Faculty of Engineering & Information Technology

Research Project

Machine Learning for Air Quality Prediction in Sydney

Student: Anastasia SHELYUKHINA (13094824)

Supervisor: Adriana-Simona MIHAITA

Table of Contents

Table of Figures

Introduction

The air pollution problem is a growing concern in the urban areas in the face of climate change, industrialisation and the progression of the urban population. The UN estimates the share of the world's urban population to be 68% by 2050 versus 30% in 1950 [1]. Ambient air pollution poses a significant threat to human health. The 2015 Global Burden of Diseases Study indicated that the exposure to PM2.5 is the fifth-ranking global mortality risk factor and a cause to 4.2 million deaths. [2] built a data integration model fitted with a Bayesian hierarchical framework to estimate global exposure to air pollution. They estimated that 92% of the world's population reside in the areas where the annual average concentration of particulate matter is higher than advised by the World Health Organisation (WHO). Premature deaths, hospitalisations, emergency ambulance dispatches and other health risks associated with ambient air pollution result in economic losses which, as a result, drive decision-makers and researchers to investigate air quality.

Various models were built by academics to study the association of air pollution in Sydney with daily hospital admissions and daily mortality [3]. Single and multiple pollutant models and time-series analysis based on Poisson regression showed an association between air pollution and increased hospitalisation rate for respiratory and heart diseases, mainly due to the elevated level of nitrogen dioxide. The analysis of particulate and ozone pollution levels showed a high correlation with all-cause mortality and cardiovascular mortality, particularly for ozone [3]. Morgan et al. stressed that even though their model cannot predict how many days, months, years of life lost the air pollution in Sydney represents, the associations between air pollution and daily mortality and hospitalisation are linked. An immediate call to action to reduce the dependence on motor-vehicle transportation as the primary contributor to air pollution was clearly announced.

Statistical modelling was also used by [4] to understand the association between mortality and air pollution in Sydney. The researchers developed the model with a larger time-span dataset, covering the years from 1997 to 2004. They found out that over eight years (1997-2004), 24-hour city-wide PM10 concentrations exceeded the 99th percentile 52 times, resulting in 48 events attributable to bushfire, 6 to dust storm and two affected by both [4]. By using logistic regression adjusted for influenza epidemics, same day and lagged temperature and humidity, researchers demonstrated an increase of 5% in non-accidental mortality at a lag of 1 day for smoke events (bushfires) and 15% for dust events at a delay of 3 days. Worth mentioning that the same-day temperature is an essential contributor during bushfire events that should be taken into consideration.

Broome at al. [5] used the BenMAP-CE software tool to estimate the health impact and economic benefits of an air pollution reduction in Sydney. The researchers proved that the decrease in PM2.5 level by 10% would result in 650 fewer premature deaths (95% CI 4230 - 850), an increase of 3,500 life-years (95% CI 2300 - 4600) and a drop in hospitalisations by 700 patients (95% CI 450 - 930) due to respiratory and cardiovascular diseases [5].

Salimi et al. explained the association between emergency ambulance dispatches (EAD) and PM2.5 concentrations using generalised linear quasi-Poisson regression model. They concluded that an increase in PM2.5 concentration was associated with EADs for heart and breathing problems [6].

Authors in [7] concluded that within a city, localised differences in the concentration of particulate matter could occur. Their focus was predominantly on the impact of urban forestry and greenspace on the ambient air pollutants levels. Even when traffic is taken into consideration, the areas with a higher concentration of greenspace have lower concentrations of particulate matter [7]. This is why several authors have proposed optimisation techniques to reduce traffic congestion and improve traffic control ([18], [20], [23], [35], [38], [39]) or even more proposed advanced techniques for traffic prediction modelling to estimate the next stage of the traffic congestion in the future ([27]) and prepare for the uptake of more connected and autonomous vehicles that would help reduce emissions and avoid traffic collisions ([30], [31], [32]).

The recent study of Wadlow et al. compared concentrations of PM2.5 near roads and intersections with the measurements of air quality stations over a four-day experiment in February 2017. There is a statistically significant difference (2x higher) between PM2.5 measurements on Anzac Parade and nearest air quality stations in Earlwood and Rozelle. Average PM2.5 concentrations along Anzac Parade were 30% higher than those on perpendicular residential and walking streets. Over the experiment that lasted four days, the researchers found out the concentration of PM2.5 varies depending on the time of the day with the highest record in the morning. Traffic lights and intersections create hotspots for PM2.5 concentrations [8]. Additionally, Rivas et al. in their income inequality study In London established that concentrations of various pollutants are higher in the morning (+13% than in the afternoon, +43% than in the evening) [9].

The literature review of the academic papers related to air quality showed that research methodology mostly comprised of traditional statistical modelling. Researchers found an association between events and possible consequences such as an increase in hospitalisation, mortality or emergency dispatches ([10], [4], [6]). Others conducted experiments to evaluate the accuracy of mobile devices to measure air pollution within cities [8]. Even though the findings of these researchers are significant in attracting stakeholder's attention to the problem of air pollution, the development of the prediction model would have made their cases stronger. With data mining techniques becoming more accessible, we can find hidden patterns and build reliable prediction models based on historical data to get insights on how severe air pollution problem is. For instance, Li et al. used a spatiotemporal deep learning model to predict air quality in China. Proposed STDL-based model could predict more accurately the air quality of all stations simultaneously and showed temporal stability in all seasons compared to traditional statistical methods, time-series models and ANNs and SVRs [11].

Neural networks proved to be an effective technique for regression modelling of real-time air pollution health indexes (APHI) mapping in Hamilton, Canada. By combining mobile and stationary air pollution data, meteorological records, land use characteristics, and traffic information [12] predicted the unseen conditions for PM2.5 with a Pearson's r of 0.88, a coefficient of determination of 0.78, and an RMSE of 3.5 mg/m3. The predictions for nitrogen dioxide were less accurate: with a Pearson's r of 0.59, a coefficient of determination of 0.34, and an RMSE of 10 ppb.

Out of four objectives of the study of [13] on smart mobile pollution monitoring in France, the findings associated with data-driven predictions are of great importance to our investigation. The scientists established that decision trees and neural networks algorithms are capable of accurately predicting ambient nitrogen dioxide concentrations measured with portable devices (81% accuracy). Weather conditions, noise, day profile and mobility are sensitive features for any air quality prediction model. Other techniques have been also applied in various set-ups that include a combination of advanced natural language models, agent-based simulations ([21],[22], [36], [37]).

Research Aims and Objectives

In recent years, researchers investigated the reliability of mobile air quality data that can improve the decision making of citizens ([12], [13]). We estimate that the state-of-the-art prediction models built using emerging machine learning techniques will serve as a solid ground for future research when widespread mobile air quality data become available in large cities such as Sydney.

Aim

The aim is to build innovating air quality prediction models for the city of Sydney that will combine air quality monitoring measurements with ancillary data such as traffic, holidays and weather conditions.

Objectives

1) build baseline air quality prediction model;
2) test modern machine learning regressors using available features;
3) compare the models to select the most important features which will influence the prediction results and the most effective regression model;
4) analyse the air quality in Sydney, identify pollution patterns and problematic areas.

Research Methodology

A data-driven approach was adopted to analyse the air quality in Sydney. Data mining techniques were used to construct the features and machine learning regression algorithms - to build air quality prediction models.

Data

The available data sets for this study consist of:

- **Air pollution data**: 11 years of collected data from 18 fixed monitoring stations in the city of Sydney, Australia with hourly granularity. The air quality concentrations received for this study are PM2.5, PM10, NEPH.
- **Meteorology**: humidity, temperature, wind speed, wind direction.
- **Traffic counts** - permanent traffic counts scattered around permanent count locations in Sydney.
- **Holidays** - school vacation and bank holidays.

Air pollution and meteorology data was provided by NSW Smart Sensing Network (NSSN). Traffic data was obtained from open source NSW Transport (Roads and Maritime Services) permanent traffic volume counters which have been a rich source of valuable information for various studies ([16], [19], [24], [25], [26], [29]).

Data Profiling

Exploratory data analysis (EDA) of air pollution data and ancillary data such as meteorology and traffic was conducted. The goal of EDA was to identify anomalies and outliers, understand the distribution of air pollution indicators for each monitoring station, identify trends and seasonality patterns, check data quality: percentage of negative values, zeros and missing values.

Predictive Modelling

A linear regression model was used as a baseline. This straightforward model that doesn't require any hyper-parameter tuning and is easy to interpret and understand. Various machine learning models were trained and tested to predict the air quality of the unseen dataset.

1. **Gradient Boosting Regressor (XGBoost)**. A popular technique in the machine learning field, frequently used by Kaggle competition winners. The model is characterised by high model performance and execution speed [14].
2. **Random Forest.** User-friendly, easy to explain the insights, robust to overfitting [15].

Randomised search and **four-fold time series cross-validation** was used for hyper-parameter tuning. The models were trained on data points that were registered before the observations in the test set. Therefore, to predict future data, the model has to learn only from the data points in the past. Figure 1 illustrates the process of time series split. The exercise was repeated four times (4-fold cross-validation).

Fold 1: Train on 2013, validate on 2014

Fold 2: Train on 2013-2014, validate on 2015

Fold 3: Train on 2013-2015, validate on 2016

Fold 4: Train on 2013-2016, validate on 2017.

Using 6 years of data (2013-2018), we managed to use 5 full years of hourly data (2013-2017) to select the hyperparameters. 2018 data was set aside for models' evaluation only.

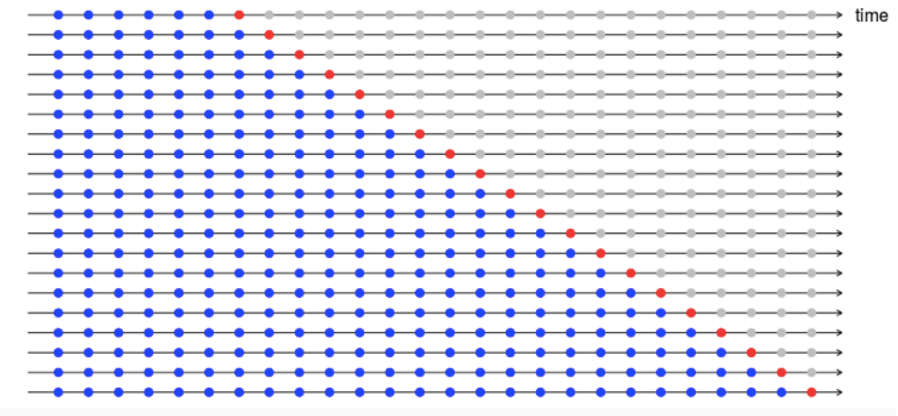


Figure 1: Rolling forward time-series cross-validation concept

Cross-validation was embedded into randomized search which allowed us to select hyper-parameters and also avoid data leakage from the test set to the train set.

Model Evaluation

Following metrics were used to evaluate the performance of the predictors.

1. Mean absolute percentage error (MAPE)

$$MAPE = \frac{100\%}{n}\sum_{t=0}^{n} |\frac{A_t - F_t}{A_t}|$$

A_t is the actual value, F_t is the predicted value, n is the number of data points. This error shows accuracy as a percentage and is used frequently as loss function in regression models.

2. Mean square error (RMSE);

$$MSE= \frac{1}{n}\sum_{t=0}^{n}(A_t - F_t)^2$$

3. Symmetric mean absolute percentage error (SMAPE)

$$SMAPE = \frac{100\%}{n}\sum_{t=0}^{n}\frac{|F_t - A_{t|}}{|A_t| + |F_t|}$$

4. R-squared (coefficient of determination)

$$\mathrm{R}^2 = 1 - \frac{\sum_{t=0}^{n}(A_t - F_t)^2}{\sum_t(A_t - \overline{A_t})^2}, \text{ where } \overline{A_t} = \frac{1}{n}\sum_{t=1}^{n} A_t$$

Exploratory Data Analysis

Air Quality Indicators

Three air quality indicators have been studied: PM2.5, PM10 and NEPH. According to the National Air Quality Standards, reporting standards for the particles (PM2.5 and PM10) are set nationally in Australia as shown in the table below:

Pollutant	*Averaging period*	*Standard*
PM2.5	24-hour	25 µg/m3
	Annual	8 µg/m3
PM10	24-hour	50 µg/m3
	Annual	25 µg/m3

Figure 2: The national standards of air quality indicators for particulate matter

These standard limits have been used for assessing the daily and yearly evolution of our most important air quality concentrations PM2.5 and PM10 used in this study and represent important thresholds for anomaly detection.

Anomaly detection

Time series of air quality indicators were visualised for each air quality station. When we looked at the stations that operated active since the year 2008, we have noticed a significant outlier at the end of 2009 that disturbed the visualisation of the hourly and daily trends (see Figure 3).

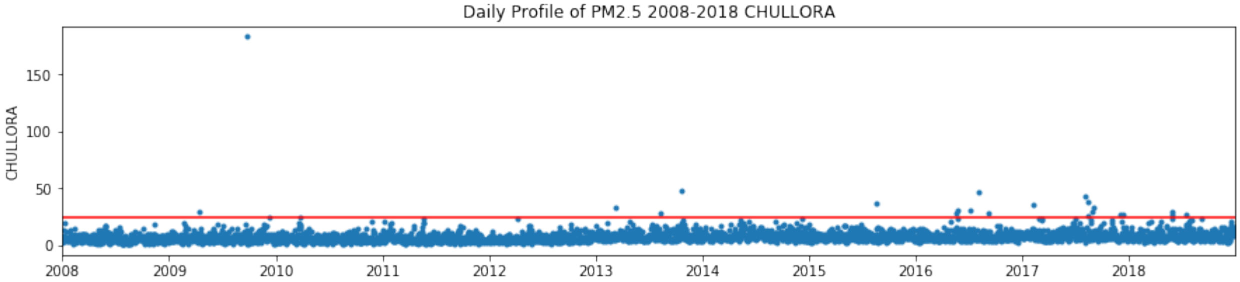


Figure 3: Hourly trend of PM2.5 in Chullora over 2008-2018 with outliers

We found out that there was a dust storm (Eastern Australian dust storm) in 2009 that hit NSW on September 23rd (Figure 4-5 represents the PM10 and PM2.5 recorded for the detected period on 5 of the most important air quality stations around Sydney). Such an event is considered to be an infrequent phenomenon with very rare repeatability (once in 11 years) as the maximum limits that have been reached were 2,000µg/m3 for PM2.5 and 14,000µg/m3 for PM10 at Liverpool. For further exploration, this date was removed from the dataset.

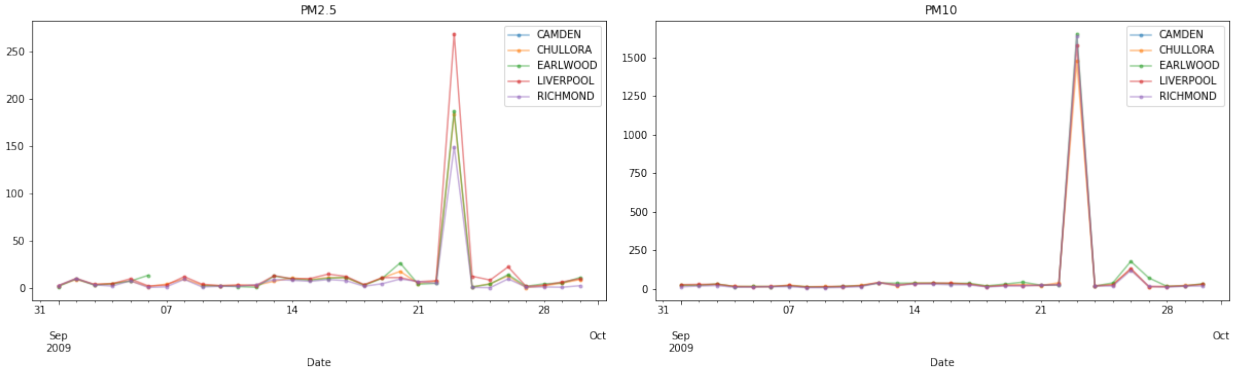


Figure 4: PM2.5 and PM10 trend in September 2009 - anomaly detection

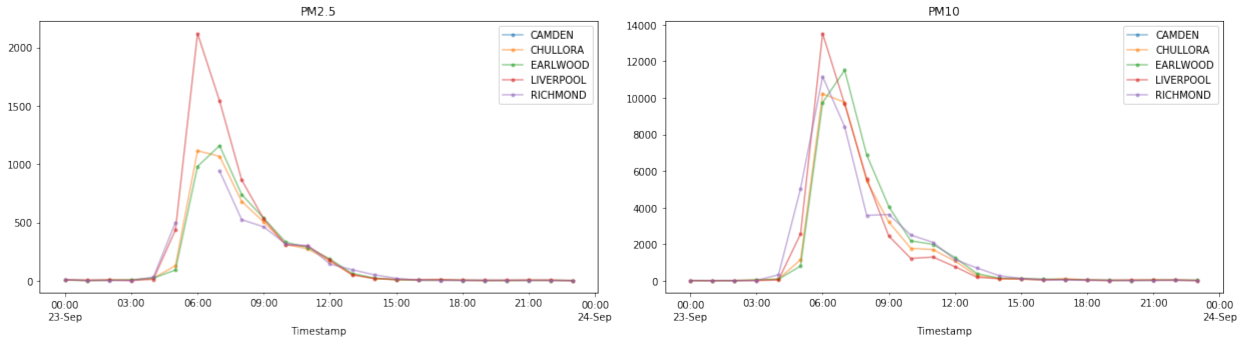


Figure 5: PM2.5 and PM10 trend on 23rd of September 2009 - anomaly detection

Station Selection

Based on the data quality assessment, five air quality stations were selected for this data exploration: **Camden**, **Chullora**, **Earlwood**, **Liverpool** and **Richmond**. The period of time for analysis was also reduced to 2013 - 2018 to include Camden, which started recording data in 2013 (Figure 6 showcases the total number of available data points for each air quality stations assessed for both air quality and weather indicators availability which led to the time range selection).

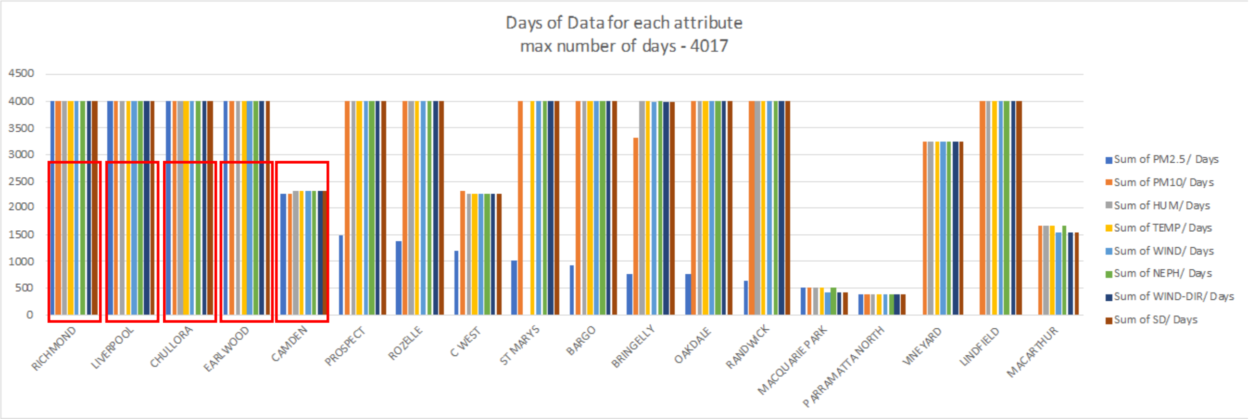


Figure 6: Days of data per station per attribute

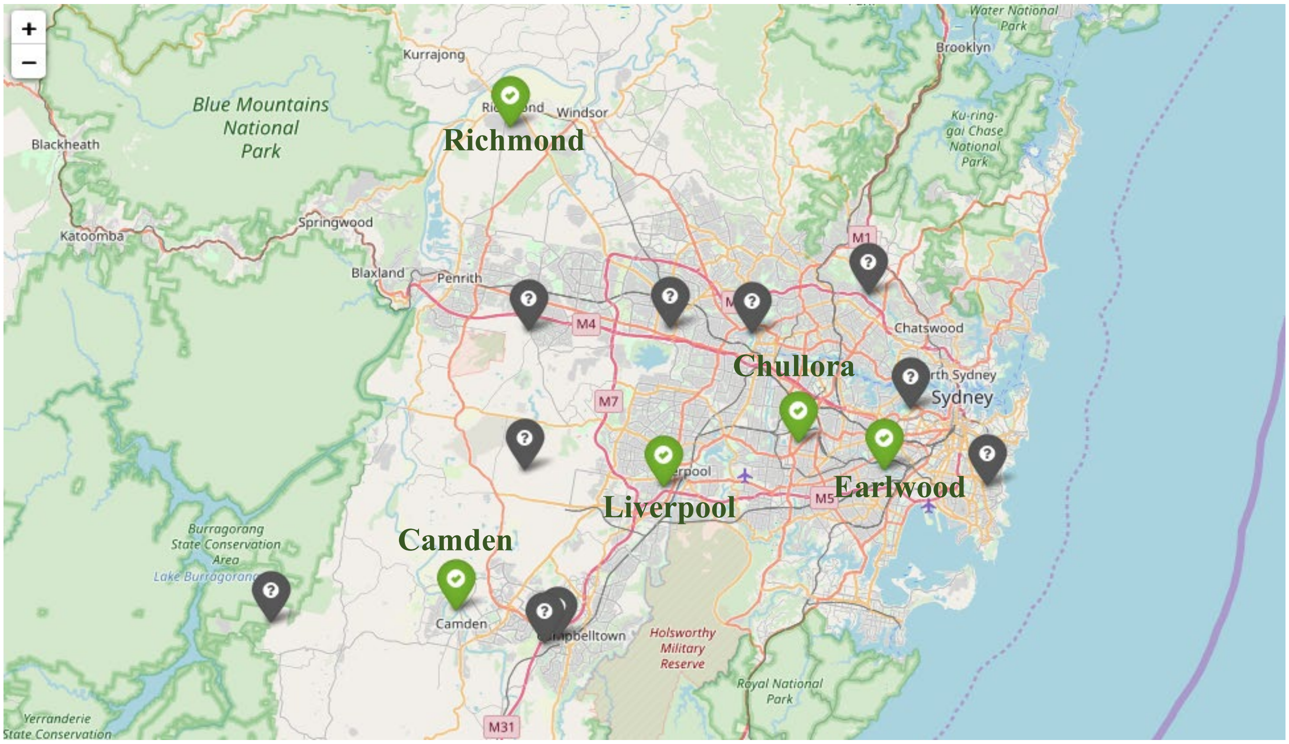


Figure 7: Location of Air Quality Monitoring Stations in the city of Sydney

In addition to data quality, the location of the station was taken into consideration. Figure 7 shows the location of all air quality stations across the city of Sydney. **Chullora** and **Earlwood** Air Quality Stations (AQS) are the closest stations to the CBD. Earlwood station is close to a city park and the river. At the same time, the station is less than 50 meters from the road which makes this station non-conforming. Chullora site is non-conforming as well because the trees have grown nearby since the establishment of the station.

Camden and **Richmond** AQS are closer to the Blue Mountains in New South Wales. Camden site is located within an Aerodrome. Richmond - on the campus of the University of Western Sydney. **Liverpool** site is a city station and can be influenced by the daily traffic of commuters going to/returning from work. By selecting all these stations, we ensure that all parts of the city of Sydney are analyzed, which may have different air pollution behavior.

Daily Profiles

The 23rd of September 2009 was dropped to remove the noise from the daily data. As represented in Figure 8, we can see that over 11 years of data collection, the daily average of PM2.5 and PM10 exceeded the national standard of 20µg/m3 and 50µg/m3 a few times. For instance, at Richmond station, the exceedance was registered 34 times which accumulates to more than one month of abnormal level of particulate matter concentrations. The number of exceedances of PM10 is similar to PM2.5. The station with the highest number of such events is Chullora (33 days). PM10 has a more apparent seasonality than PM2.5. It is especially visible for the Camden site (Figure 8).

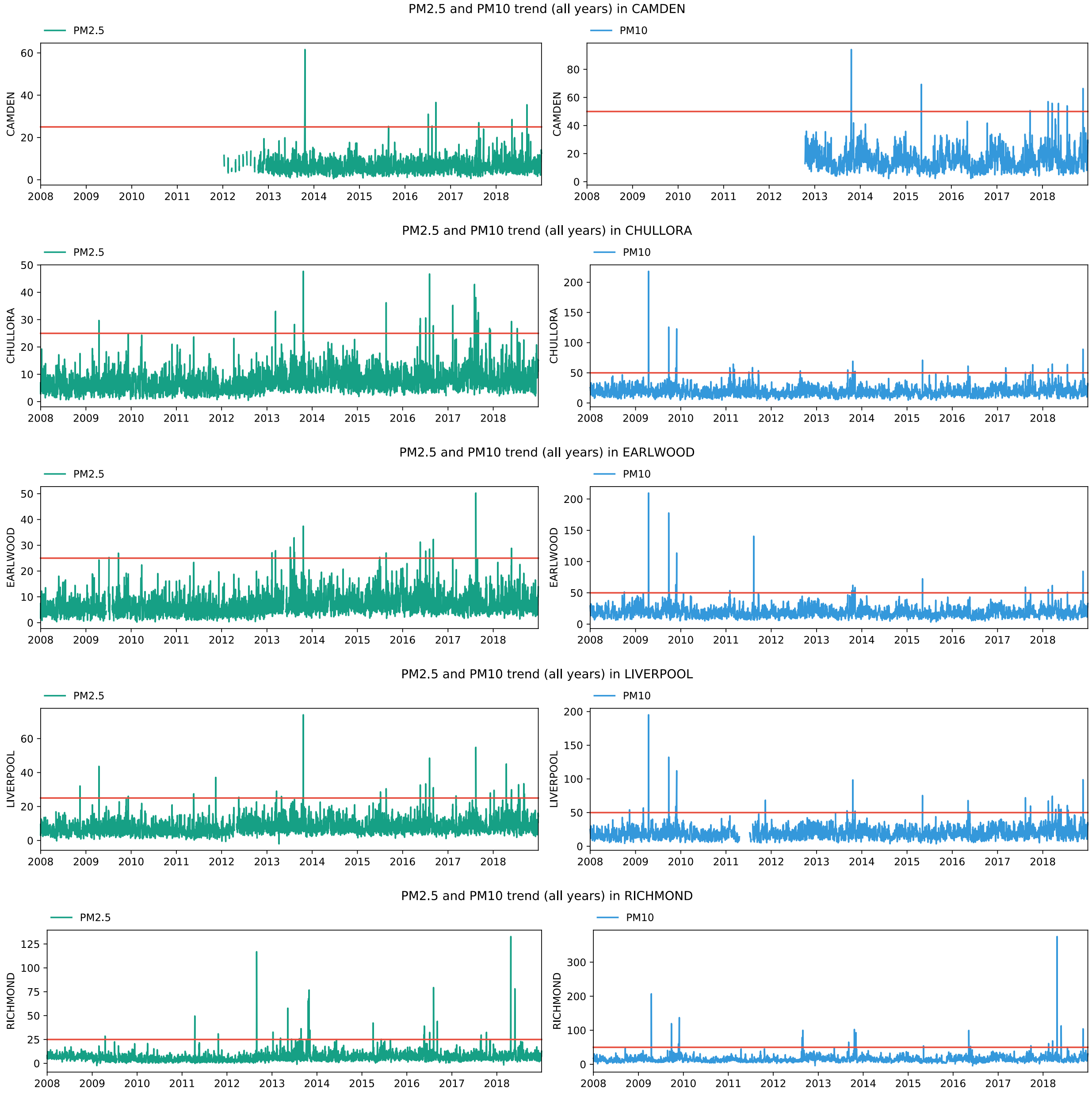


Figure 8: Daily means of particulate matter over 2008-2018 by AQ station with the national daily standard

Yearly Profiles

Overall, when analysing the yearly concentrations, we observed that those of PM2.5 increased by **38%**, of PM10 by **23%** and of NEPH by **15%** since 2008.

As one-day anomaly doesn't create excessive noise for yearly means, the anomaly event was kept for analysis. Figure 9 depicts the annual trends of PM2.5, PM10 and NEPH (visibility).

Chosen air quality sites behave differently one from another, with Liverpool and Chullora exceeding the annual standard for PM2.5 almost every year since 2012. Annual standard of PM10 was exceeded once in 11 years - due to the dust storm in 2009. In 2018, PM10 showed a growing trend and Liverpool site approached the annual national standard.

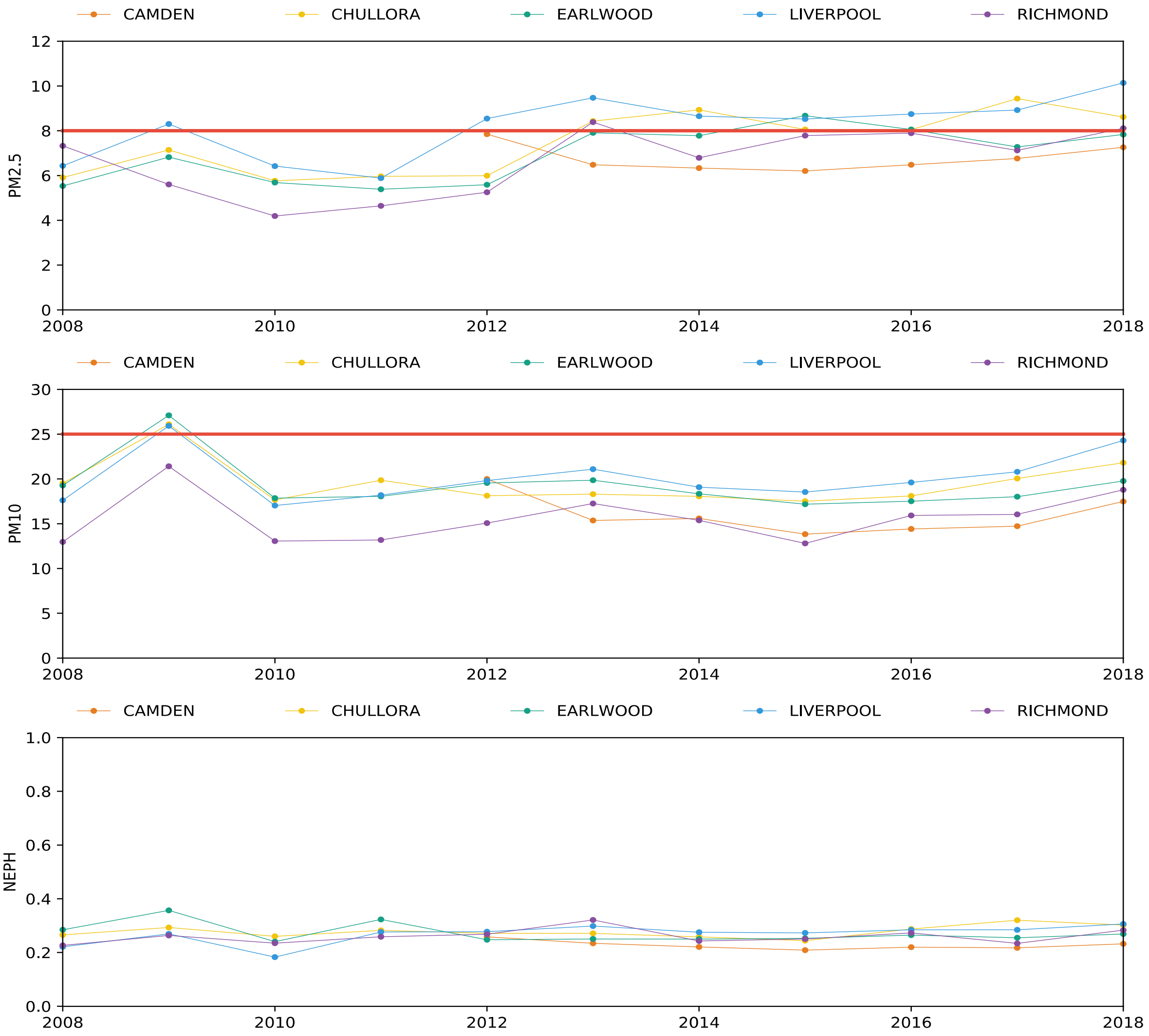
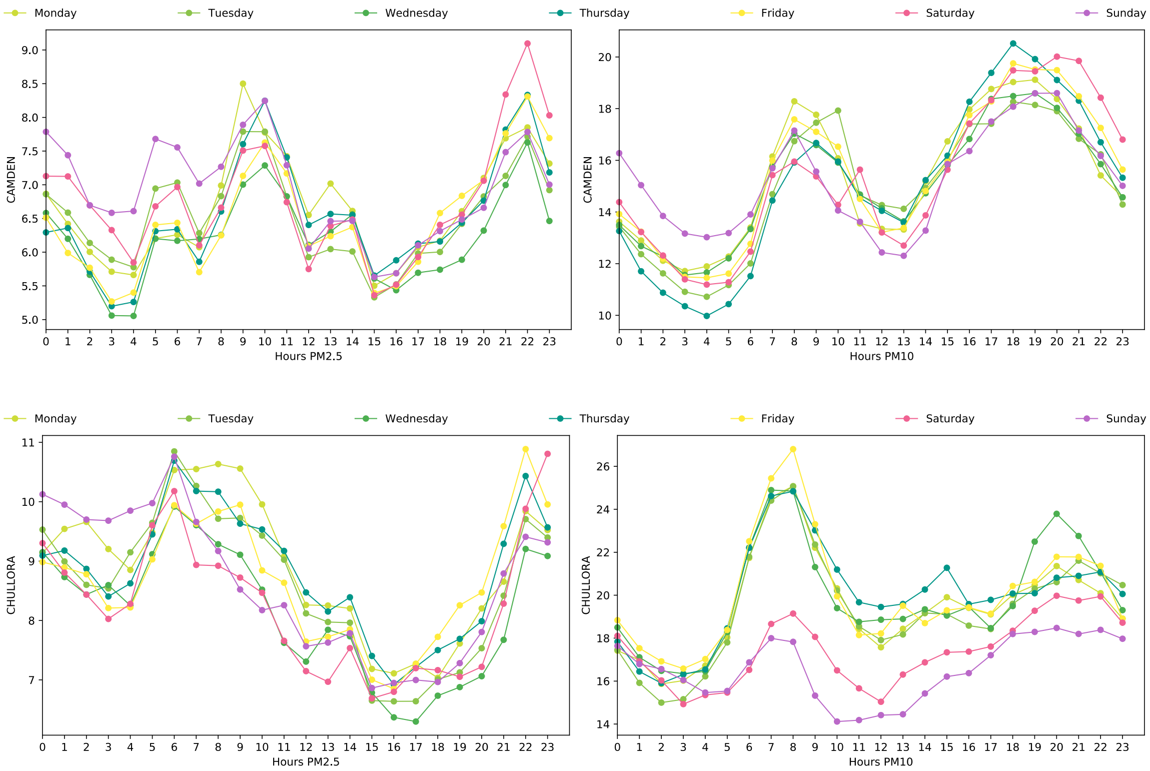


Figure 9: Annual means of air quality indicators by AQ station with the annual national standard

Hourly Profile by day of the week

Based on all historical data sets used for this study, we have also analysed the daily trends of all pollutants. This has been done to analyse the influence of other daily activities of citizens on the air pollution concentrations. PM2.5 tends to fluctuate over 24 hours, independently of the day of the week. The time of the highest value varies for each AQS, but generally, PM2.5 reaches its peak around 9:00 AM in Camden and 5-6 AM at other sites; also, there is a high increase of PM2.5 concentrations observed for all stations which starts around 8 PM in the evening and reaches a high peak around 10 PM. The concentration of PM2.5 is higher on the weekdays except for Richmond, where Sunday seems to be the most polluted day of the week. Weekend pollution also exceeds weekday pollution at night at all sites (Figure 10).



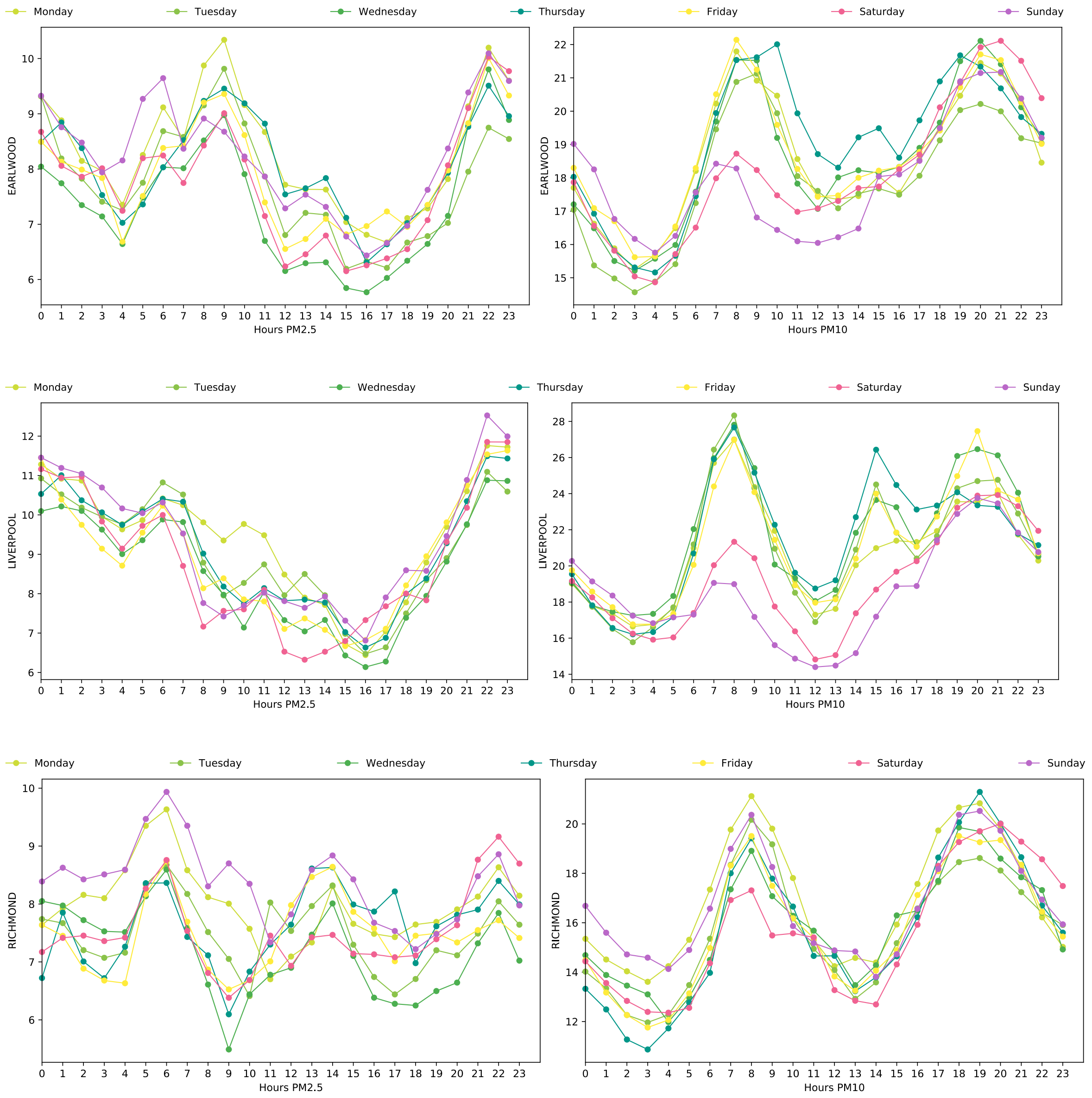


Figure 10: Daily profiles of particulate matter by day of the week by station (average of all years 2008-2018)

PM10 shows distinctively different daily behaviour for weekdays and weekends except for Camden and Richmond, where the trend is the same every day of the week. PM10 has two peaks during the day: in the morning around 8 AM and in the evening around 7 PM. Pollution on the weekend is usually lower than on the weekday for all stations, but the peaks are at the same time as for weekdays.

Distribution

In order to evaluate the quality of our data sets, we have analysed the distribution of the air quality indicators (see left figures in Figure 11) which appears to be normal distributed and right-skewed. Density plot of PM2.5 looks bimodal due to a large number of missing values (replaced by zeros to visualise distribution), negative values and zero values in the original dataset. The Empirical Cumulative Distribution Function (ECDF) shows that most of the data point for PM2.5 are values below 25µg/m3 and for PM10 - below 50 µg/m3 (see right figures in Figure 11).

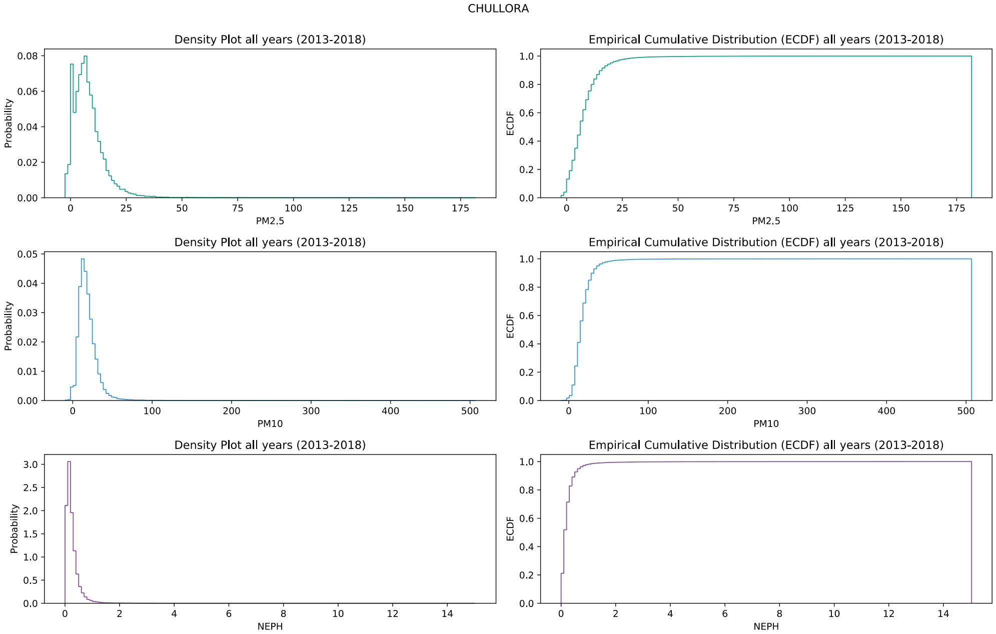


Figure 11: PDF and ECDF of air quality indicators at Chullora station (all years together)

Data Quality

To better understand the data sets, we looked at the data quality by air pollution indicator, by year andby each air quality station (see Figure 12). Data for PM2.5 has a large number of missing values and negatives across all AQS (sometimes reaching event 14-15%). Missing values are distributed randomly over the years. Negative values are considered as errors, as the concentration of particulate matter cannot be negative. We cleaned and removed negative values from the dataset before we started building the prediction models.

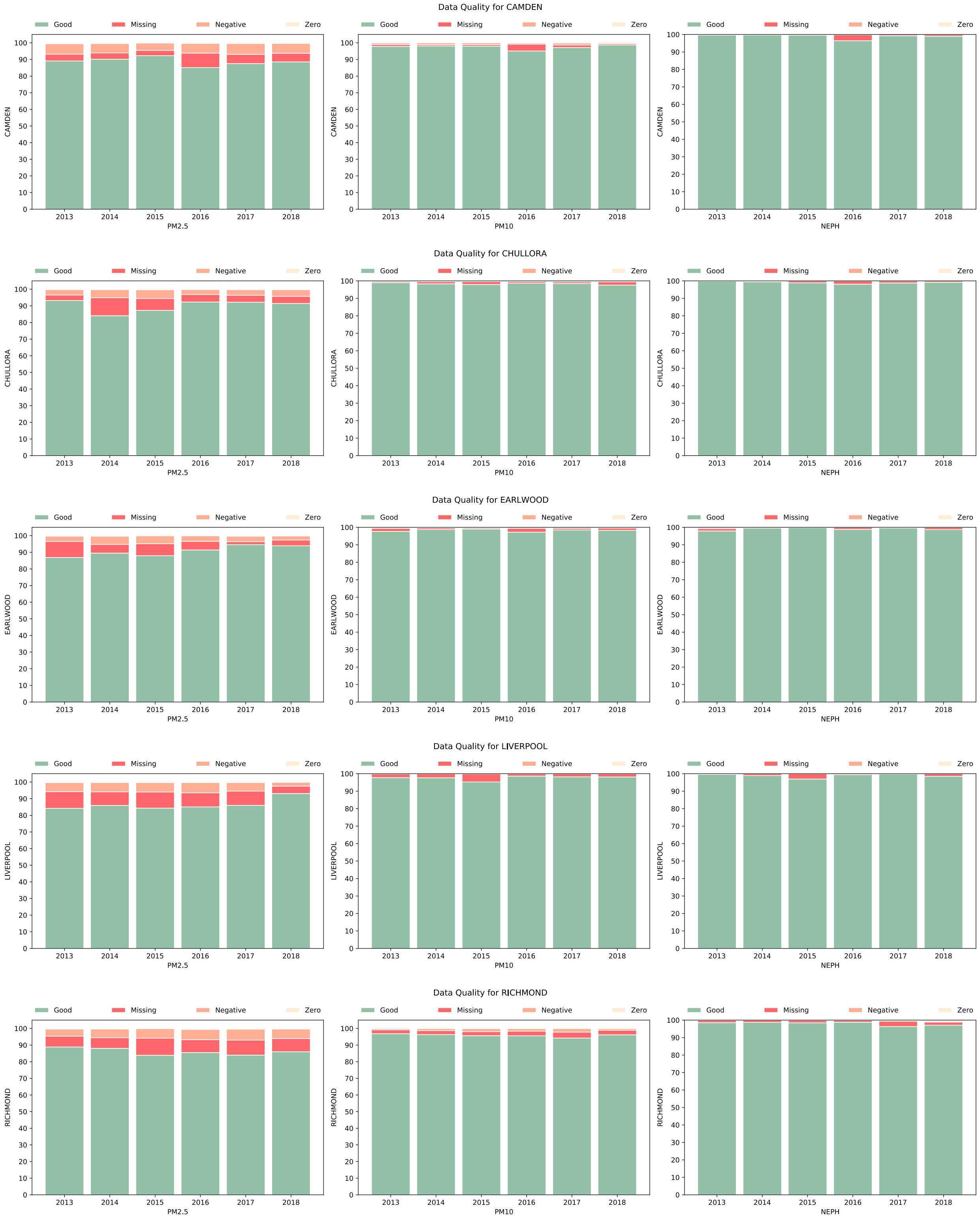


Figure 12: Data quality of air quality indicators (missing values, zeros, negatives) by station by year

Meteorology Data

When studying air pollution, weather conditions should be taken into consideration. Meteorology can have a significant impact on air pollution due to various external factors which can help accumulate or disperse the air pollution. For example, when the weather is calm (no wind), the pollutants cannot disperse resulting in the pollution build-up. In windy conditions, pollutants disperse promptly, therefore causing a positive effect on air quality. Weather data can also help with predicting air pollution and we will use it as some of the most important features in the model construction.

Four meteorology indicators were used for this research:

- temperature in Celsius degrees,
- humidity in percentages (0-100%),
- wind direction in degrees (0-360 degrees) and
- wind speed in m/s.

Temperature plays an essential role in air quality forecasting as it can cause chemical reactions in the atmosphere resulting in favourable conditions for smog. Humidity is an important feature, as well. In addition to creating chemical reactions in the air, just like temperature or solar radiation, humidity can have an impact on visibility. During events of the high level of pollution, wind direction can be used to identify the source of pollution as well as to predict and reduce the impact of high pollution episodes.

Yearly Profiles

All selected air quality monitoring stations have different weather patterns (see Figure 13). Earlwood site, the closest station to the CBD, registers the highest level of temperature throughout the year. Camden site is the coldest. The temperature in Liverpool, Camden and Richmond sometimes goes below zero degrees (163 events for Camden site over 6 years 2013-2018). Humidity yearly profile has a downward trend at all locations. The air in Sydney tends to get dryer year over year (3.5pp in Camden, 1.6pp in Chullora, 2.4pp in Earlwood, 3.3pp in Liverpool and 4pp in Richmond) since 2013. Liverpool site registers the highest wind speed out of the five air quality monitoring stations. The weather measurement can be impacted by the location of the monitoring station (next to the park vs. next to the tall buildings). The yearly mean of wind direction was calculated using the mean of circular quantities:

$$\bar{\alpha} = atan2(\frac{1}{n}\sum_{j=1}^{n}\sin\alpha_j, \frac{1}{n}\sum_{j=1}^{n}\cos\alpha_j)$$

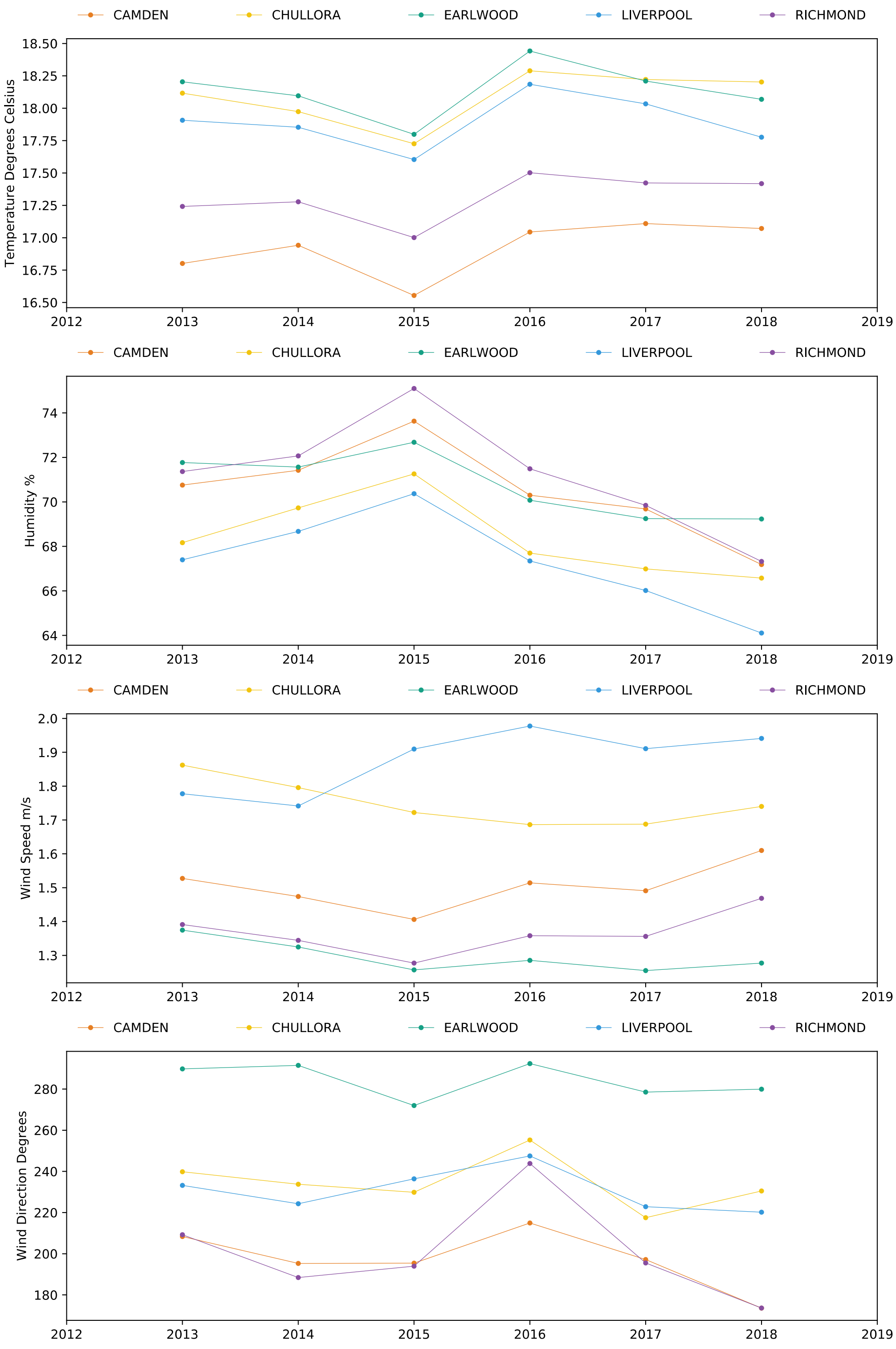


Figure 13: Annual means of meteorological data (2013-2018) by station

Monthly Profiles

All meteorological attributes have seasonality (see data profiling results from Figure 15). Temperature tends to decrease during winter months with the minimum in July. The month with the maximum temperature is January. The difference between the coldest and the hottest months is around 14 degrees Celsius.

Humidity reaches its peak in June, then slows down to reach its lowest level in September. We can assume that such behaviour correlates to the rainy season in Sydney, but we lack rainfall data to make a specific statement. Autumn and winter are characterised by the lowest wind speed, which can result in air pollutants concentration build-up. The wind direction in summer is different from wind direction in other months (Figure 15).

Data Quality

Data quality of the selected meteorological features is overall very good (see Figure 14). Missing values are not common. For instance, at Chullora site, missing values were registered mostly in 2013, and the percentage of missing values versus useful data is less than 5% for wind direction and wind speed. There are also a few missing values in 2015. The volume of missing values isn't critical either for temperature or humidity. Therefore the weather data features have been used in their original format.

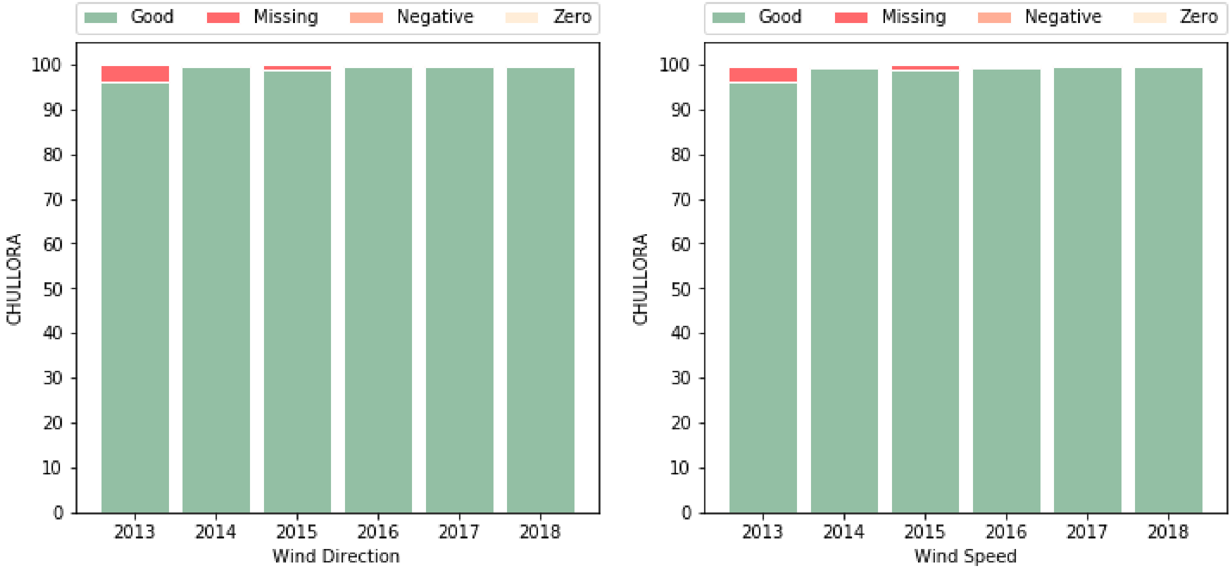


Figure 14: Data Quality for Wind Speed and Wind Direction at Chullora site

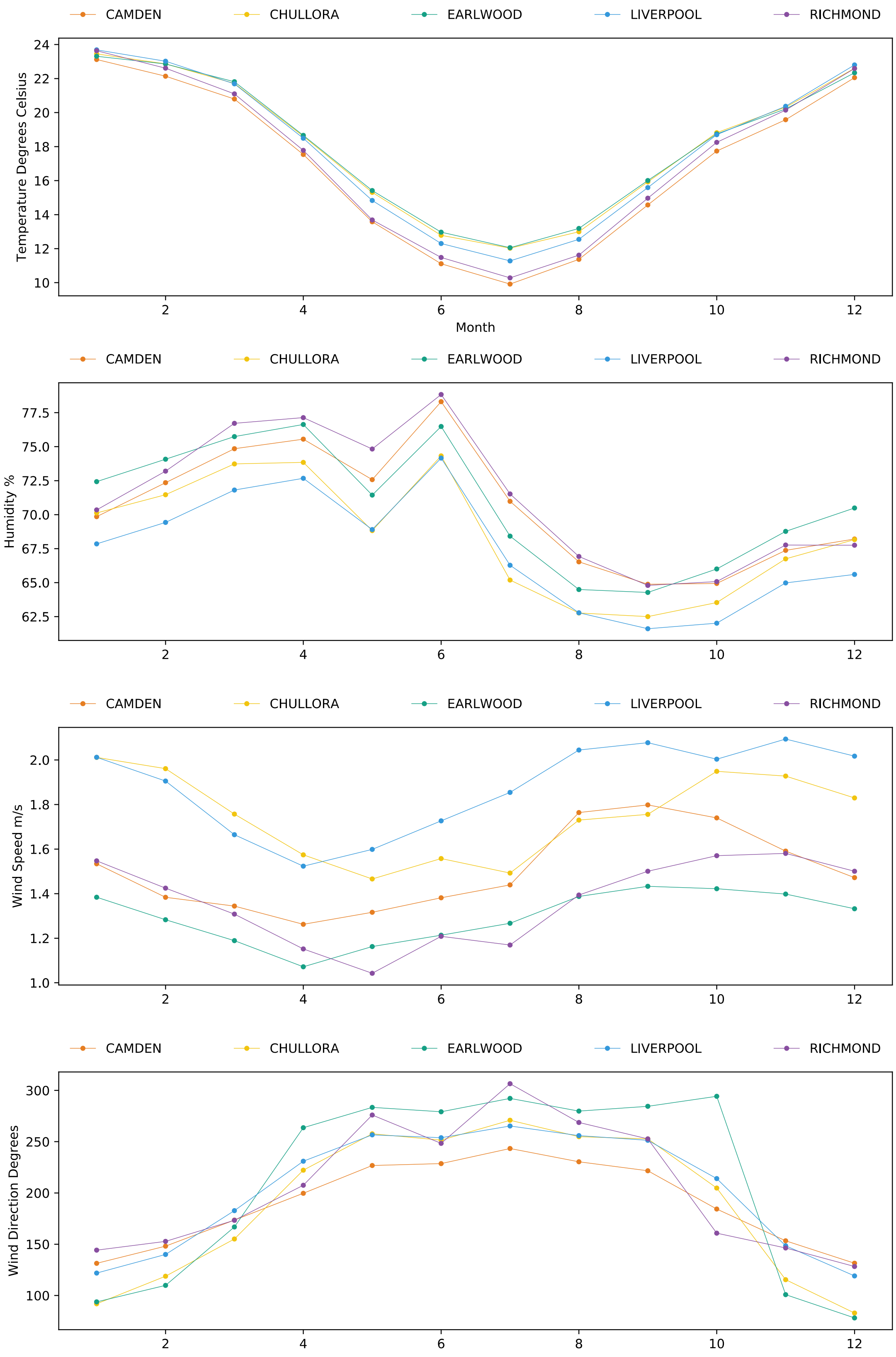


Figure 15: Monthly profiles of meteorology data by station (all years)

Distribution

The distribution of meteorological features is different depending on the weather indicator (see Figure 16). Humidity distribution is slightly left-skewed with averages recorded around 70-80%. Humidity level tends to be on the upper end of the humidity range in Sydney.

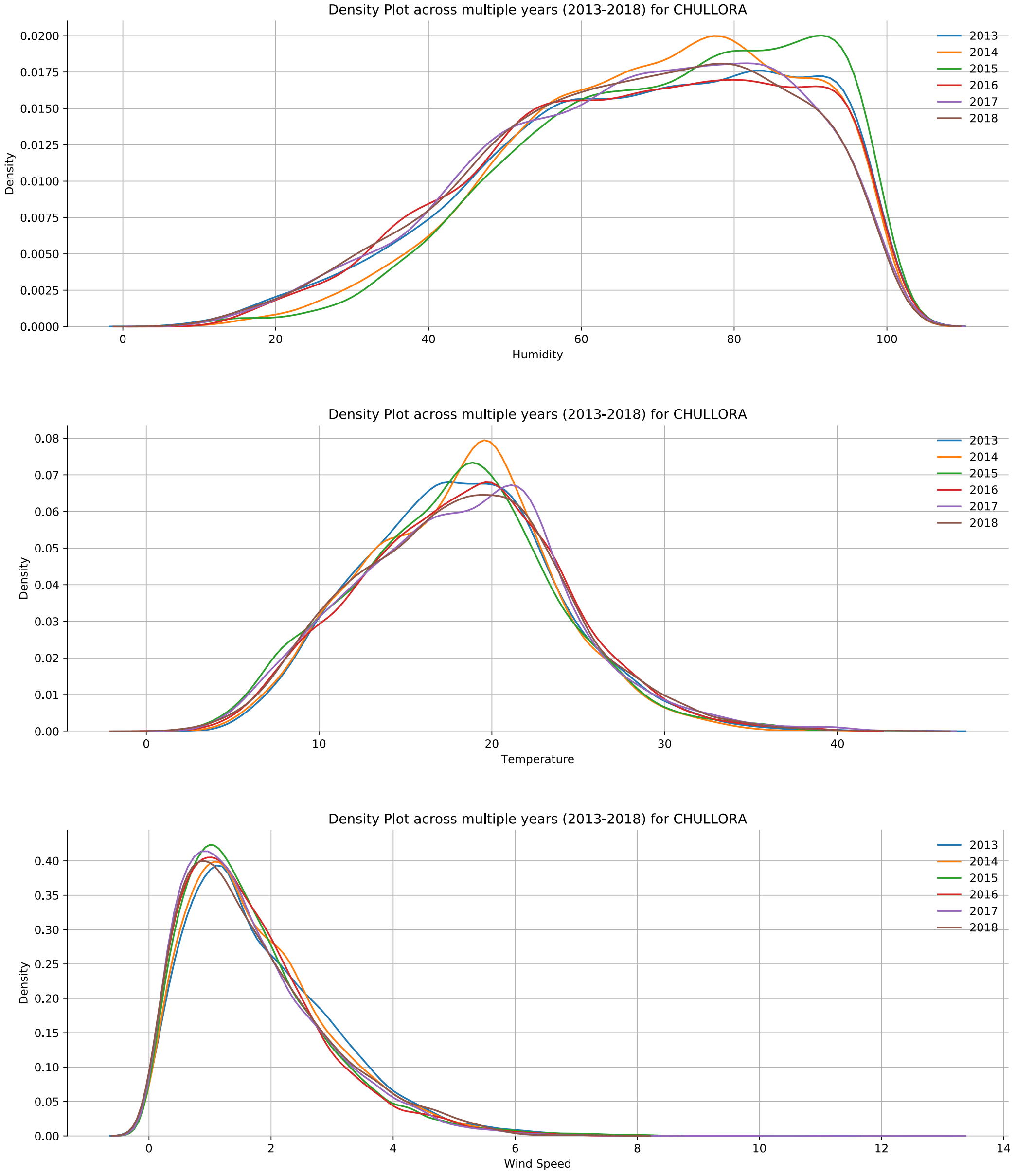


Figure 16: Density plot for humidity, temperature and wind speed at Chullora (all years)

Temperature distribution has almost perfect bell shape with a slight skew towards the right and overall average records registered around 20 °C. The temperature can exceed 40 °C . There were 81 events (hours) of 40 °C threshold exceedance at Richmond site, 55 events at Camden, 43 events at Liverpool, 31 events at Chullora and 25 events at Earlwood over six years (2013-2018).

Wind speed distribution is normal and right-skewed with average recorded around 1m/s. The transformation of the features will be discussed in detail in the pre-processing section of the report.

Traffic Data

The impact of traffic congestion on air pollution was previously explored by many researchers ([7], [8], [9], [17], [28], [33], [34]). They discovered that traffic has a severe impact on the concentration of particulate matter in the air. The concentration of PM2.5 near roads and intersections can be twice higher than the measurements of permanent air quality monitoring stations.

Therefore, for our study, we have started to collect traffic flow counts as reported by the official traffic count stations around the city of Sydney, reporting hourly measurements. We do make the observation that the number of such traffic count locations is limited and does not include the entire traffic loop count detectors which belong to the SCATS monitoring system, proprietary to Transport for NSW.

We have therefore extracted all traffic flow counting from the areas surrounding our 5 chosen AQS and have looked at various ways to integrate this into any air pollution prediction model.

Hourly Profiles by Day of week

Traffic data has an evident pattern by day of the week (see Figure 17). During weekdays, there are two peaks in traffic count at all locations: at 6-7 AM and at 5-6 PM. The pattern is similar to PM10 profile by the day of the week. Traffic count on the weekend has different behaviour. There is just one peak in the late morning / early afternoon. Interestingly, at the Camden site, Saturday is the busiest day in terms of traffic counts, possibly explained by high number of local residents travelling for weekend shopping/activities.

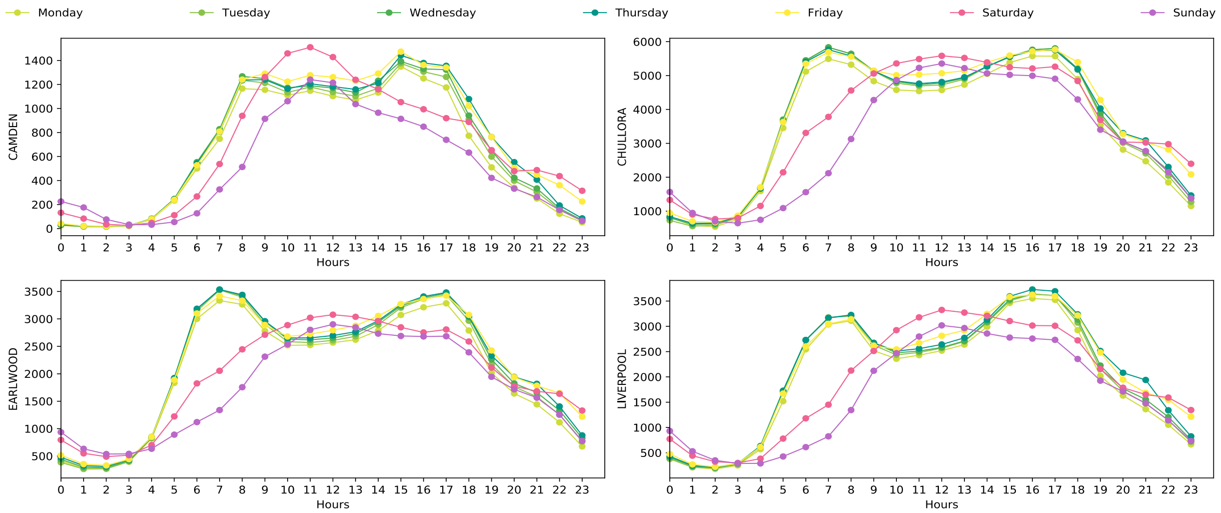


Figure 17: Daily profiles of traffic data by station (hourly mean across all years) by day of the week

Traffic Data Quality

Traffic data was downloaded from open source NSW Transport portal. Firstly, we analysed the spread of traffic counting stations (TCS) around the selected air quality monitoring stations and selected the permanent counters based on the different proximity to AQ stations:

- less than 2km radius,
- less than 2-3km,
- less than 3-4km and
- less than 4-5km.

For Chullora, only TCS in the vicinity of 3-4km was aligned with our data quality needs (counter 28008) (see Figure 18). The TCS in less than 2km radius (43239) had too many missing values in 2013 and 2017. The traffic count also dropped for several months in the end of 2016. Data produced by this TCS is considered unreliable (see Figure 19 for a representation of daily means of traffic count for the traffic counter 43239 (<2km from Chullora AQ station)).

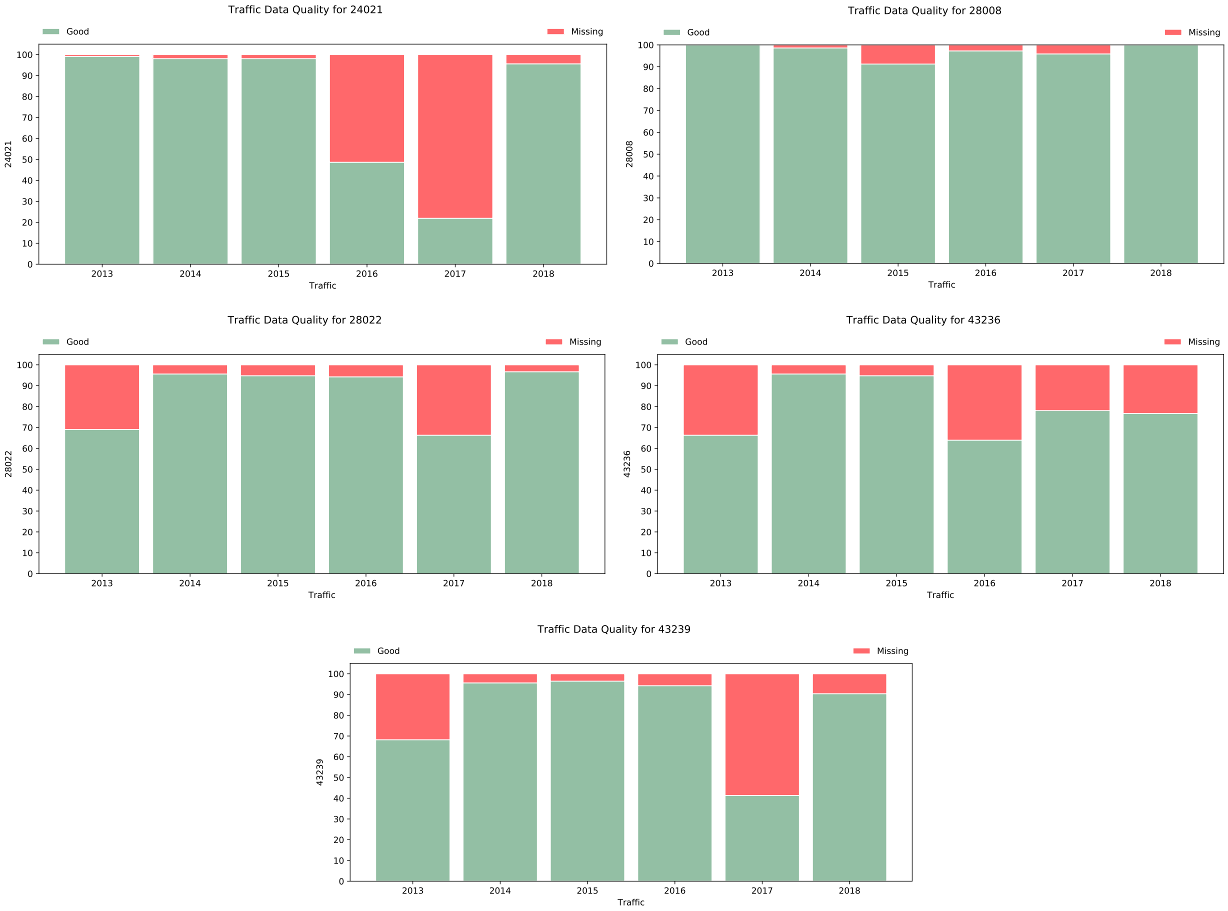


Figure 18: Data Quality for traffic feature for Chullora station by counter

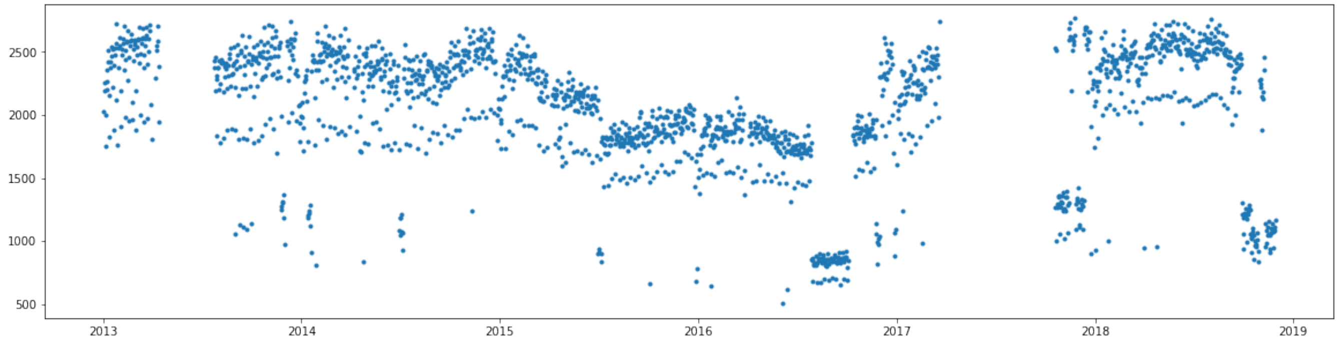


Figure 19: Daily means of traffic count for the traffic counter 43239 (<2km from Chullora AQ station)

Calculating the mean across all five traffic counters for Chullora also produced unreliable results with complicated multimodal distribution and different shape for each year (see Figure 20).

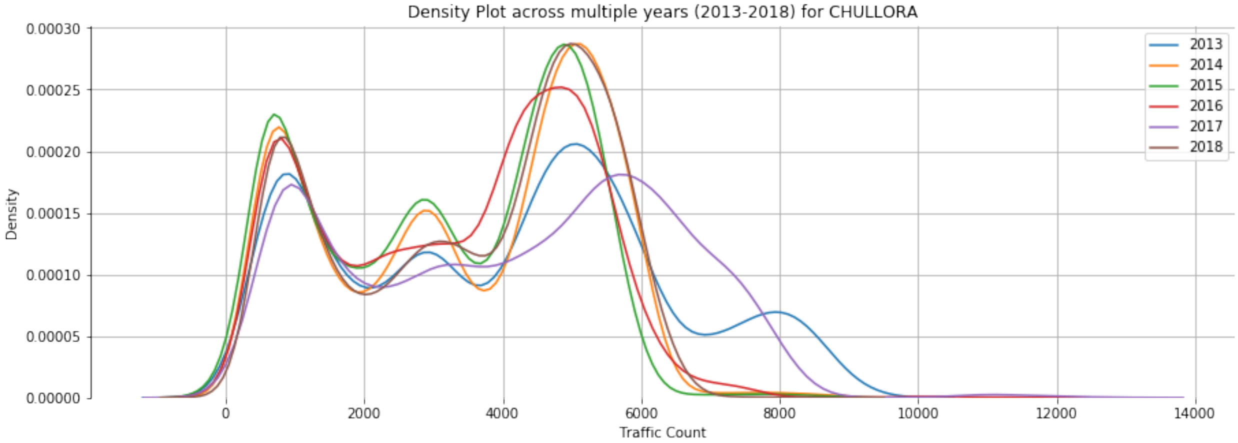


Figure 20: Distribution of the average traffic count of all traffic counters combined for Chullora

Traffic counter 28008 was selected to be included as a feature for prediction model as it's the only traffic counter in proximity of Chullora air quality monitoring station that produced reliable and consistent data.

This type of analysis has been carried out for all air quality stations, and all possible combinations of traffic flow information in their vicinity.

Data Preprocessing

Preprocessing of some features and target variable was necessary before building any prediction model. As discovered in the data exploration stage, the following issues needed to be resolved:

- Missing values imputation for air pollution indicators and features,
- Removal of outliers,
- Transformation of features with skewed distribution.

Missing Values Imputations

During exploratory data analysis, we discovered different patterns in pollution data depending on the day of the week, the hour of the day or month. Firstly, we created daily profiles of pollution data for each air monitoring station for each day of the week and month. Figure 21 shows the visualisation of average daily PM10 profile for Chullora on Tuesdays in month of February. Grey dotted lines are different years. The orange line is the average across all these years, and the green line around it is a 95% confidence interval. Secondly, we created a vector of hourly averages based on the station, day of the week and the month. Lastly, we used these vectors to replace any missing values in our data set (example – missing values for PM10 in February 2018 at 8 AM in the morning will be replaced with average values of PM10 at 08AM calculated across all years of data records). As the number of missing values was not critical, this approach solved the problem of missing values without skewing the data.

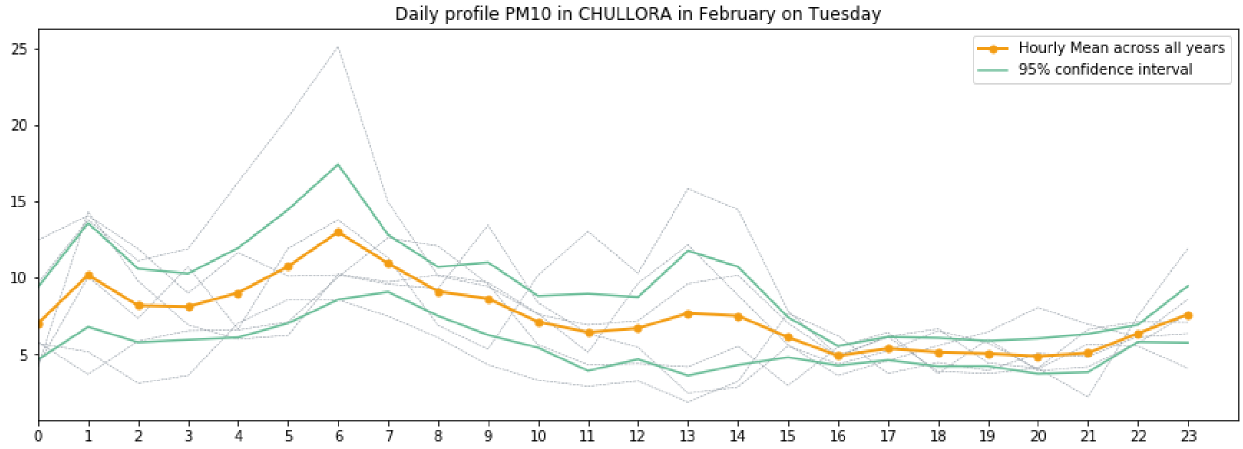


Figure 21: Daily profile of PM10 at Chullora by hour by day of the week by month averaged across all years with 95% confidence interval

Outliers

When building a prediction model, we need to make sure that the model will not be overfitted and will have a good capability to generalise. Therefore, it was decided to remove the outliers. We removed data points above 99th percentile (>63.7 µg/m3) resulting in 526 data points being removed (1% of the dataset). Figure 22-Figure 23-Figure 24 provide an outlier representation by hour, month and year.

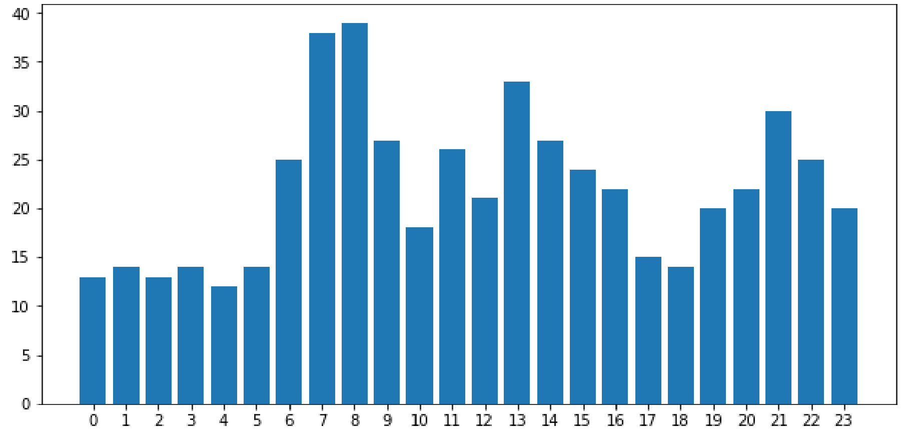


Figure 22: Number of PM10 outliers above 99th percentile at Chullora AQ station by hour of the day from 2013 to 2018

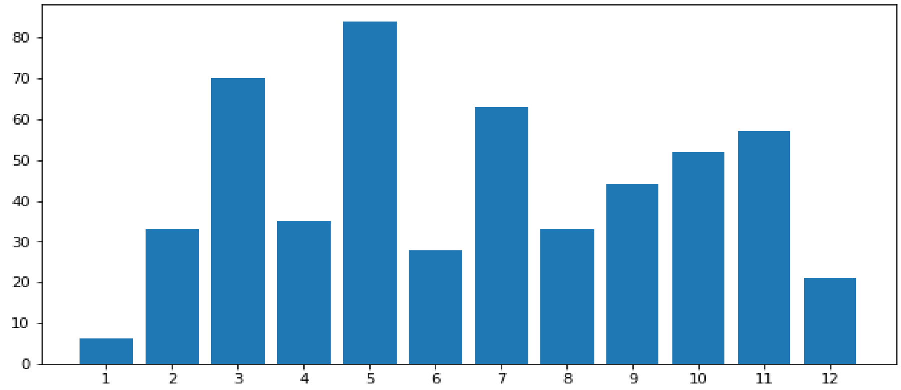


Figure 23: Number of PM10 outliers above 99th percentile at Chullora AQ station by month from 2013 to 2018

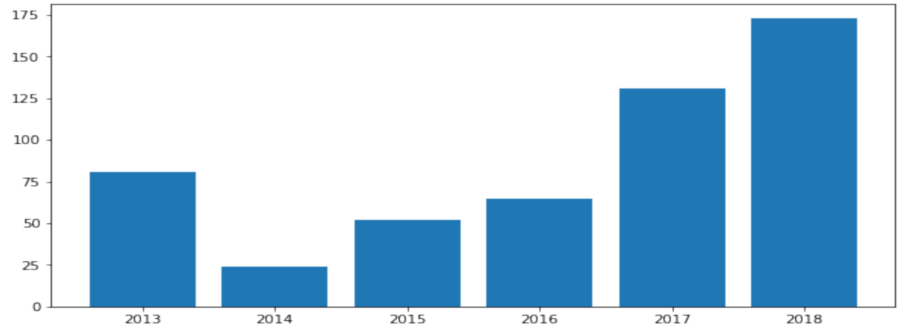


Figure 24: Number of PM10 outliers above 99th percentile at Chullora AQ station by year

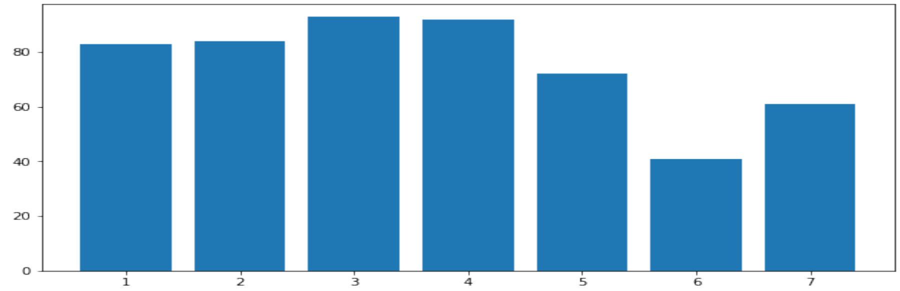


Figure 25: Number of PM10 outliers above 99th percentile at Chullora AQ station by day of the week from 2013 to 2018

Negative values and zeroes were dropped as well. Such data is considered faulty, the prediction of errors produced by the equipment is out of the scope of this research project. There were 244 data points below or equal zero. The biggest number of outliers was recorded at 8AM when we analyse at the aggregation by hour (see Figure 22), in May when we look at the aggregation by month (see Figure 23) and during weekdays when we look at the aggregation by day of the week (see Figure 25). Number of outliers increase year over year, which reflect a deterioration in the data quality (see Figure 24).

Data Transformation

The target variable for our prediction, the meteorology data and the traffic features were scaled with MinMaxScaler to optimise the computational power and to make sure we have similar ranges of scalability in our models training. The new range of these features is between [0, 1].

Wind speed feature was transformed to achieve normal bell-shape distribution by using box cox library and lambda = 0.0, which corresponds to the natural logarithm. The distribution of the transformed wind speed feature is shown on Figure 26.

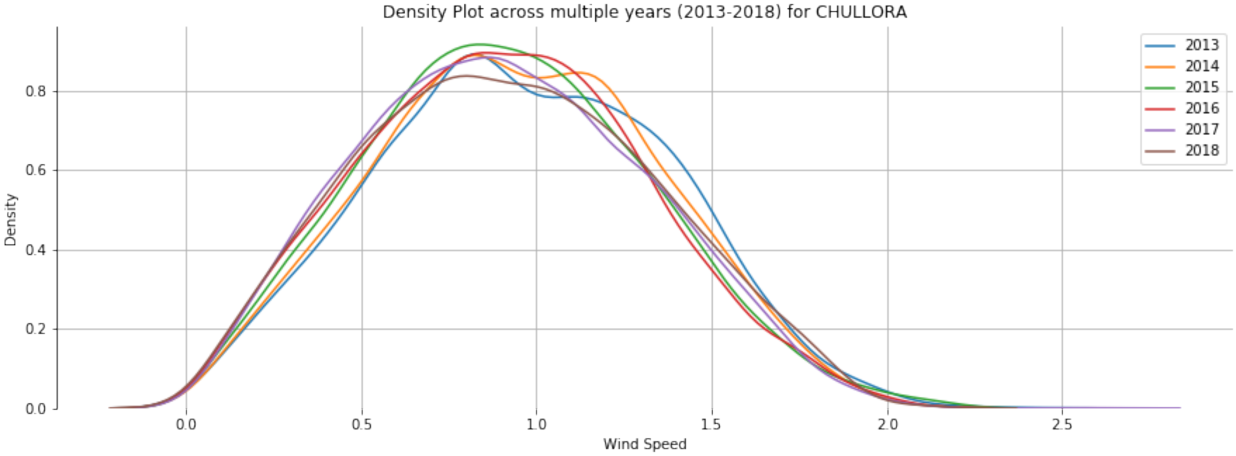


Figure 26: Density plot for log transformed wind speed at Chullora (all years)

Predictive Modelling

Within the research project, we built a model to predict the concentration of PM10 at Chullora air quality monitoring station. Chullora is one of the air quality stations located near CBD. Data quality for this station was better than for other stations. PM10 was chosen as a predictor due to its possible correlation with traffic.

To summarize, the prediction modelling approach we present in this study aims to predict future concentrations of PM10 based on historical data in conjunction with meteorology data, traffic, holidays, day of the week and month.

Baseline Model

To better understand which features are more important, we started with a simple Linear Regression predictor. Firstly, the model had only one feature - year. As shown in Figure 27, the model couldn't predict anything, and R squared was negative. Then, we started adding features one by one to see which attribute would have the most significant impact. Grey dotted line separates ancillary data like meteorology and traffic from historical pollution data. From the graph, we can see that wind speed and traffic flow had a positive impact on the performance of the model, but it was very slim. After we started adding historical data, the performance improved slowly with the steep peak after adding a 3-hour step.

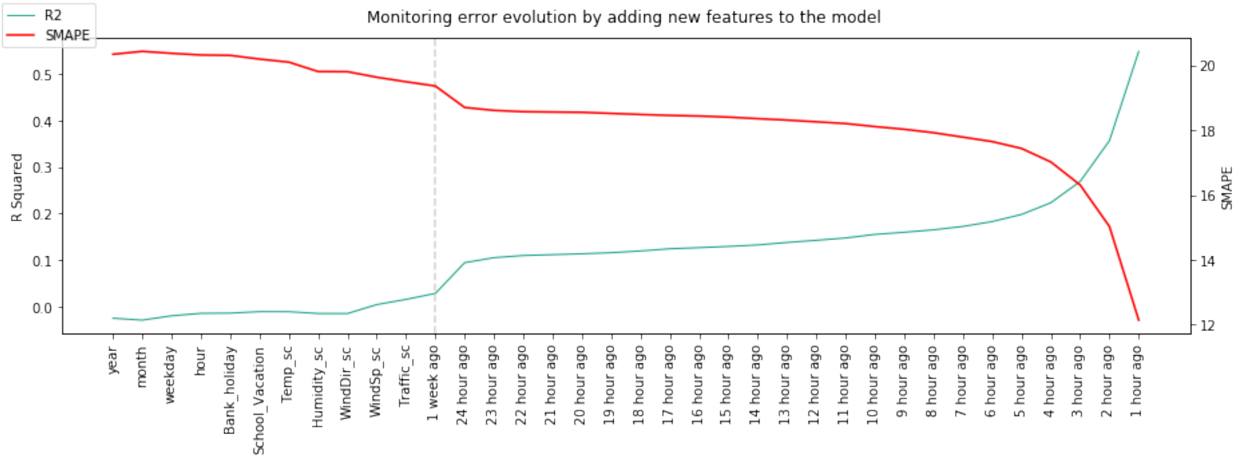


Figure 27: The evolution of R-squared and SMAPE for linear regression through adding more features one-by-one.

The final Linear Regression model was trained with all the features above on five years of historical data (2013-2017). The test set consisted of PM10 concentrations for 2018. We managed to achieve 0.61 coefficient of determination on train set and 0.58 on the test set. MAPE and SMAPE were 37.8% and 12.4% respectively.

Random Forest

Random Forest model was trained and tested within the same timeframe: 5 years of training data and one year of testing. For hyperparameter tuning, we used randomised search with 100 iterations and time-series 4-fold split. The search was optimising for the coefficient of determination. This way, we have identified the best parameters for the model, which are as follows:

Bootstrap: True,

max_depth: 10,

max_features: 'auto',

min_samples_leaf: 2,

min_samples_split: 5,

n_estimators: 800.

Performance of Random Forest Regressor exceeded the results achieved with linear regression with a coefficient of determination of 0.59 on the tests set. MAPE and SMAPE were slightly better than for linear regression: 37.2% and 12.1% respectively.

By using feature importance method for Random Forest (see Figure 28), we can see that regardless of the performance improvement, the random forest might not be the best fit for this problem as the predictor mostly learned on a single feature - 1-hour step.

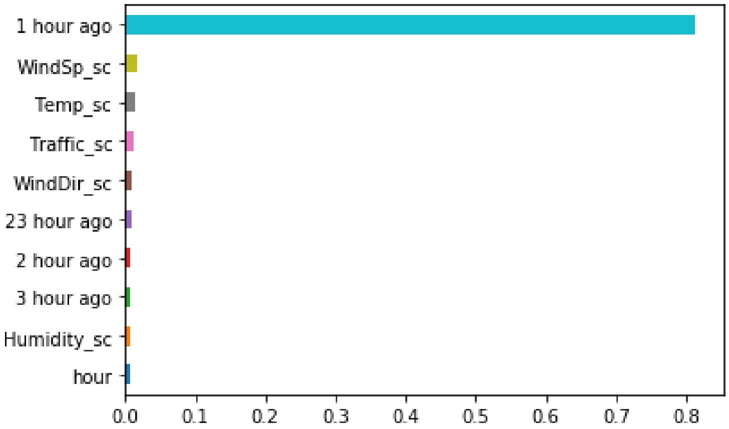


Figure 28: Feature importance of Random Forest Regressor (PM10/Chullora)

XGBoost

Finally, we used XGBoost regressor to solve the problem of air quality prediction in Sydney. For hyperparameter tuning, we used the randomized search with time series 4-fold split and coefficient of determination scoring. We ran 100 iterations to find the best parameters, which were:

{'min_child_weight': 7,

'max_depth': 6,

'learning_rate': 0.1,

'gamma': 0.0,

'colsample_bytree': 0.5}

With these parameters, we managed to achieve R squared of 0.62 on the test set.

Mean Absolute Percentage Error was 36%, and Symmetric Mean Absolute Percentage Error was 11.8%. 1-hour step is the most crucial feature. Compared to random forest regressor (see Figure 29), XGBoost has taken into consideration other historical data like the 3-hour step, 2-hour step and 24-hour step, but also ancillary features like wind speed, traffic and temperature.

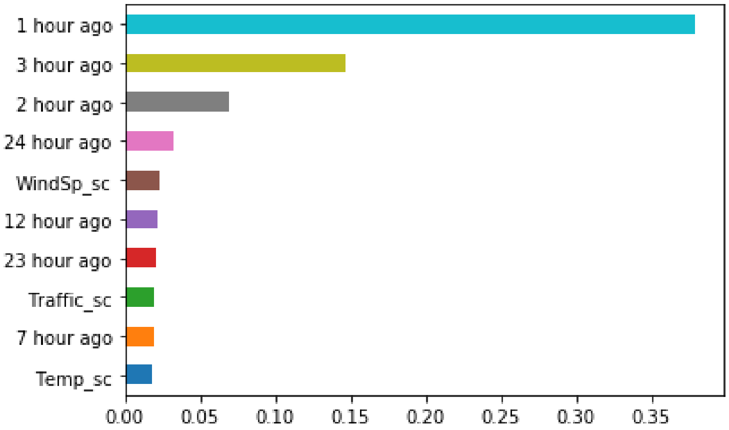


Figure 29: Feature importance of XGBoost Regressor (PM10/Chullora)

Evaluation and model comparison

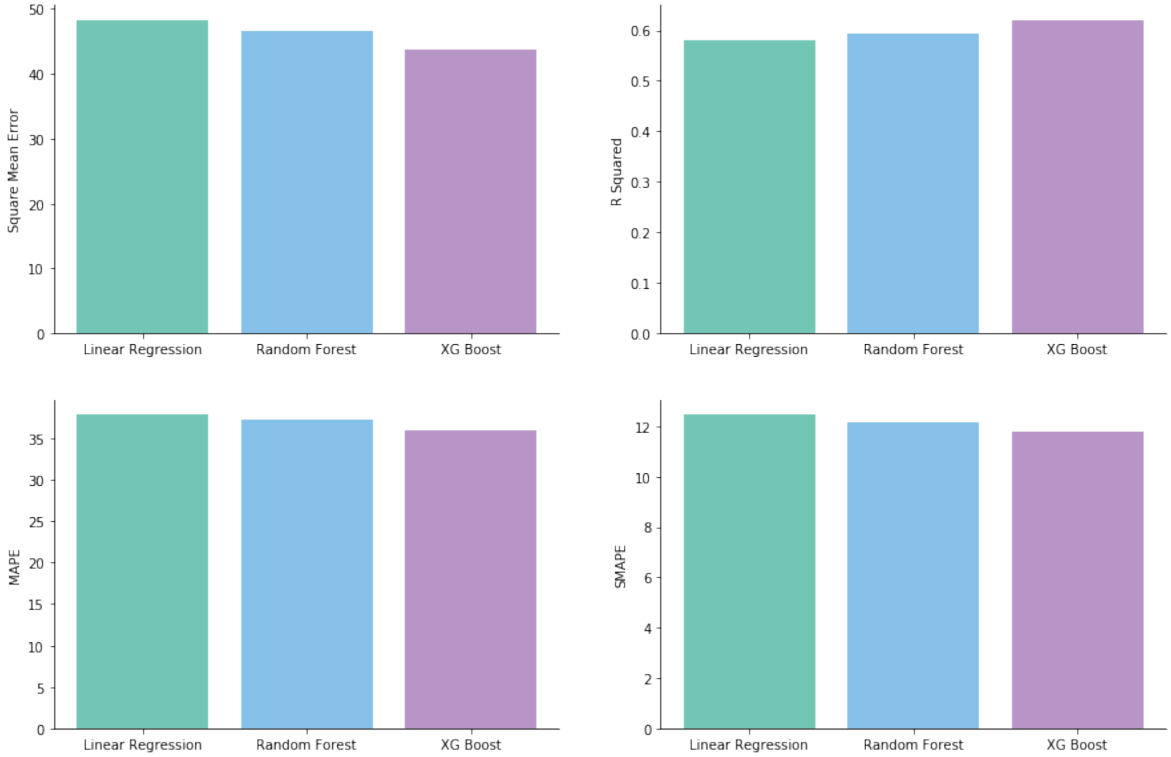


Figure 30: Comparison of regression evaluation metrics (MSE, R Squared, MAPE, SMAPE) by model

All models were evaluated using regression metrics such as MSE, R-squared, MAPE and SMAPE. Figure 30 summarises the performance of each model by each error metric. XGBoost showed the best performance compared to baseline linear regression and random forest regressor for all metrics. By using XGBoost, we decreased the mean squared error from 48.2 for linear regression to 43.6. Coefficient of determination for XGBoost was 0.62, which is 0.04 higher than for the baseline model. Percentage errors also decreased with XGBoost (MAPE 36% and SMAPE 11.8%).

Knowing that 1-hour step is the most important feature, we calculated the same error metrics between real data and 1-hour step feature to see if the model predicts better than just taking the data of 1 hour ago (dummy). Dummy regressor showed worse results than the models (MSE - 59.1, coefficient of determination 0.49, SMAPE 13.4%) except for MAPE - 36.3%. Therefore, we can conclude that this metric might not be representative for this problem.

Figure 31 is a scatter plot of real data versus prediction results provided by different models deployed in this study. XGBoost has a narrower spread around the ground truth line than other regressors. The spread of Linear Regression points is broader and more randomised. All models struggle to predict real values after a certain threshold (>40µg/m3) and perform rather well for shorter PM10 predictions (<40µg/m3). Maximum prediction achieved for each model is 55.4 for Linear Regression, 47.9 for Random Forest and 51.2 for XGBoost, whereas the maximum real concentration of PM10 in this data sample was 63.7.

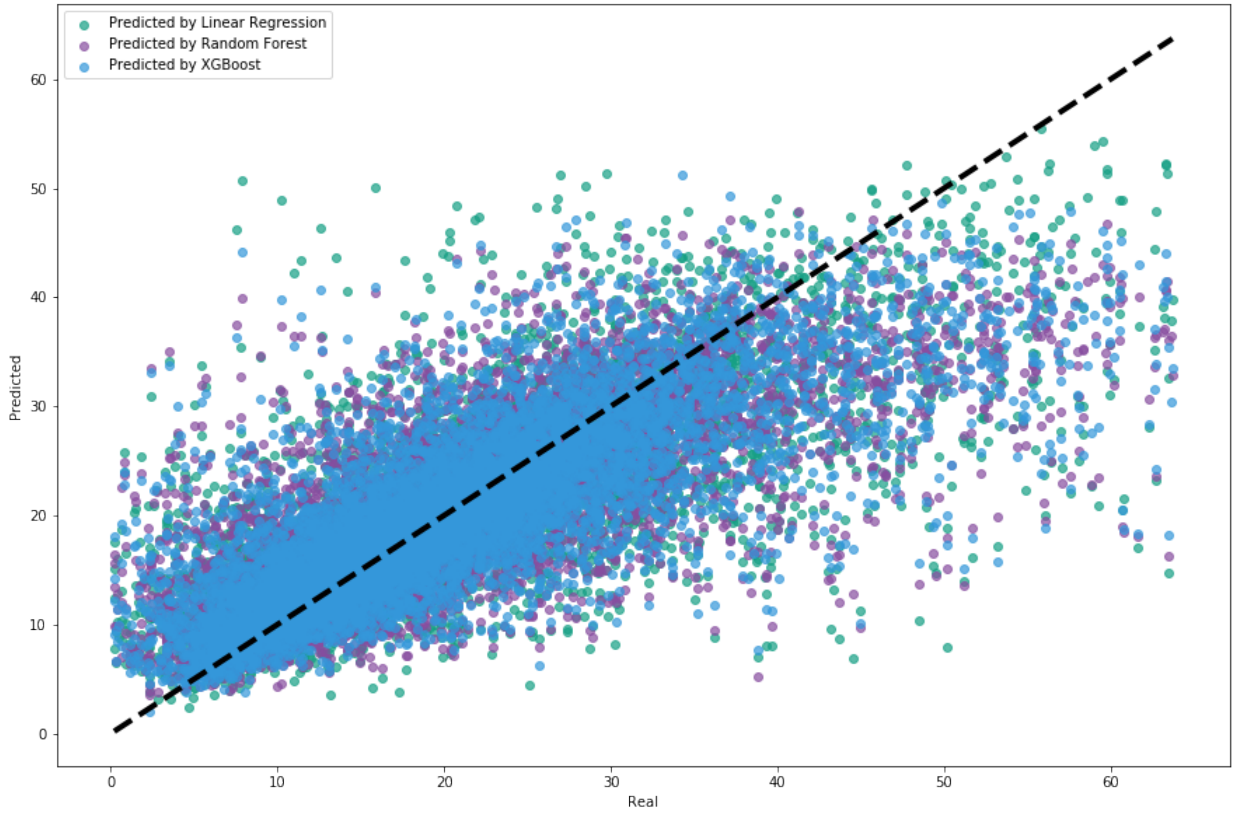


Figure 31: Scatter plot predicted by models versus real

Figures 32-34 show as well the comparison between actual means by month, hour and day of the week with predicted means. Almost every month of 2018 was underpredicted by the models except June, and October. December was overpredicted by XGBoost.

6 AM, 7 AM, and 8 AM are all associated with the highest PM10 pollution. Neither of the models could predict these hours closely enough. Weekend means were correctly predicted by the models, whereas weekdays were under predicted during the morning AM peak hours.

If we filter the data and prediction by the specific hour, weekday or month and then calculate mean squared errors, we get slightly different picture compared to average values in Figure 32-34.

For instance, we confirm that XGBoost performed well in June and October (MSE around 20), but predictions for February and May are subpar (Figure 35). The highest mean squared error was seen for 8 AM, 3 PM and 5 PM. On the other hand, XGBoost performed fine for night hours with MSE around 30. MSE is high for all weekdays except Sundays.

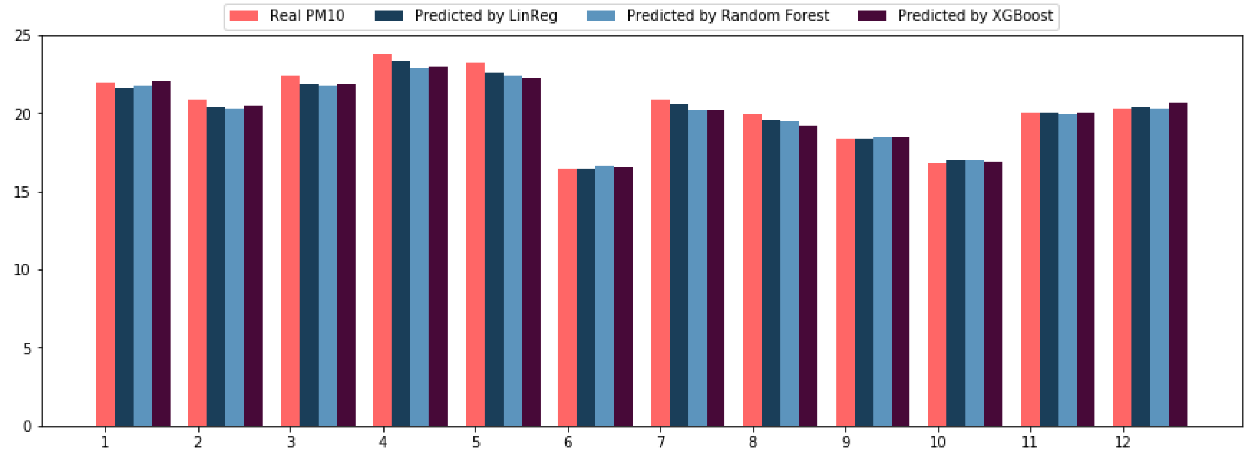


Figure 32: Monthly means - comparison between ground truth and predictions

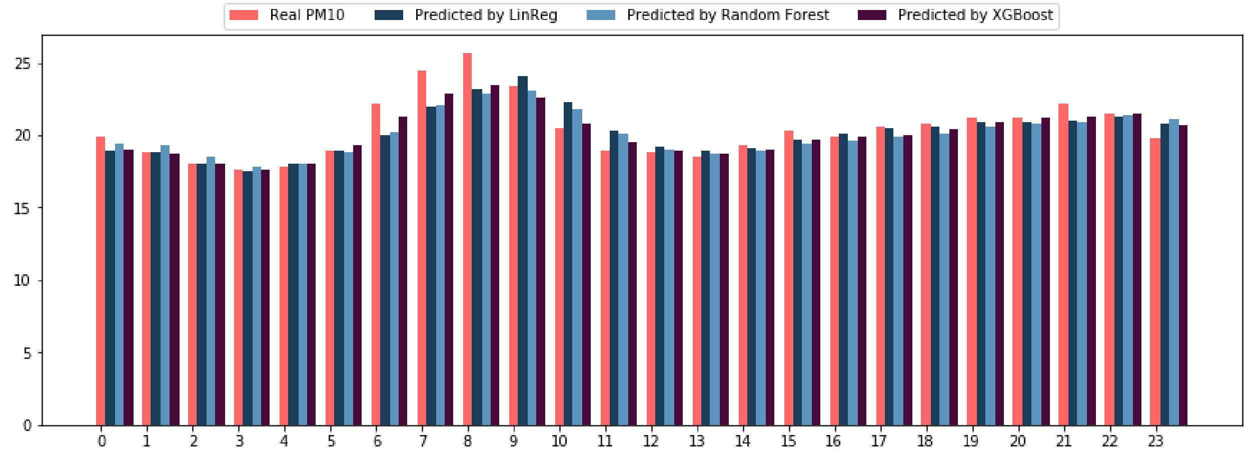


Figure 33: Hourly means - comparison between ground truth and predictions

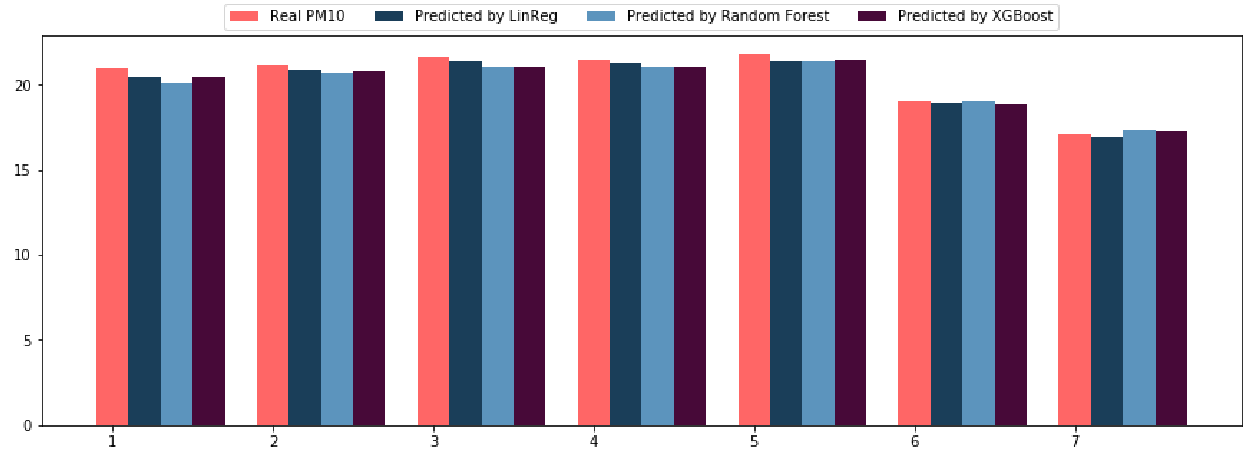


Figure 34: Means by day of the week - comparison between ground truth and predictions

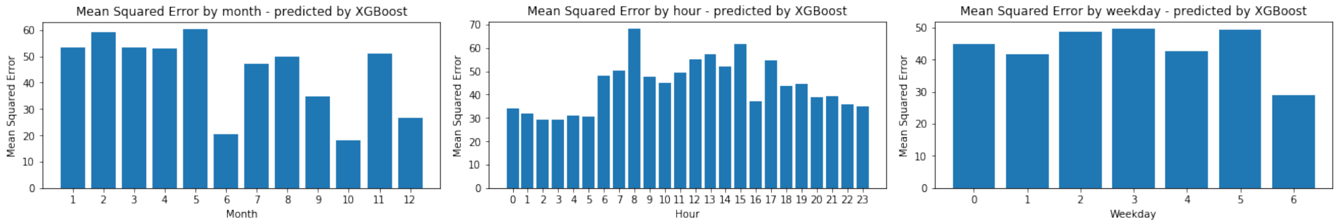


Figure 35: Mean Squared Error by month, hour and day of the week - predicted by XGBoost

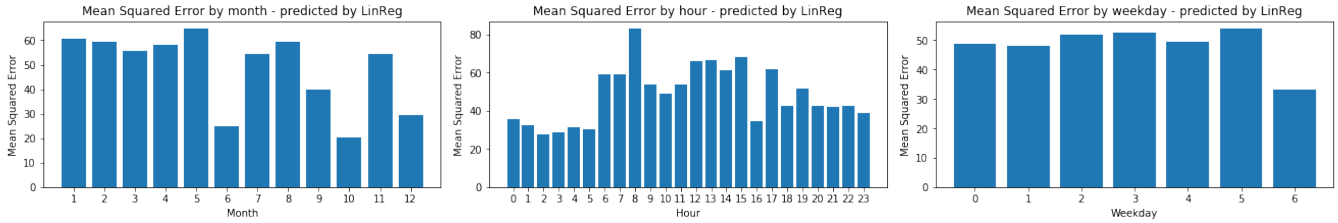
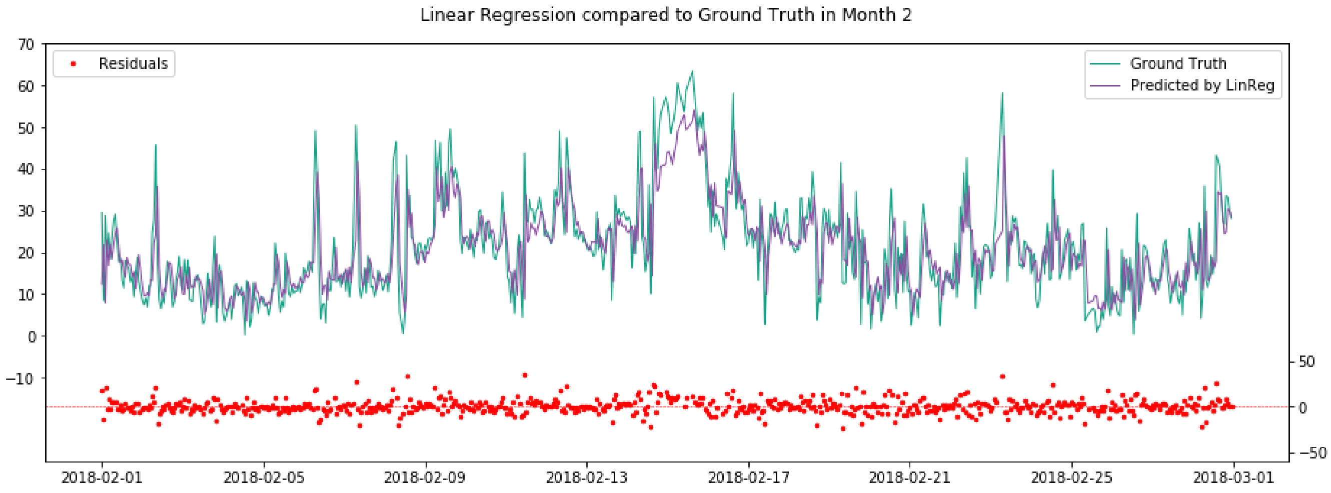


Figure 36: Mean Squared Error by month, hour and day of the week - predicted by Linear Regression

The linear Regression model was worse at predicting 8 AM concentrations (Figure 36). On the other hand, this model performed better than XGBoost in February. Otherwise, the errors have the same pattern as XGBoost: low MSE in June, October; on Sundays and at 4 PM.

If we zoom in into February (Figure 37), we can notice the significant increase in PM10 concentrations between the 13th and 17th of the month. It is an excellent example of when the slightly overfitted model of Linear Regression performed better than a tuned XGBoost model that tends to generalise better. This event causes poor MSE for February, and we can see the flip above zero in residuals in mid-February.



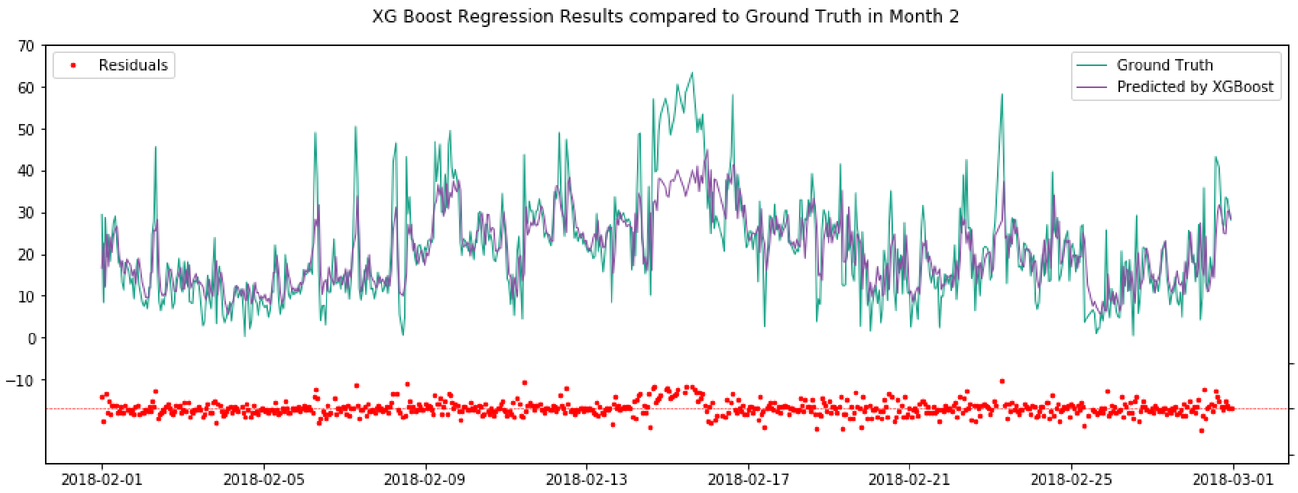


Figure 37: Residuals: XGBoost and linear regression versus ground truth in February 2018

Conclusion

Air pollution is a crucial issue in the face of urbanisations and climate change. There is a statistically significant correlation between the concentration of fine particulate matter and number of premature deaths, hospitalisations and emergency dispatches. Researchers all over the world work on the problem of air pollution prediction to help governments, organisations and regular citizens make better decisions. Even though the issue is persistent, the approach to solving the problem is somewhat outdated. The use of modern prediction techniques like machine learning is underused. This research project had an aim to implement state of the art machine learning model to predict air quality in Sydney to provide a travelling decision-making tool for the citizens.

The first step of the research was an exploratory data analysis of PM2.5, PM10 and NEPH. We also looked at ancillary data like meteorology, traffic and holidays. The level of pollution in Sydney is healthy overall except for the special events that happen due to fire or dust storms. It is worth to mention that the concentration of pollutants in the air grows year on year. PM2.5 increase by 38% since 2008, PM10 by 23% and NEPH by 15%. Also, the number of outlier events is increasing. For PM10, the number of outliers above 99th percentile more than doubled since 2013 (173 hours vs. 81 hours). The yearly average of PM2.5 exceeds annual national standards in Liverpool since 2012. Chullora site exceeds PM2.5 standards since 2013 with 2014-2015 being on the border. The majority of sites are getting closer to the limit each year. In terms of data quality, PM10 and NEPH datasets have a tolerable number of missing values, whereas the number of missing and faulty values (negative or zero) for some stations in years is around 15%. Also, only four stations have a long history of registering the level of PM2.5, two of them are non-conforming anymore.

Meteorological data for this study consisted of temperature, humidity, wind speed and wind direction. All features have seasonality. Humidity has a downward trend - the air in Sydney is getting dryer year over year. Wind speed at night is generally slower than during the day. Data quality of the selected meteorology attributes is good.

Traffic count was taken from the open-source traffic portal for NSW - Traffic Volume Viewer. The data comes from permanent traffic counters. Consequently, it was challenging to find traffic counters that would be in close vicinity of air quality monitoring stations with decent data quality. For Chullora, there was just one traffic counter on Liverpool road, which data quality was reliable and consistent. It is located 2.7km away from the station. Different behaviour was identified for each station by the day of the week. During weekdays there are always two peaks - in the morning and in the evening when citizens travel to/from work or school. On the weekend there is just one peak later in the morning / early in the afternoon. The pattern resembles the behaviour of PM10.

The prediction model was trained on five years of data (2013-2017) and tested on one year of data (2018). We built the model to predict the concentration of PM10 at the Chullora site. The baseline model was linear regression. We also trained Random Forest and XGBoost Regressors. Even though Random Forest and XGBoost outperformed baseline model, Random Forest ignored all the features of the model except the 1-hour step, and it also took hours to tune its hyperparameters. XGBoost achieved the highest coefficient of determination (0.62) and the lowest mean squared error (43.6). The model performed better on specific months - June and October with a coefficient of determination of 0.71 and 0.72 respectively. Night hours also had better accuracy versus the whole model, as well as Sunday predictions, which can be explained by a naturally lower concentration of PM10. There are fewer outliers during these periods. The model learned to generalise well, but with values of PM10 going higher than 40µg/m3, all models started to underpredict.

Due to the low correlation between PM10 level and meteorology or traffic, ancillary features didn't have a big impact on the model, but at the same time, we wouldn't be able to achieve this accuracy without them.

Recommendations and Future Work

In the face of the increasing trend of air pollution in Sydney, it is essential to have a reliable state of the art prediction model. Even though we managed to achieve a decent level of accuracy and minimise prediction errors with machine learning algorithms like XGBoost (coefficient of determination 0.62), the performance of the model can be improved. Adams & Kanaroglou 2016 and Mihăiţă et al. achieved higher accuracy (up to 0.81) in predicting air pollution with neural networks.

Long short-term memory is an artificial recurrent neural network (RNN) architecture. It proved to be a good fit for time series forecasting.

In addition to building a deep learning algorithm, designing different models for different seasons of the year might improve the performance as we have seen that linear regression performed better than XGBoost in February 2019. Due to different pattern in ancillary and target data depending on the day of the week, it is possible that building two predictors: one for weekdays and another for weekends with different hyperparameters and learning rates to improve the overall performance.

For now, the model is capable of prediction one pollution indicator PM10 at one air quality monitoring station - Chullora. To produce a more wholesome solution, it is possible to build a model that will predict all air quality indicators at all sites simultaneously like STDL-based model developed by (Li et al. 2016).

ACKNOWLEDGEMENT:

The authors of this study would like to thank NSSN (NSW Smart Sensing Network), in particular Zhitao Xiong for the help, advice and data provided.

References

1. Nations, U. 2019, World Urban Prospects. The 2018 Revision, New York.
2. Shaddick, G., Thomas, M.L., Green, A., Brauer, M., van Donkelaar, A., Burnett, R., Chang, H.H., Cohen, A., Van Dingenen, R. & Dora, C. 2018, 'Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution', Journal of the Royal Statistical Society: Series C (Applied Statistics), vol. 67, no. 1, pp. 231-53.
3. Morgan, G., Corbett, S., Wlodarczyk, J. & Lewis, P. 1998, 'Air pollution and daily mortality in Sydney, Australia, 1989 through 1993', American journal of public health, vol. 88, no. 5, pp. 759-64.
4. Johnston, F., Hanigan, I., Henderson, S., Morgan, G. & Bowman, D. 2011, 'Extreme air pollution events from bushfires and dust storms and their association with mortality in Sydney, Australia 1994–2007', Environmental research, vol. 111, no. 6, pp. 811-6.
5. Broome, R.A., Fann, N., Cristina, T.J.N., Fulcher, C., Duc, H. & Morgan, G.G. 2015, 'The health benefits of reducing air pollution in Sydney, Australia', Environmental research, vol. 143, pp. 19-25.
6. Salimi, F., Henderson, S.B., Morgan, G.G., Jalaludin, B. & Johnston, F.H. 2017, 'Ambient particulate matter, landscape fire smoke, and emergency ambulance dispatches in Sydney, Australia', Environment international, vol. 99, pp. 208-12.
7. Irga, P., Burchett, M. & Torpy, F. 2015, 'Does urban forestry have a quantitative effect on ambient air quality in an urban environment?', Atmospheric Environment, vol. 120, pp. 173-81.
8. Wadlow, I., Paton-Walsh, C., Forehead, H., Perez, P., Amirghasemi, M., Guérette, É.-A., Gendek, O. & Kumar, P. 2019, 'Understanding Spatial Variability of Air Quality in Sydney: Part 2—A Roadside Case Study', Atmosphere, vol. 10, no. 4, p. 217.
9. Rivas, I., Kumar, P. & Hagen-Zanker, A. 2017, 'Exposure to air pollutants during commuting in London: Are there inequalities among different socio-economic groups?', Environ Int, vol. 101, pp. 143-57.

10. Morgan, G., Corbett, S. & Wlodarczyk, J. 1998, 'Air pollution and hospital admissions in Sydney, Australia, 1990 to 1994', American journal of public health, vol. 88, no. 12, pp. 1761-6.
11. Li, X., Peng, L., Hu, Y., Shao, J. & Chi, T. 2016, 'Deep learning architecture for air quality predictions', Environmental Science and Pollution Research, vol. 23, no. 22, pp. 22408-17.
12. Adams, M.D. & Kanaroglou, P.S. 2016, 'Mapping real-time air pollution health risk for environmental management: Combining mobile and s air pollution monitoring with neural network models', J Environ Manage, vol. 168, pp. 133-41.
13. Mihăiţă, A.S., Dupont, L., Chery, O., Camargo, M. & Cai, C. 2019, 'Evaluating air quality by combining stationary, smart mobile pollution monitoring and data-driven modelling', Journal of Cleaner Production, vol. 221, pp. 398-418.
14. Friedman, J.H. 2001, 'Greedy function approximation: a gradient boosting machine', Annals of statistics, pp. 1189-232.
15. Liaw, A. & Wiener, M. 2002, 'Classification and regression by randomForest', R news, vol. 2, no. 3, pp. 18-22.
16. Yuming Ou, Adriana-Simona Mihăiţă, Fang Chen,14 - Big data processing and analysis on the impact of COVID-19 on public transport delay, Editor(s): Utku Kose, Deepak Gupta, Victor Hugo C. de Albuquerque, Ashish Khanna, Data Science for COVID-19, Academic Press, 2021, Pages 257-278, ISBN 9780323907699, https://doi.org/10.1016/B978-0-323-90769-9.00010-4.
17. Mihaita A. S., Benavides, M., Camargo, C. Cai, Predicting air quality by integrating a mesoscopic traffic simulation model and air pollutant estimation models, International Journal of Intelligent Transportation System Research (IJITSR), DOI: 10.1007/s13177-018-0160-z, Online ISSN1868-8659, Online in 15 May 2019, Volume 17, Issue 2, pp 125–141.
18. Mihaita A. S., Dupont L., Camargo M., Multi-objective traffic signal optimization using 3D mesoscopic simulation and evolutionary algorithms, Simulation Modelling Practice and Theory (SIMPAT), https://doi.org/10.1016/j.simpat.2018.05.005, Volume 86, August 2018, Pages 120-138.
19. Wen T, Mihăiţă A-S, Nguyen H, Cai C, Chen F. Integrated Incident Decision-Support using Traffic Simulation and Data-Driven Models. Transportation Research Record. 2018;2672(42):247-256. doi:10.1177/0361198118782270.
20. Mihaita A.S., Mocanu S., Lhoste, P., "Probabilistic analysis of a class of continuous-time stochastic switching systems with event-driven control", European Journal of Automation (JESA), July 2016, H5 =17.
21. Monticolo, D., Mihaita, A.S., Darwich, H., Hilaire, V., "An Agent Based System to build project memories during engineering projects", Knowledge Based Systems Journal (KBS), January 2014.
22. Monticolo, D. Mihaita A.S. "A multi Agent System to Manage Ideas during Collaborative Creativity Workshops", International Journal of Future Computer and Communication (IJFCC), vol 3., nr 1, February 2014, P66-71, (extended version of the paper presented in ICFCC 2013).
23. Mihaita A.S., Mocanu S., "Simulation en temps continu pour la commande orientée événements des systèmes stochastiques à commutation", European Journal of Automation (JESA), 45 1-3 (157-172), Oct 2011. (selected for publication after the MSR Conference Lille 2011).
24. A-S. Mihaita, H. Li, Z. He and M. -A. Rizoiu, "Motorway Traffic Flow Prediction using Advanced Deep Learning," 2019 IEEE Intelligent Transportation Systems Conference (ITSC), 2019, pp. 1683-1690, doi: 10.1109/ITSC.2019.8916852.
25. Mihaita, A.S., Liu, Z., Cai, C., Rizoiu, M.A "Arterial incident duration prediction using a bi-level framework of extreme gradient-tree boosting", ITS World Congress 2019, Singapore, 21-25 Oct 2019.
26. Mao, T., Mihaita, A.S., Cai, C., Traffic Signal Control Optimisation under Severe Incident Conditions using Genetic Algorithm, ITS World Congress 2019, Singapore, 21-25 Oct 2019.
27. Shaffiei, S. Mihaita, A.S., Cai, C., Demand Estimation and Prediction for Short-term Traffic Forecasting in Existence of Non-recurrent Incidents, ITS World Congress 2019, Singapore, 21-25 Oct 2019.
28. Mihaita A. S., Dupont L., Cherry O., Camargo M., Cai C., Air quality monitoring using stationary versus mobile sensing units: a case study from Lorraine, France, 25th ITS World Congress (ITSWC 2018), Copenhagen, Denmark, 17-21st of September 2018, (H5=11).
29. Wen Tao, Mihaita A.S., Nguyen Hoang, Cai Chen, Integrated Incident decision support using traffic simulation and data-driven models. Transportation Research Board 97th Annual Meeting (TRB 2018), Washington D.C., January 7-11, 2018, H5=48.
30. Mihaita A. S., Tyler Paul, Wall John, Vecovsky Vanessa, Cai Chen, Positioning and collision alert investigation for DSRC-equipped light vehicles through a case study in CITI, 24th World Congress on Intelligent Transportation Systems (ITSWC 2017), Montreal, Canada, 29 October - 2 November 2017, H5=11.
31. Mihaita A. S, Cai Chen, Chen Fang, Event-triggered control for improving the positioning accuracy of connected vehicles equipped with DSRC, International Federation of Automatic Control World Congress (IFAC WC 2017), 9-14 July 2017, Toulouse, France, H5 = 32.

32. Mihaita A. S, Tyler Paul, Menon Aditya, Wen Tao, Ou Yuming, Cai Chen, Chen Fang, "An investigation of positioning accuracy transmitted by connected heavy vehicles using DSRC", Transportation Research Board 96th Annual Meeting (TRB 2017), Washington D.C., January 8-12, Paper number 17-03863, 2017, https://pubsindex.trb.org/view/2017/C/1438533 H5=48.
33. Mihaita A. S., Benavides, M., Camargo, M., "Integrating a mesoscopic traffic simulation model and a simplified NO2 estimation model", 23rd World Congress on Intelligent Transportation Systems (ITSWC 2016), Melbourne, Australia, 10-14 October 2016.
34. Mihaita A.S., Camargo, M., Lhoste, P. , " Evaluating the impact of the traffic reconfiguration of a complex urban intersection ", 10th International Conference on Modelling, Optimization and Simulation (MOSIM 2014), Nancy, France, 5-7 November 2014 (accepted on 18th of July 2014).
35. Mihaita A.S., Camargo, M., Lhoste, P. "Optimization of a complex urban intersection using discrete-event simulation and evolutionary algorithms", International Federation of Automatic Control World Congress (IFAC WC 2014), Cape Town, Africa, 24-29 August 2014.
36. Issa, F., Monticolo, D., Gabriel, A. , Mihaita, A.S., "An Intelligent System based on Natural Language Processing to support the brain purge in the creativity process", IAENG International Conference on Artificial Intelligence and Applications (ICAIA 2014), Hong Kong, 12-14 March, 2014.
37. Monticolo, D., Mihaita A.S., "A Multi Agent System to manage ideas during Collaborative Creativity Workshops", 5th International Conference on Future Computer and Communication (ICFCC 2013), Phuket, Thailand, 26 May 2013.
38. Mihaita A. S., Mocanu S., "Un nouveau modéle de l'énergie de commande des systèmes stochastiques à commutation", Septième Conférence Internationale Francophone d'Automatique (CIFA 2012) Grenoble, France, 4-7th of July, 2012.
39. Mihaita A. S., Mocanu S., "An Energy Model for the Event-Based Control of a Switched Integrator", International Federation of Automatic Control World Congress (IFAC WC 2011), Milano, September 2011.