)	0.999+/-0.001	0.854+/-0.004
Naive Bayes	0.983+/-0.005	0.851+/-0.017

Table B.5: F1 score of the classifier trained with Feature set 4 from Table 5.1

Classification Algorithm	**Validation F1 Score (labelled)**	**Validation F1 Score (unlabelled)**
Logistic Regression	0.991+/-0.001	0.889+/-0.011
SVM (RBF Kernel) (c=100)	0.995+/-0.000	0.847+/-0.004
Random Forest (n=500)	0.997+/-0.000	0.873+/-0.005
Naive Bayes	0.963+/-0.008	0.852+/-0.004

Table B.6: F1 score of the classifier trained with Feature set 5 from Table 5.1

Appendix C

Source Apportionment Data

Model	Individual Pollution Source Measured	Multiple Pollution Source Measured
K- Nearest Neighbour (1 Nearest Neighbour)	0.814+/-0.011	0.791+/-0.003
Logistic Regression	0.554+/-0.009	0.864+/-0.004
Naive Bayes	0.631+/-0.008	0.760+/-0.028
Neural Network (L=3, U=500)	0.800+/-0.009	0.859+/-0.021
Random Forest ($n_{est} = 51$)	0.769+/-0.006	0.833+/-0.009
Support Vector Machine (Linear Kernel, c=2500)	0.779+/-0.006	0.815+/-0.023
Support Vector Machine (RBF Kernel, c=100000)	0.803+/-0.005	0.878+/-0.009
Support Vector Machine (Sigmoid, c=10)	0.225+/-0.009	0.426+/-0.042

Table C.1: Validation F1 scores for classifiers trained using the measured single source data

Model	**Individual Pollution Source Measured**	**Multiple Pollution Source Measured**
K- Nearest Neighbour (5 Nearest Neighbour)	0.732+/-0.006	0.819+/-0.016
Logistic Regression	0.647+/-0.008	0.788+/-0.003
Naive Bayes	0.630+/-0.008	0.760+/-0.028
Neural Network (L=1, U=100)	0.767+/-0.009	0.849+/-0.015
Random Forest (n_{est}=51)	0.739+/-0.006	0.811+/-0.016
Support Vector Machine (Linear Kernel, c=2500)	0.775+/-0.005	0.775+/-0.012
Support Vector Machine (RBF Kernel, c=100000)	0.791+/-0.024	0.823+/-0.024

Table C.2: Validation F1 scores for classifiers trained using the data set sampled from the individual MVG random variables

Model	**Individual Pollution Source Measured**	**Multiple Pollution Source Measured**
K- Nearest Neighbour (1 Nearest Neighbour)	0.814+/-0.011	0.791+/-0.003
Logistic Regression	0.632+/-0.008	0.803+/-0.004
Naive Bayes	0.630+/-0.009	0.762+/-0.029
Neural Network (L=3, U=500)	0.832+/-0.006	0.873+/-0.004
Random Forest (n_{est}=51)	0.769+/-0.008	0.830+/-0.009
Support Vector Machine (Linear Kernel, c=2500)	0.779+/-0.005	0.816+/-0.008
Support Vector Machine (RBF Kernel, c=100000)	0.803+/-0.003	0.839+/-0.013

Table C.3: Validation F1 scores for classifiers trained from the Mixed PSD data

Model	**Individual Pollution Source Measured**	**Multiple Pollution Source Measured**
K- Nearest Neighbour (1 Nearest Neighbour)	0.814+/-0.011	0.791+/-0.002
Logistic Regression	0.633+/-0.009	0.804+/-0.004
Naive Bayes	0.631+/-0.008	0.763+/-0.028
Neural Network (L,=3, U=500)	0.825+/-0.003	0.852+/-0.014
Random Forest (n_{est}=51)	0.773+/-0.012	0.816+/-0.017
Support Vector Machine (Linear Kernel, c=2500)	0.780+/-0.006	0.814+/-0.008
Support Vector Machine (RBF Kernel, c=100000)	0.798+/-0.002	0.836+/-0.014

Table C.4: Validation F1 scores for classifiers trained from the Mixed count data

Model	**Individual Pollution Source Measured**	**Multiple Pollution Source Measured**
K- Nearest Neighbour (3 Nearest Neighbour)	0.565+/-0.006	0.731+/-0.007
Logistic Regression	0.538+/-0.004	0.746+/-0.003
Naive Bayes	0.555+/-0.013	0.859+/-0.008
Neural Network (L=1, U=250)	0.649+/-0.005	0.903+/-0.012
Random Forest (n_{est}=31)	0.612+/-0.008	0.915+/-0.003
Support Vector Machine (Linear Kernel,c =2500)	0.618+/-0.006	0.898+/-0.002
Support Vector Machine (RBF Kernel, c=1000)	0.633+/-0.006	0.875+/-0.014

Table C.5: Validation F1 scores for classifiers trained from the Mixed MVG random variables

Model	**Bin Values**			**Bin Values and IO**		
Data set used to train classifers	**Individual**	**PSD Mixed**	**Count Mixed**	**Individual**	**PSD Mixed**	**Count Mixed**
kNN	0.791+/-0.004	0.791+/-0.005	0.791+/-0.004	0.854+/-0.003	0.854+/-0.004	0.854+/-0.004
Logistic Regression	0.557+/-0.010	0.677+/-0.006	0.675+/-0.006	0.729+/-0.007	0.786+/-0.004	0.786+/-0.004
Naive Bayes	0.589+/-0.008	0.591+/-0.008	0.589+/-0.008	0.700+/-0.010	0.698+/-0.010	0.698+/-0.008
Neural Network	0.799+/-0.007	0.816+/-0.006	0.814+/-0.006	0.851+/-0.003	0.854+/-0.004	0.858+/-0.005
Random Forest	0.806+/-0.005	0.808+/-0.007	0.805+/-0.006		0.836+/-0.006	0.846+/-0.002
Support Vector Machine (Linear Kernel)	0.811+/-0.004	0.817+/-0.006	0.818+/-0.005	0.854+/-0.002	0.852+/-0.003	0.848+/-0.002
Support Vector Machine (RBF kernel)	0.826+/-0.004	0.827+/-0.006	0.825+/-0.005	0.860+/-0.007	0.861+/-0.005	0.863+/-0.005

Table C.6: Validation results for the source apportionment hybrid models on measured individual pollution source data

Model	**Bin Values**			**Bin Values and IO**		
Data set used to train classifers	**Individual**	**PSD Mixed**	**Count Mixed**	**Individual**	**PSD Mixed**	**Count Mixed**
kNN	0.743+/-0.008	0.745+/-0.009	0.743+/-0.008	0.784+/-0.002	0.784+/-0.002	0.784+/-0.002
Logistic Regression	0.799+/-0.003	0.798+/-0.004	0.801+/-0.005	0.853+/-0.002	0.860+/-0.003	0.861+/-0.003
Naive Bayes	0.673+/-0.009	0.676+/-0.010	0.678+/-0.009	0.777+/-0.010	0.777+/-0.011	0.778+/-0.011
Neural Network	0.750+/-0.005	0.793+/-0.004	0.803+/-0.008	0.863+/-0.013	0.874+/-0.008	0.887+/-0.006
Random Forest	0.756+/-0.005	0.762+/-0.007	0.755+/-0.008		0.823+/-0.010	0.819+/-0.011
Support Vector Machine (Linear Kernel)	0.813+/-0.005	0.808+/-0.005	0.810+/-0.006	0.882+/-0.009	0.892+/-0.007	0.879+/-0.011
Support Vector Machine (RBF kernel)	0.787+/-0.005	0.790+/-0.005	0.791+/-0.005	0.855+/-0.005	0.864+/-0.006	0.851+/-0.003

Table C.7: Validation results for the source apportionment hybrid models on measured multiple pollution source data

Appendix D

Source Apportionment Mixing Algorithms

Algorithm 1: Generating the mixed data set through linear combinations of the PSDs

initialisation: n=1, minimum_composition=0.5, n_components=3 and mixed_PSD_dataset;
while $n = N$ **do**
  initialise mixture_weights to **0**;
  mixed_PSD to **0**;
  for *component in n_components* **do**
    if *component = 1* **then**
      randomly select the major component index and randomly assign a weight to the mixture weights between the minimum composition and 1;
      $rem = 1 - sum(mixture)$;
    else if *component = 2 and component n_components* **then**
      randomly select the next component index and randomly assign a weight to the mixture weights between the 0 and rem;
      $rem = 1 - sum(mixture)$;
    else
      randomly select the last component index and with weights in the mixture is given by rem;
    end
  end
  for *mixture_weight in mixture_weights* **do**
    sampled_PSD = sample a measurement for the corresponding pollution source;
    $mixed\_PSD = mixed\_PSD + (mixtureweight * sampledPSD)$;
  end
  append mixed_PSD to mixed_psd_dataset ;
  $n = n + 1$;
end

Algorithm 2: Generating the mixed data set through linear combinations of the bin count data

initialisation: n=1, minimum_composition=0.5, n_components=3 and mixed_PSD_dataset;

while $n = N$ **do**

  initialise mixture_weights to **0**;

  mixed_COUNTS to **0**;

  for *component in n_components* **do**

    if *component = 1* **then**

      randomly select the major component index and randomly assign a weight to the mixture weights between the minimum composition and 1;

      $rem = 1 - sum(mixture)$;

    else if *component = 2 and component n_components* **then**

      randomly select the next component index and randomly assign a weight to the mixture weights between the 0 and rem;

      $rem = 1 - sum(mixture)$;

    else

      randomly select the last component index and with weights in the mixture is given by rem;

    end

  end

  for *mixture_weight in mixture_weights* **do**

    sampled_COUNTS = sample a measurement for the corresponding pollution source;

    $mixed\_COUNTS =$

    $mixed\_COUNTS + (mixture\_weight * sampled\_COUNTS)$;

    $mixed\_PSD = \frac{mixed\_COUNTS}{sum(mixed\_COUNTS)}$

  end

  append mixed_PSD to mixed_psd_dataset ;

  $n = n + 1$;

end

Algorithm 3: Generating the mixed data set through linear combinations of parameters of the MVG random variables

initialisation: n=1, minimum_composition=0.5, n_components=3 and mixed_PSD_dataset;
while $n = N$ **do**
  initialise mixture_weights to **0**;
  mixed_PSD to **0**;
  for *component in n_components* **do**
    if *component = 1* **then**
      randomly select the major component index and randomly assign a weight to the mixture weights between the minimum composition and 1;
      $rem = 1 - sum(mixture)$;
    else if *component = 2 and component n_components* **then**
      randomly select the next component index and randomly assign a weight to the mixture weights between the 0 and rem;
      $rem = 1 - sum(mixture)$;
    else
      randomly select the last component index and with weights in the mixture is given by rem;
    end
  end
  for *mixture_weight in mixture_weights* **do**
    $\mu_{mixed} = \mu_{mixed} + (mixture\_weight * \mu_{component})$;
    $\Sigma_{mixed} = \Sigma_{mixed} + (mixture\_weight * \Sigma_{component})$;
    $X \sim \mathcal{N}(\mu, \Sigma^2)$;
  end
  Sample mixed_PSD from X and append mixed_PSD to mixed_psd_dataset ;
  $n = n + 1$;
end

Appendix E

Monitoring Air Quality in London

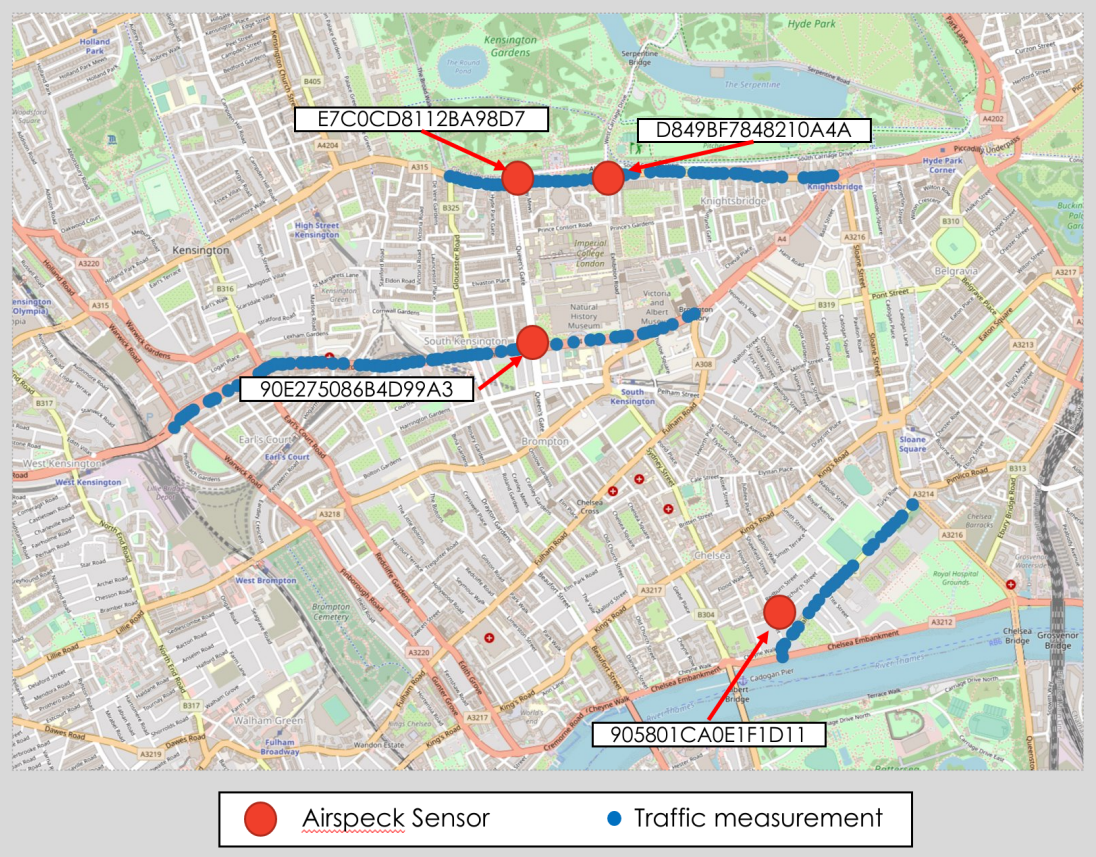


Figure E.1: Map of London showing Airspeck-S sensor locations and HERE traffic monitoring locations

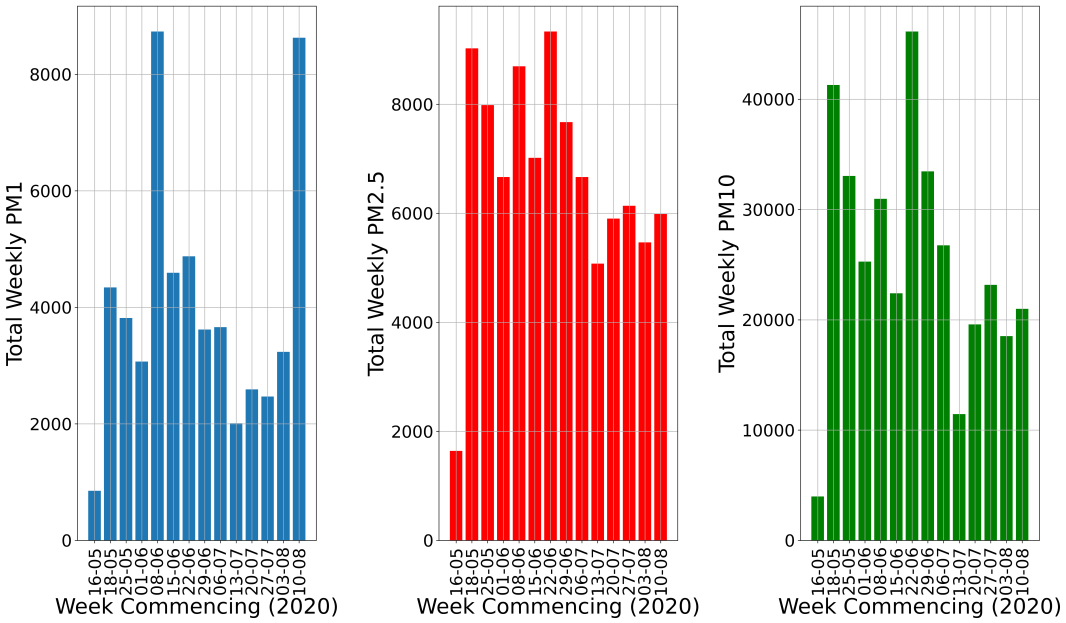


Figure E.2: Weekly cumulative PM_1, $PM_{2.5}$ and PM_{10} for Sensor 90E275086B4D99A3

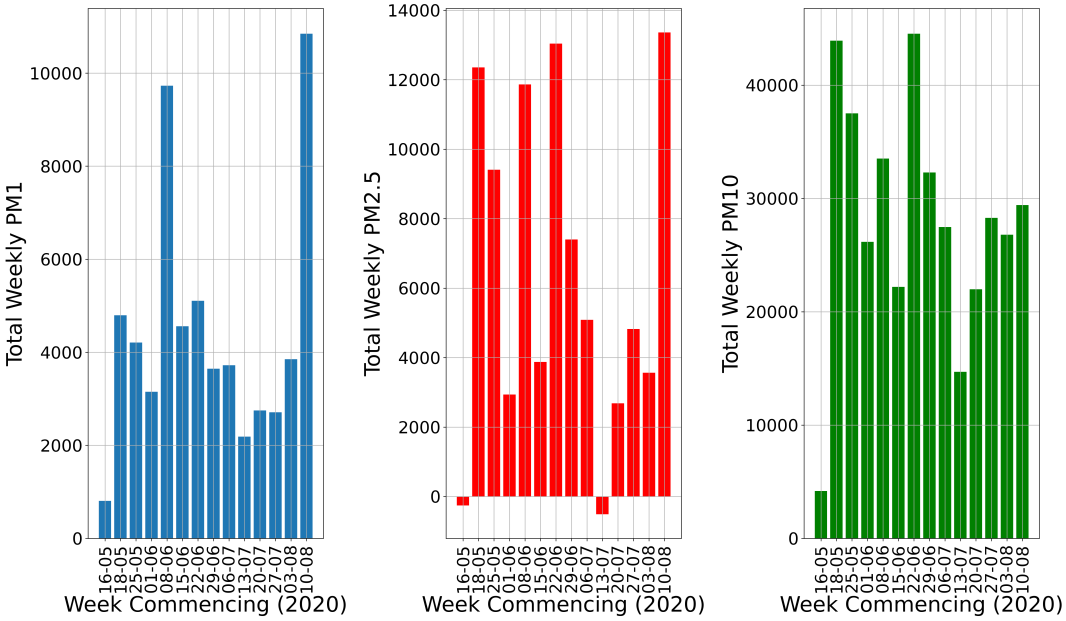


Figure E.3: Weekly cumulative PM_1, $PM_{2.5}$ and PM_{10} for Sensor D849BF7848210A4A

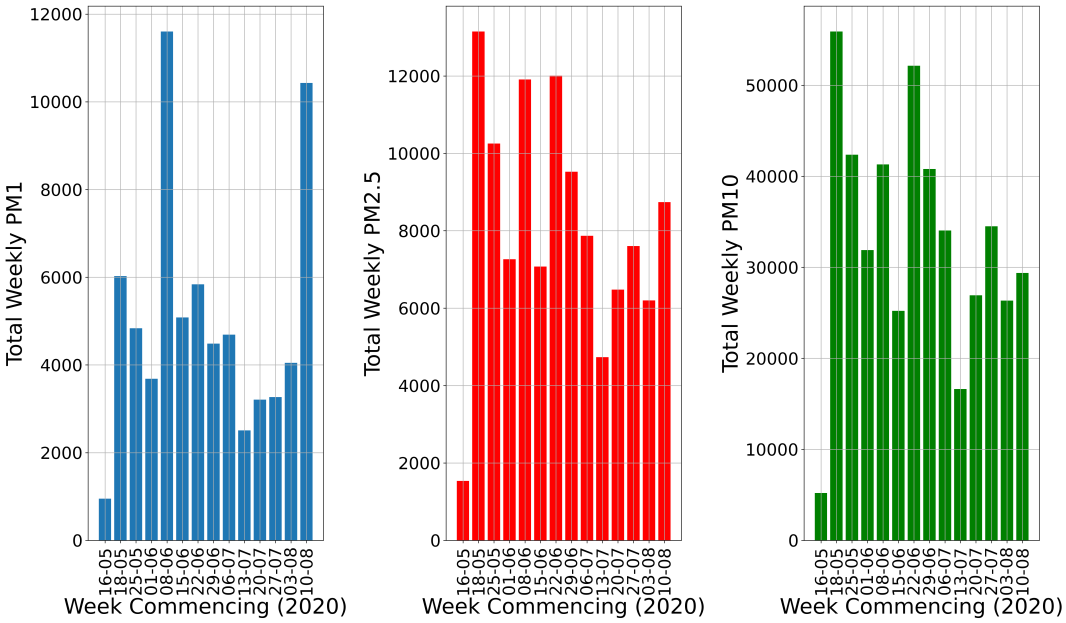


Figure E.4: Weekly cumulative PM_1, $PM_{2.5}$ and PM_{10} for Sensor E7C0CD8112BA98D7

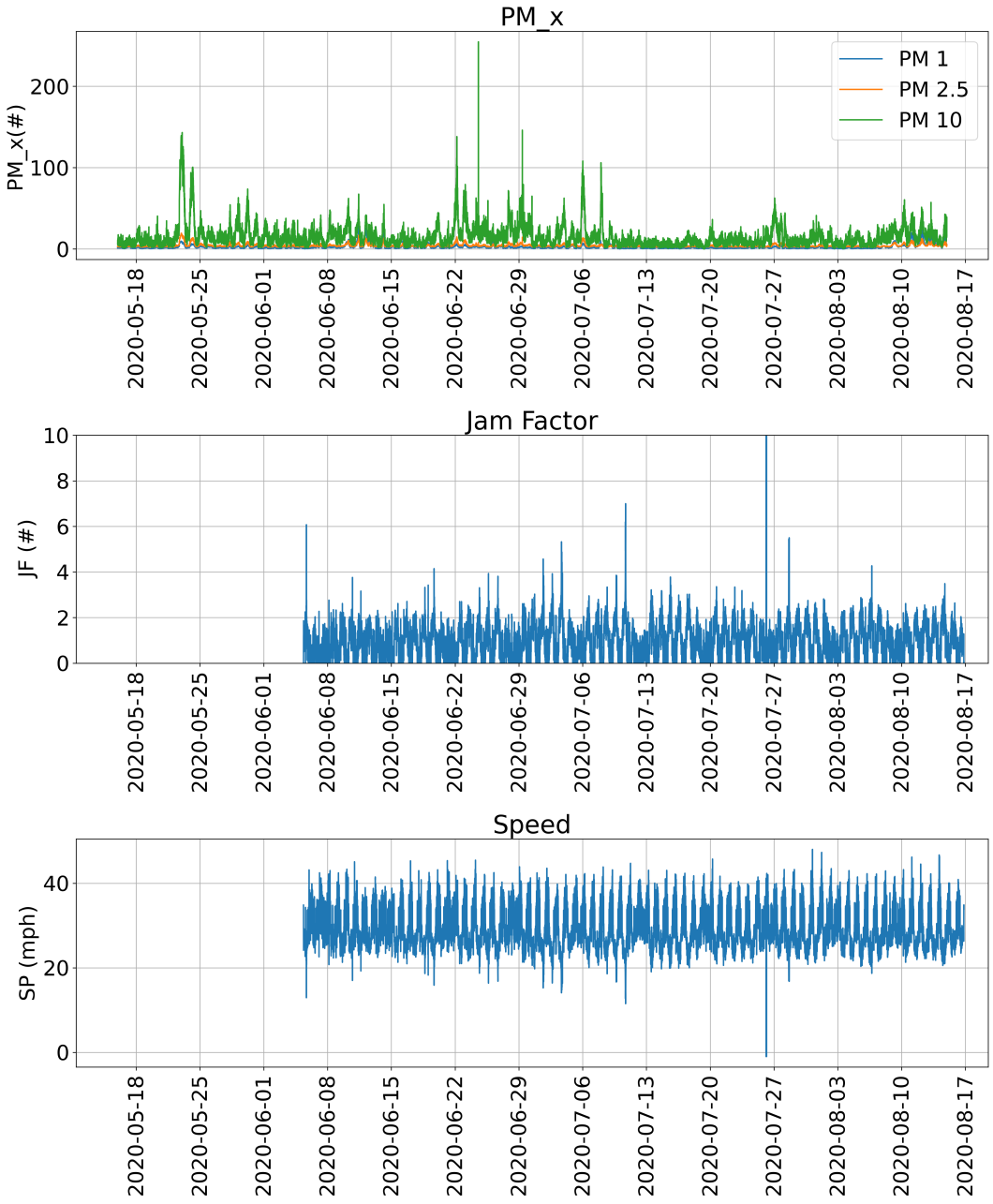


Figure E.5: PM_1, $PM_{2.5}$ and PM_{10} and traffic data for Sensor 90E275086B4D99A3

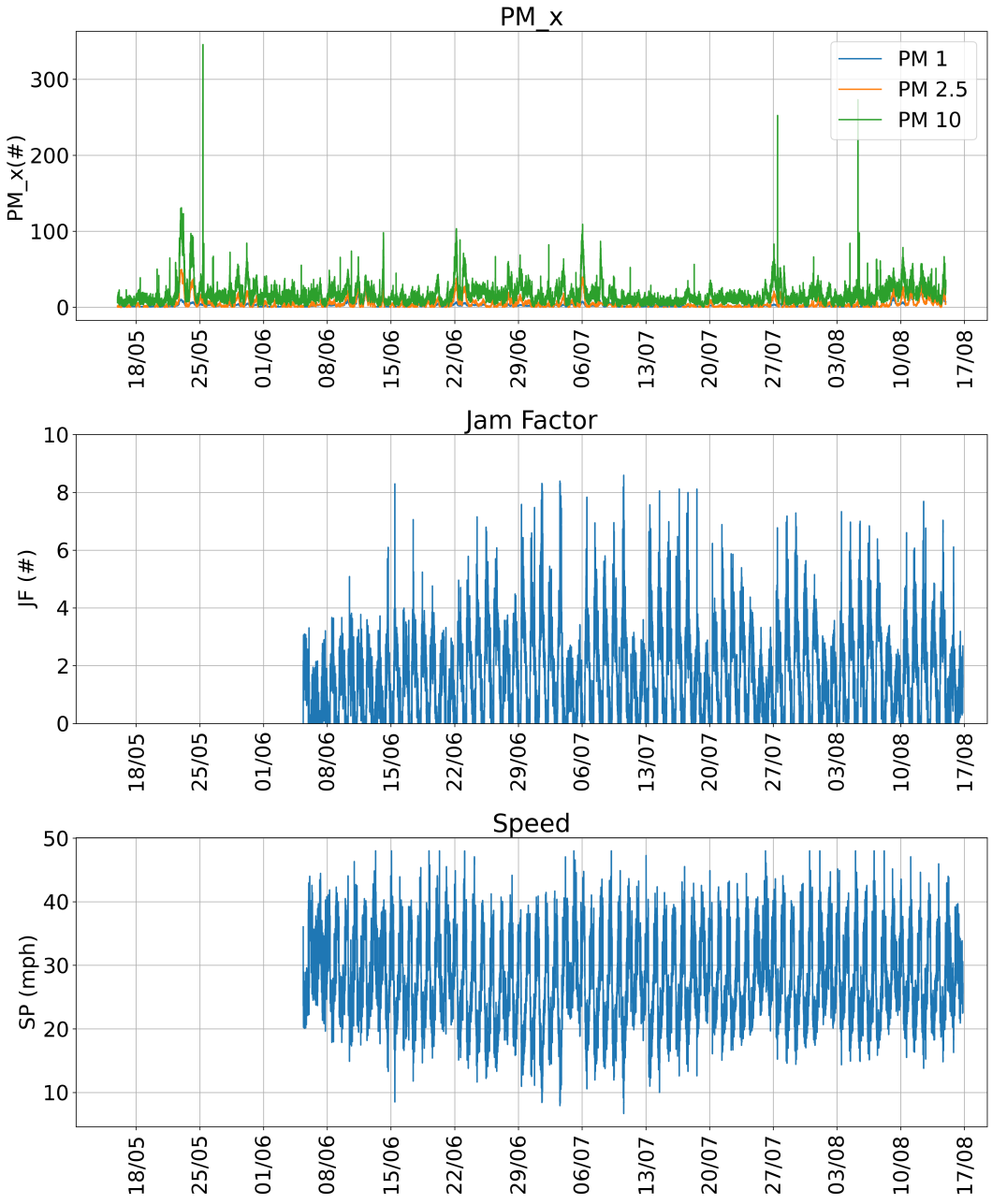


Figure E.6: PM_1, $PM_{2.5}$ and PM_{10} and traffic data for Sensor D849BF7848210A4A

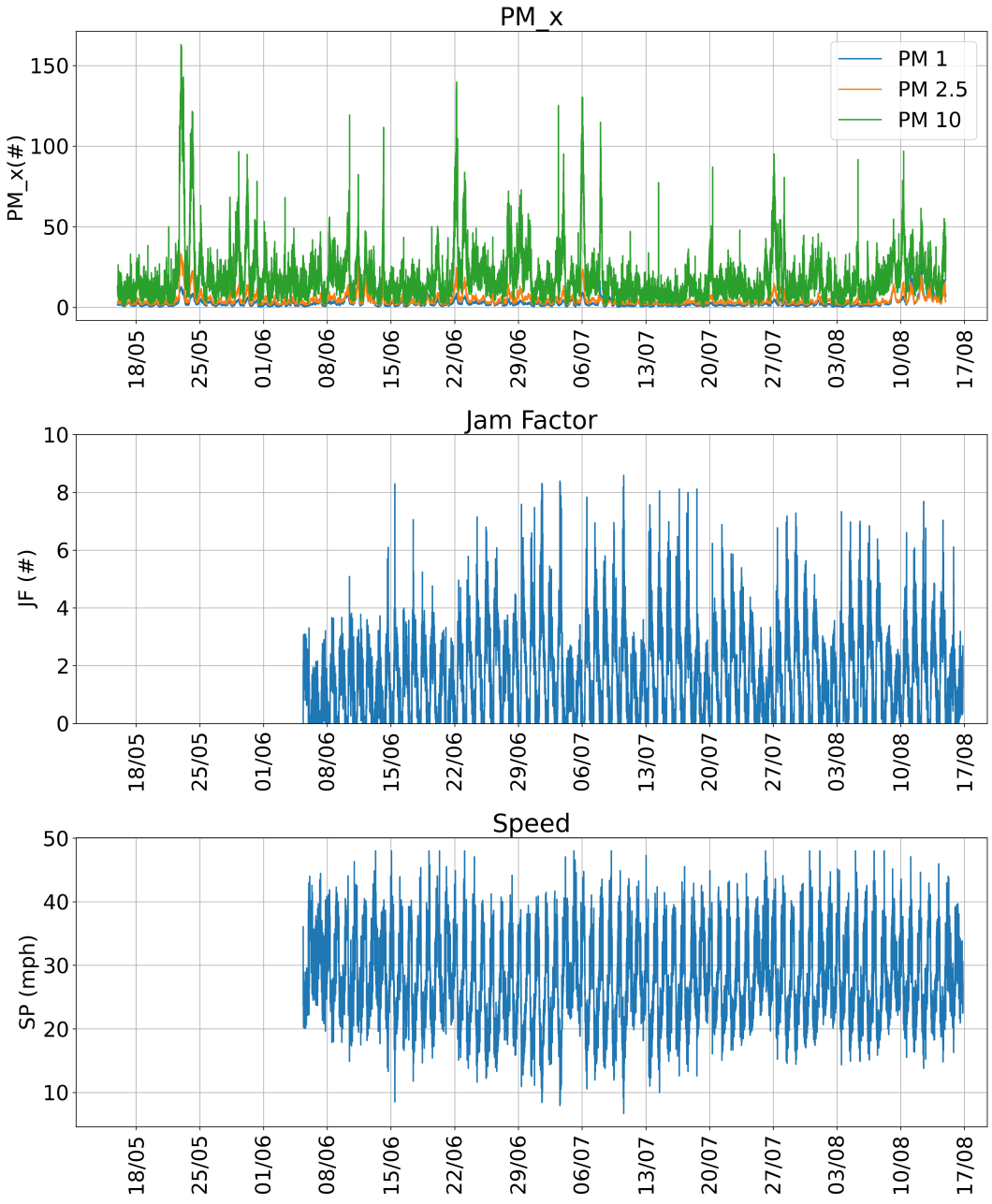


Figure E.7: PM_1, $PM_{2.5}$ and PM_{10} and traffic data for Sensor E7C0CD8112BA98D7