Coarse Particulate Matter Distribution and Variability over South Island, New Zealand

Surafel Tilahun
Auckland University of Technology
Auckland, New Zealand

NOVEMBER 4, 2019
AUT

Contents

1 Introduction

Due to its detrimental health effects, air pollution has become a very topical issue around the globe. WHO reports (WHO, 2016), that around 3 million people perish each year, due to the various health complications caused by toxic micro-particles spread through air pollution. Relative to other countries, New Zealand generally has low air pollution levels and is considered to have cleanest and safest air on the planet. However, some regions have shown degraded air quality with 35 out of 51 PM monitoring sites exceeding the daily national air quality standard (StatsNZ, 2018).

Particulate matter (PM) is a mixture of microscopic particles in liquid or solid form, suspended in the atmosphere. PM particles of many shapes and sizes are grouped into two sub-types, namely PM10 and PM2.5. The former (PM10) consists of coarse particles which have a diameter smaller than or equal to 10μm, while PM2.5 particles are fine with diameter less than or equal to 2.5 μm (EPA, 2018). Owing to its microscopic nature, particulate matter can penetrate deep into the bloodstream and vital organs such as lungs which can cause DNA mutations, respiratory diseases, heart diseases and more. In addition to direct inhalation, Particulate Matter can also affect human health via precipitation and climate.

In this paper, we analyse the spatial distribution and temporal variability of PM10, spatial and spatio-temporal interpolation, and relationship between PM10 and other meteological factors and pollutants. The data were collected at 4 different ground monitoring sites across Canterbury between 10/12/2014 and 01/01/2019. The goal of this paper is:

- To analyze the seasonal variation of PM10 concetraiton across the study area.
- To impliment different types of spatial and spatio-temporal interpolation methods and benchmark their performance.
- To understand the most important factors to predict the concentration of PM10 using Multilinear Regression Model (MLR), wlong with Principal Compoent Analysis (PCA).
- To perfrom seasonal and monthly clustering analysis and insight their transition period.

In general, this paper aims to anser the following main research questions:

- Which reagion have highest PM10 concentration?
- Which interpoation methods provide best interpolated serface area?
- Does PCA improve the prediciton accuracy of MLR?

This paper is organized as follows: section two provides a review of papers related to the subject matter, followed by section 3 where each of the methods we adopt in this paper are described in detail. Section 4 contains key results obtained from the analysis, along with significant findings. Section 5 comprises discussion of each key results and findings and finally, we have conclusion and future scope work the can be done to enhance the work even better in section 6.

2 Literature Review

Particulate matter (PM) is known to vary according to time and space. In New Zealand, PM concentrations are worse during wintertime as coal and wood fires are used for indoor home heating (StatsNZ, 2018). Due to their adverse health effects, particulate matter and aerosols have attracted much attention in NZ academic research circles. Research conducted by (Biancofiore, Busilacchio et al. 2017) examined the ambient air pollution in Wellington during the three weeks of springtime. Indoor and outdoor measures for PM10, PM2.5 and other meteorological variables concentrations indoors outdoors used for elemental speciation analysis. Receptor Receptor modelling and appointment of PM mass by PMF was applied through the EPAPMF program. The results reveal

results indicated that when students are present at school during weekdays (between 9 AM and 3 PM), the indoor average for PM2.5 was significantly higher than when students were not present at school. On the other hand, when students were present at school, the average indoor concentration of PM10 was significantly higher than outdoor concentration. Additionally, the seasonal relationship with climate drivers and weather evolution was investigated in some urban and rural areas across New Zealand. The findings suggest that a correlation exists between low wind speed and higher PM10 concentrations. Another conducted Another paper examines and analyzes the such synoptic factors which vary seasonally, along with their effects on PM10 concentrations (Fiddes, Pezza et al. 2016). To analyze these factors PM10 concentrations were measured by TEOM (Tapered Element Oscillating Microbalance) series and FH62 BAM. Most monitoring sites recorded a seasonal peak for PM10 concentrations in winters. They also support the finding that there is a correlation between increasing PM10 events and anticyclonic circulation that influences NZ's air.

High concentrations of airborne particulate matter are a worldwide problem; thus it has been discussed extensively in literature. For example, an investigation was conducted to examine the influence of several meteorological factors on the accumulation of PM10 and PM2.5 over Shenyang, a city in northeast China (Li, Ma et al. 2017). The paper analyzed the concentrations of PM between January 2014 to May 2016 at 11 different stations. The results showed that PM concentrations peaked during October and November, which is believed to be caused by the burning of crop residues. Spatial distribution analysis suggested that the highest PM2.5 concentration was recorded in central urban areas, while the lowest concentrations were observed in forested areas with little human activity. Relationships between meteorological conditions and PM concentrations were assessed using correlation coefficients. Both PM2.5 and PM10 concentrations were negatively correlated with atmospheric visibility and wind speed. However, air pressure, air temperature and relative humidity showed positive correlations with PM concentrations. (Biancofiore, Busilacchio et al. 2017) also performed predictive analysis on concentration levels of PM2.5 and PM10 on the Adriatic coast of Italy. The authors employed three different methodologies in their study to obtain three days ahead forecasts for PM concentration levels, which includes multiple linear regression, recursive neural network and a non-recursive neural network. The input variables included meteorological factors and carbon monoxide (CO) concentration which improved the forecast accuracy of the models. The model was able to predict 98% the days correctly when PM10 concentration is below 50 micrograms per cubic meter. This work also predicted concentrations levels for PM2.5 using meteorological conditions, CO and PM10 levels as input variables. This is quite useful as PM10 is more commonly measured and can be used to predict PM2.5 when it is not measured at all.

Sometimes data collected from different monitoring stations is not representative of PM concentrations in an entire area. In such cases, spatial interpolation can be used to estimate values at other unknown points. This method was applied to PM concentrations collected from 35 monitoring stations in Beijing (Jie, Kepeng et al. 2016). The data contained PM2.5 and PM10 concentrations averaged over one-hour from March 2013 to February 2014. The authors used MATLAB interpolation and compared its result against Kriging interpolation using k-fold cross validation accuracy metric. Although the original data was only collected from 35 different sites, the interpolation method allowed PM concentrations to be estimated at any latitude-longitude point in Beijing. Results from cross-validation showed that both kriging and radial basis function provided similar interpolation effects. The study also acknowledges that further research and satellite retrieval is required to verify the accuracy of the interpolation function. Finally, the paper also stated that there was mild air pollution throughout spring, summer and autumn, while PM levels were moderate in winter. Overall, there was higher pollution concentration in southern areas compared to northern. In another systematic study, two widely adopted interpolation methods (Inverse Distance Weight (IDW) and Kriging.) are employed to interpolate PM10 concentration of unknown areas in Yard city. The data

was collected from 13 monitoring sites. According to the RMSE and percentage RMSE metric, kriging better interpolated surface area and the highest (297μg) as well as the lowest (35μg) concentration of PM10 was found in spring (Mohammad, Sara, Mohammad, & Miri, 2017).

Interpolation methods have a wide range of applications and several researchers employed different types of interpolation methods outside the air ambient related domains. For instance, Mathew, Jane and Guangquan used Bayesian Spatio-temporal analysis to comprehend the Regional Municipality of Waterloo, which is located in Ontario seasonal property crime trends (Matthew, Jane, & Guangquan, 2017). The researchers were able to discover that most of the crimes in the region are related to a location in a central business district, commercial land use, eating and drinking establishments, schools, parks, and public transit stations. In another paper, researchers investigated the risk of malaria in Kenya using spatial-temporal map by considering 36 years of data on *Plasmodium falciparum*. They used Spatio-temporal model to predict annual malaria risk for children from 2 to 5 years of age in the country, with 1 km spatial resolution between 1990 and 2015. The researchers claimed that the model was efficiently predicting the risk of malaria for children in the regarding age range (Macharia, et al., 2018).

One of the main issues that rise related to applying MLR is that the mode can severely affected by the multicollinearity of the predictors, as a result produces a poor predictive accuracy. Therefore, it is important to deal with this issue prior to fitting regression. Some recent papers were able to show the bright side of PCA in terms of dealing with such issue. A paper from environmental science journal presented an application of PCA to select the most relevant and highly correlated variables, then predict the concentration of PM10 using Artificial Neural Network (ANN). The method was tested by using several time series, such as solar radiation, vertical wind speed, atmospheric pleasure, PM2.5, benzene, NO and PM10 in Varanasi, India. After PCA-ANN model was applied to the PM10 data, its result was compared against MLR. The result reveals that PCA-ANN forecasted PM10 with 9.88% of Root Mean Square Error (RMSE), which is a better prediction than the MLR model as per the result provided in the paper.

3 Study Area and Data

3.1 Study Area

For spatial interpolation we have considered the mean PM10 concentrations for each of the four sites. Figure 1 shows the area of interest which is considered for this study region, and the red points emphasize the target locations where the data is collected.

Figure 1: Map of study region and the dots represent data collection sites.

Table 1: Geographical coordinate of the study area.

Station Name	Longitude	Latitude	Average PM10
Ashburton	-43.9122	171.7552	17.032
Geraldine	-44.1002	171.2415	17.400
Anzac Square	-44.4045	171.2497	21.353
Washdyke	-44.3567	171.2363	17.642

3.2 Study Area Background

Washdyke: this is one of the suburbs where most industrial activities take place. Therefore, the smoke emission of the industries might influence the level of air ambient PM10 concentration in the region. It is also a site where state highway pass through, which means that the carbon footprint also might significantly impact the concentration of PM10 in the regarding area.

Ashburton: its known for having dry temperature climate. However, it is located at the higher altitude of the city of Christchurch, thus the city experiences a significant temperature variation, sometimes exceeding $30^0 c$, and whilst winter experiences regular frosts and annual snowfall.

Geraldine: its is one of the most agriculturally based sites, where dairy farm is at the heart of the city. Its wide area is covered by forest, which might be useful in lowering the carbon footprint.

Anzac Square: this is another agricultural site with highest population size (29,000). The population size might influence the concentration of PM10, specially during the wintertime, many majorities of the people might use home heating equipment.

3.3 Data Exploration

The data set that is being considered for this study comprises PM10, PM2.5, SO_2,CO, PMcoarse and other metrological factors such as (wind speed (m/s), relative humidity, temperature ground-top mast, temperature 6m, temperature 2m, wind max m/s and temperature ground). The data set is provided by Environment Canterbury Region Council and obtained from (Ecaht). Here we are considering a data collected from 4 different stations in southern Canterbury (Anzac Square, Ashburton, Geraldine and Washdyke), from 30/12/2014 to 1/1/2019.

One of the biggest problems that we encounter while dealing with such analysis is missing values. Data might be missing due to various reasons, such as malfunction of the monitor. In consequence, despite the huge amount of data that is being collected from the four stations, there are still a number of missing values that exist in our dataset, which might be troublesome for our knowledge discovery. SO_2 and CO are the two main variables with great deal of missing values from Ashburton station (>50% of observation is missing), majority of observation of Wind direction and SO_2 are missing from Anzac Square (>50%), Relative Humidity, PM2.5 and PMcoarse are the main features with huge number of missing values from Geraldine station (>50%) and PM2.5, PMcoarse and wind direction are the three variables with huge number of missing values (> 50%). Therefore, we will have to drop the corresponding variables from the dataset after analysing their effect on the concentration of PM10.

Table 2: Features with missing and negaative values in Anzac Square.

Feature Name With Missing Values	**Missing Values (%)**	Feature Name With Negative Values	Negative values (%)
Wind Direction	13.59	PMCoarse	7.87
SO2	6.77	PM2.5	0.87
PMCoarse	2	SO2	0.87

Table 3: Features with missing and negaative values in Geraldine.

Feature Name With Missing Values	**Missing Values (%)**	Feature Name With Negative Values	Negative values (%)
Wind Direction	40.31	PMCoarse	3.31
SO2	29.44	PM2.5	1.32
PMCoarse	28.61	SO2	0.08

Table 4: Features with missing and negaative values in Ashburton.

Feature Name With Missing Values	**Missing Values (%)**	Feature Name With Negative Values	Negative values (%)
Wind Direction	19.42	PMCoarse	3.25
SO2	6.77	PM2.5	0.44
PMCoarse	2	SO2	0.29

Table 5: Features with missing and negaative values in Washdyke.

Feature Name With Missing Values	**Missing Values (%)**	Feature Name With Negative Values	Negative values (%)
Wind Direction	94.23	PMCoarse	7.17
SO2	53.54	PM2.5	0.30
PMCoarse	2	SO2	0.27

3.4 Pre-Processing

Prior to any analysis, we prepared the data in such a way it is compatible to the as per the algorithms we implement. This is one essential aspect of our analysis because most algorithm's prediction performance relies on the quality of dataset. The type of data preprocessing methods that are applied on our dataset is presented as following:

Scaling: This method is used to make the variance of the data consistent. Features with highly varying magnitude will be problematic. Especially, this problem is even worse for K-means clustering algorithm as we consider Euclidian distance to calculate the distance between data points. Hence, this method circumvents the problem of high variation in the dataset by reducing the variance of the magnitude of each feature.

Normalization: Another method that we applied on our dataset is normalization. This is a method of making the scaled datapoint span between 0 and 1. Once we scaled our data, we do not only want to achieve reducing the variance of the data but make the values positive because some of our variables must remain strictly positive values (for example PM10). Hence, normalization is a preferable method to deal with this issue.

Missing values: The important aspect of data analysis is to get the clean training data for modelling. There were several instances where we encountered various missing values for some attributes in Anzac, Ashburton due to which we had to remove all the corresponding values in the rows for certain missing values in the columns. In Geraldine, relative humidity had 14166 missing values, in this case we removed the whole column. In Washdyke, CO had 18789 missing value and SO2 has 33063, both columns were eliminated from the data frame as removing rows would also delete important data from other variables.

Negative Values: this is problematic for knowledge discovery purpose because negative PM10 concentration is not realistic as the minimum it can get is zero, but not less than that (in real world scenario). Therefore, we had several negative values in our dataset, which must be removed to make our result interpretable.

Table 6: Over All data points before and after removal of missing values.

	Anzac Square	*Ashburton*	*Geraldine*	*Washdyke*
Complete Dataset	35316	35128	35135	35087
After Removing Missing Values	27811	25391	17652	25405
Daily Average Observations	1436	1187	906	1260

3.5 Statistical Analysis of The Data

Here we perfo

Table 7:Figure 2: Summary of the seasonally and yearly averaged data for Anzac Square region.

Anzac Square	**Mean PM10**	**Mean PM2.5**	**Mean CO**	**Wind Speed**	**SD PM10**	**CV PM10**
Summer	17.4	8.05	0.076	2.13	10.69	0.614
Autumn	21.04	12.74	0.244	1.69	19.13	0.909
Winter	33.49	24.69	0.538	1.48	31.98	0.954
Spring	17.71	8.77	0.132	2.102	13.66	0.771
2015	27.5	16.7	0.25	2.07	23.2	0.843
2016	23.1	12.6	0.23	2.08	20.4	0.883
2017	20.8	11.1	0.17	2.10	17.5	0.841
2018	19.4	9.68	0.17	2.05	16.5	0.85

PM10 and PM2.5 concentrations in ANZAC Square are at their highest in winter and autumn. The coefficient of variation for PM10 in winter is 0.95, this means that in winter PM10 concentrations are highly dispersed. Overall CV for PM10 is 0.96 but autumn and winter are the main contributors towards this number.

Table 8:Figure 2: Summary of the seasonally and yearly averaged data for Ashburton region.

Ashburton	Mean PM10	Mean PM2.5	Mean CO	Wind Speed	SD PM10	CV PM10
Summer	14.56	5.07	0.071	2.466	9.22	0.633
Autumn	17.6	9	0.193	2.042	15.69	0.891
Winter	23.22	16.89	0.375	1.915	22.84	0.983
Spring	15.31	6.94	0.11	2.417	11.02	0.719
2015	15.9	6.34	0.07	2.59	10.4	0.654
2016	16.9	8.84	0.14	2.39	14.0	0.824
2017	16.0	7.80	0.13	2.38	12.7	0.793
2018	15.1	7.73	0.13	2.34	11.3	0.748

In Ashburton PM10, PM2.5 and CO concentrations are also at their highest in winter followed by autumn. PM10 concentrations are also highly dispersed in winter compared to other seasons. Average Wind speed is highest in summer, when pollutant concentrations are low. However, in winter when average pollutant concentrations are at their highest, average wind speed is at its lowest.

Table 9:Figure 2: Summary of the monthly and yearly averaged data for Geraldine region.

Geraldine	PM10	PM2.5	CO	Wind Speed	SD PM10	CV PM10
Summer	14.19	4.91	0.05	1.27	10.04	0.70
Autumn	16.83	9.67	0.16	1.08	14.03	0.83
Winter	23.76	18.7	0.36	1.06	19.86	0.83
Spring	14.58	6.86	0.11	1.38	10.67	0.73
2016	17.1	10	0.15	1.22	13.4	0.78
2017	16.4	8.97	0.13	1.38	12.3	0.75
2018	15.5	8.72	0.11	1.43	12.7	0.81

In Geraldine, we observe the same pattern for average accumulations of pollutants. However, the dispersion of PM10 concentrations in Geraldine are lower than Anzac Square and Ashburton.

Table 10:Figure 2: Summary of the monthly and yearly averaged Washdyke data.

Washdyke	PM10	PM2.5	Wind Speed	SD PM10	CV PM10
Summer	17.90	4.85	2.71	1.78	0.72
Autumn	18.29	5.63	2.14	1.68	1.00
Winter	16.65	7.06	1.76	1.50	1.01
Spring	16.34	5.02	2.61	1.87	0.85
2015	20.9	7.16	2.93	17.4	0.83
2016	19	6.54	2.65	14.3	0.75
2017	17.9	5.85	2.71	13.2	0.73
2018	17.1	5.55	2.69	13.3	0.77

Contrary to other locations, average PM10 concentrations in Washdyke are not at their highest in winter. Although PM2.5 and CO concentrations are still highest in winter followed by autumn. This unexpected behaviour is mostly likely due to the large number of missing values present in the Washdyke dataset. Coefficient of variation for PM10 is 1 in winter and autumn, which means that the data is highly dispersed. This may also be due to the large number of missing values, which can

unbalance the data (sometimes concentrations are very high and sometimes very low). Overall, we can observe from the tables above that average PM10 concentrations are always higher than PM2.5. Wind speeds are usually high in summer and spring and during this time PM10 and PM2.5 concentrations are at their lowest. Out of the four locations, concentrations for all pollutants are highest in Anzac Square across all seasons. In the next section, we analyse PM10 concentrations with close retrospection as this is the pollutant with the highest concentration.

3.6 Time Series Analysis of PM10 And Other Pollutants

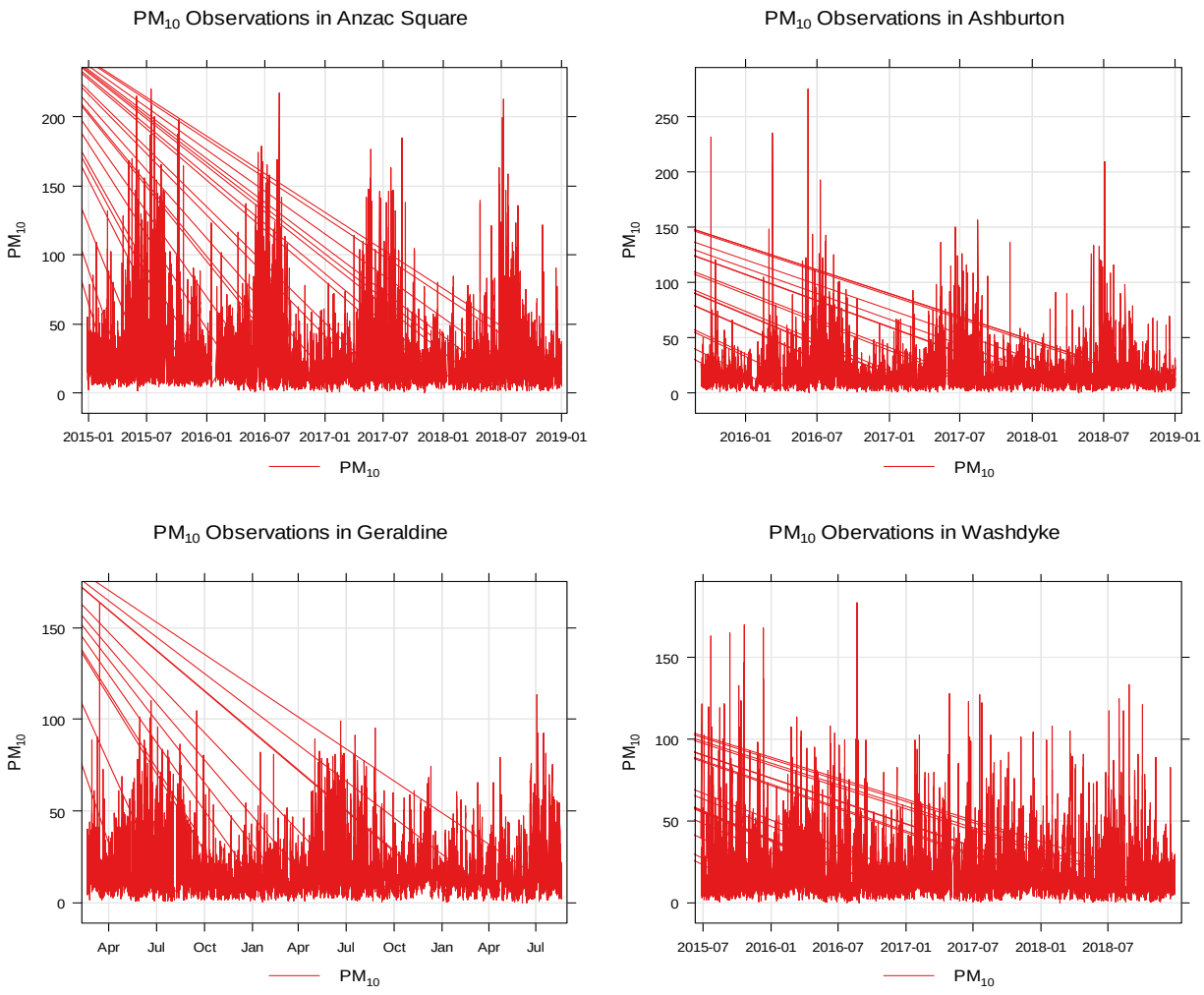

Figure 3:Time-series plot of hourly PM concentration in four regions.

As we can see from figure 2, despite the seasonal variation, there seem to be a random spike that occur over certain time interval. This is even worst in the region of Washdyke as it does not follow a seasonal pattern at all. We summarize this is occurring because Washdyke is known to be the most industrial region, compared to other monitoring sites, as a result the concentration of PM10 in this region tends to be arbitrary throughout the year. Unlike other regions, due to the removal of missing values in the region of Geraldine's data, we cannot see the seasonal pattern for the entire years of study. Furthermore, since the plot is in high resolution (hourly observation), the size of the data is

concealing the seasonal pattern. Hence, we averaged monthly PM10 concentration of the all regions and the result is presented below in figure 3.

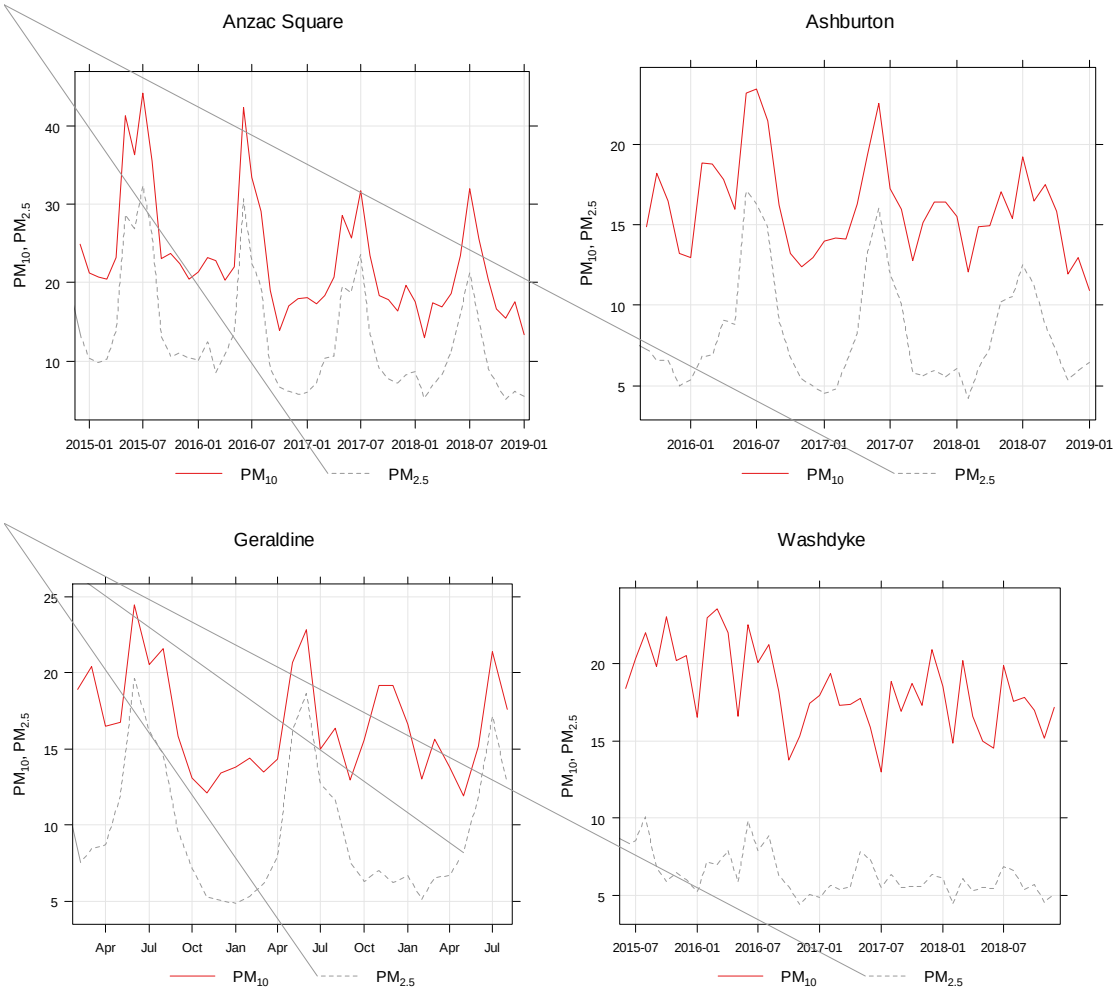

Figure 4: Monthly averaged PM10 concentration in the four regions.

Concentrations of PM10 across the four regions have distinctive seasonal patterns and variation (figure 3). Seasonal patterns are more apparent in the three regions of Canterbury (Anzac Square, Ashburton and Geraldine) than Washdyke. The general trend of PM10 concentration seems to decline as we go further in time, which might be due to the lesser utilization of traditional home heating appliances. Despite the general downward trend, the concentration of PM10 is significantly higher during the wintertime (April-September), where the highest PM10 concentration is observed across all the desired years. This plot authenticates our assumption of the ambient air PM10 hit the peak during the wintertime. On the other hand, the concentration of PM10 for all regions during the summertime is observed to be the lowest. However, PM10 concentration in the region of Washdyke is following a relatively less seasonal pattern. We assume this is because Washdyke is said to be the most industrial suburb compared to the other three regions, thus the emission of CO2 into the air of Washdyke regions seems to contribute to the level of air pollutants. A clear indication of major contribution of air pollution in all the target regions is the overall PM10 concentration in that particular region. This leads us to drill down to the hourly data and understand the major time duration contributor for PM accumulation for all the regions individually.

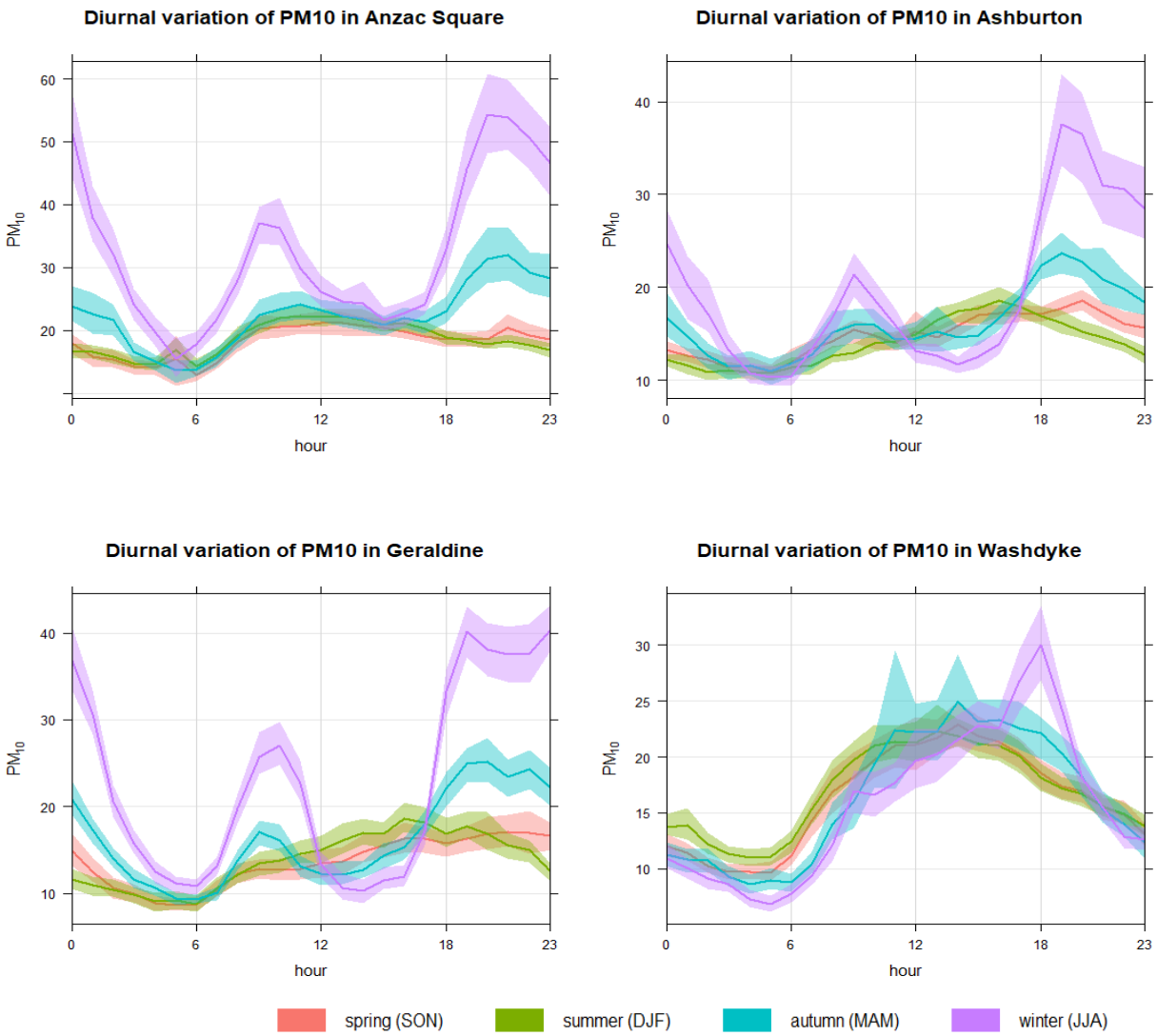

Figure 5: 24-hours of PM10 concentration of four regions.

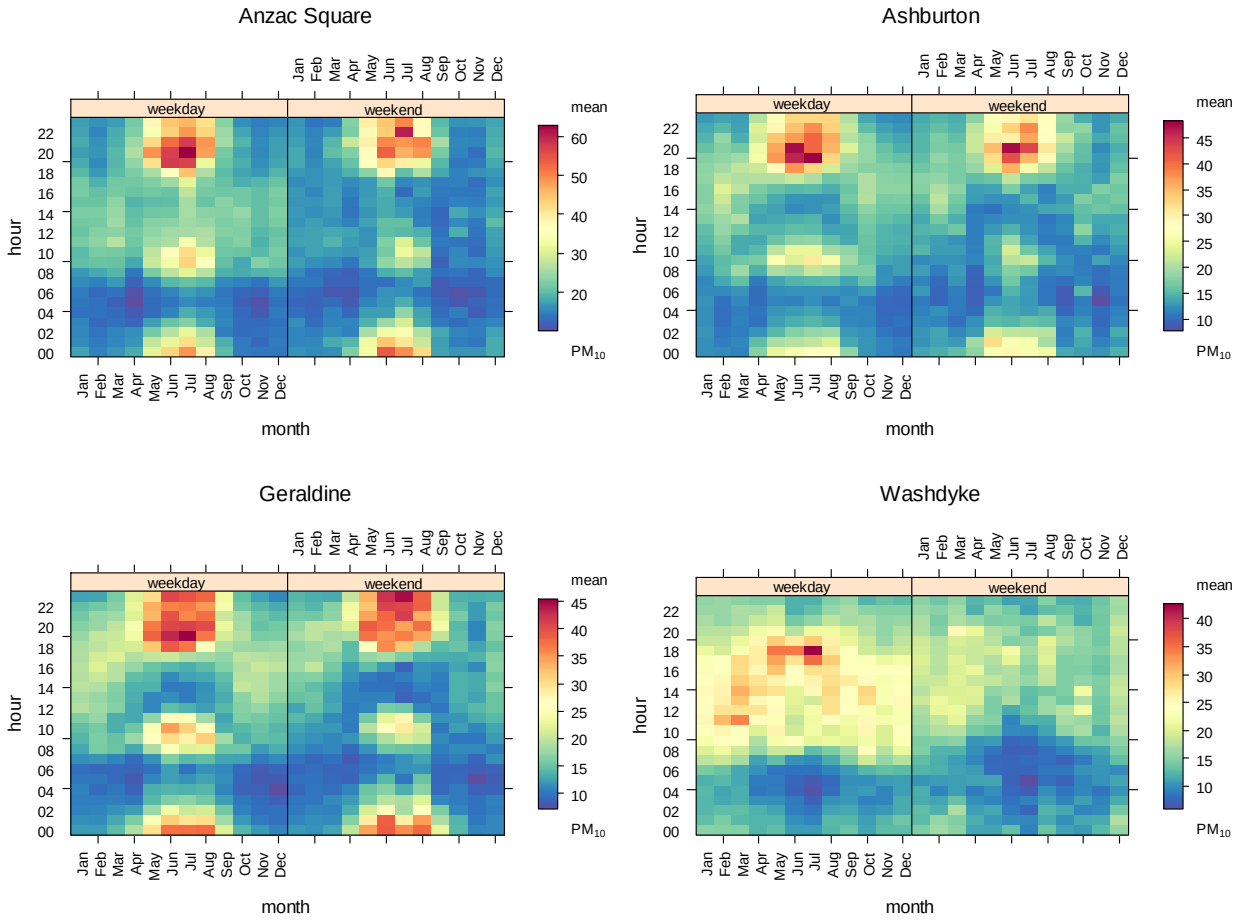

Figure 6:Heat map of the 24 hours and 12 months, along with weekend and weekday of PM10 concentration in the four regions.

The heat map in figure 5 illustrates that the ambient air PM10 hits its peak at night (> 40 μmg), in winter season (April - August). The concentration of PM10 in the weekend and weekday does not seem to show much difference for all regions. However, this aspect of the scenario is different for the region of Washdyke because ambient air PM10 is higher in the weekday than weekend. As we discussed earlier, Washdyke is an industrial location and most these industries are operated during the weekday, therefore, causing the concentration PM10 to be higher around the mid-day, during the weekday. We assume there might strong correlation between the air pollutants being emitted from the industries based in Washdyke and concentration of PM10, even though a proper statistical analysis must be considered to support our assumption.

Next, we investigate the contributing factors to PM10 concentrations using correlation analysis.

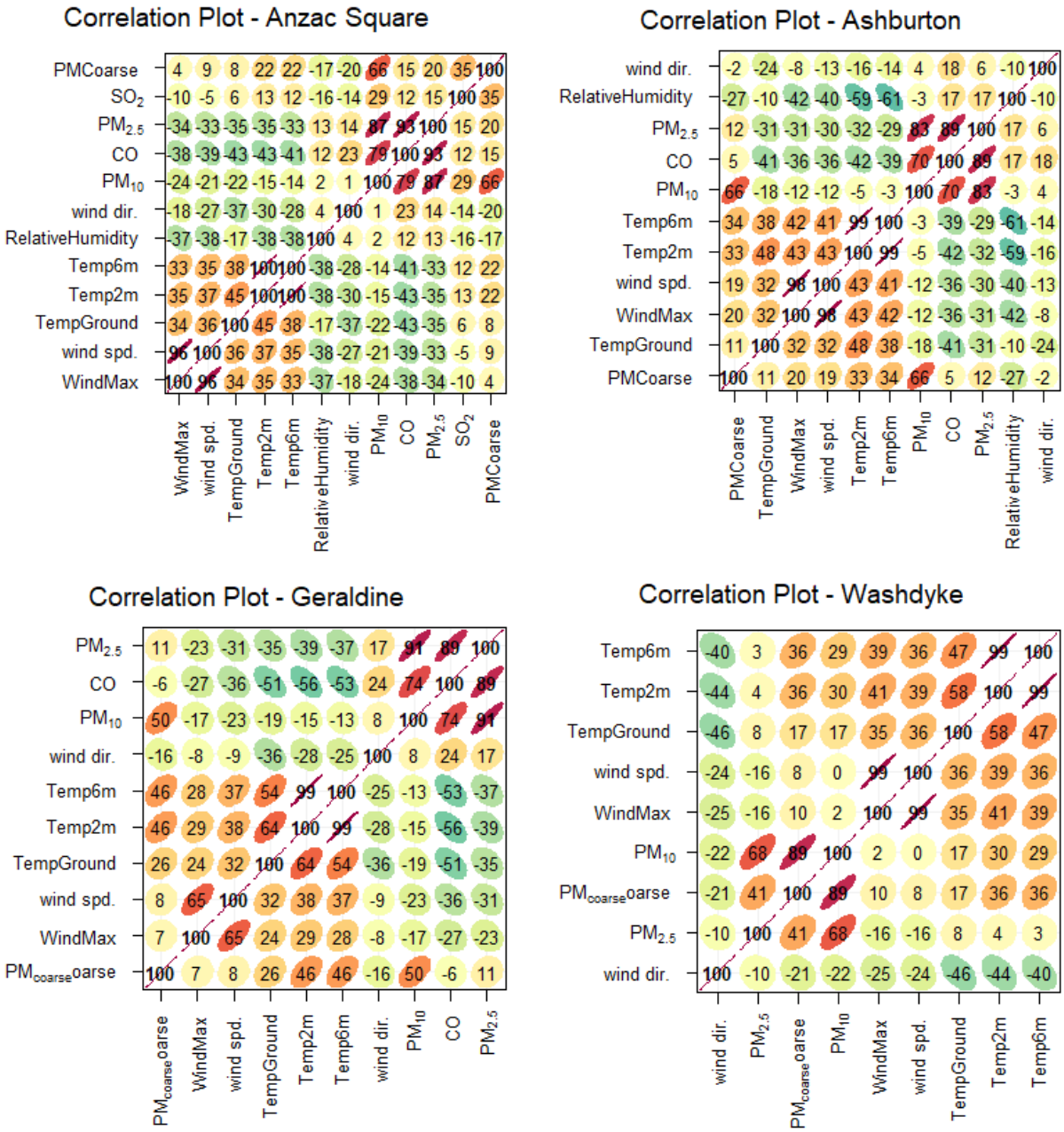

Figure 7:Correlation plot between the entire variables.

The plots above show the correlation between pollutants and metrological factors for each location considered in this study. The ellipses are representations of scatter plots, while straight line at 45 degrees represents a perfect positive correlation. Zero correlation is represented with a circle shape. Numeric or correlation coefficient shows the strength of the correlation, whereas the sign determines whether the association between the corresponding variables is positive or negative.

The relationship of CO with PM10 is significantly higher (>70%) for autumn and gets further escalated in winter (>80%). This is expected due to their common emission source such as home heating burning wood, coal, natural gas and gasoline. This illustrates a direct relationship between the incomplete combustion of carbon containing fuels and ambient air PM10. Although a strong positive correlation

might be observed for PM10 and CO, various meteorological factors such as Temperature around 6metres, and Wind Direction also indicate a weak negative correlation that increases in the course of winter season. Overall, most meteorological parameters (humidity, temperature and wind) have a very weak contribution to the degree of PM10. This means that most of the PM10 concentration across all seasons are caused the emission of CO, especially in winter and autumn.

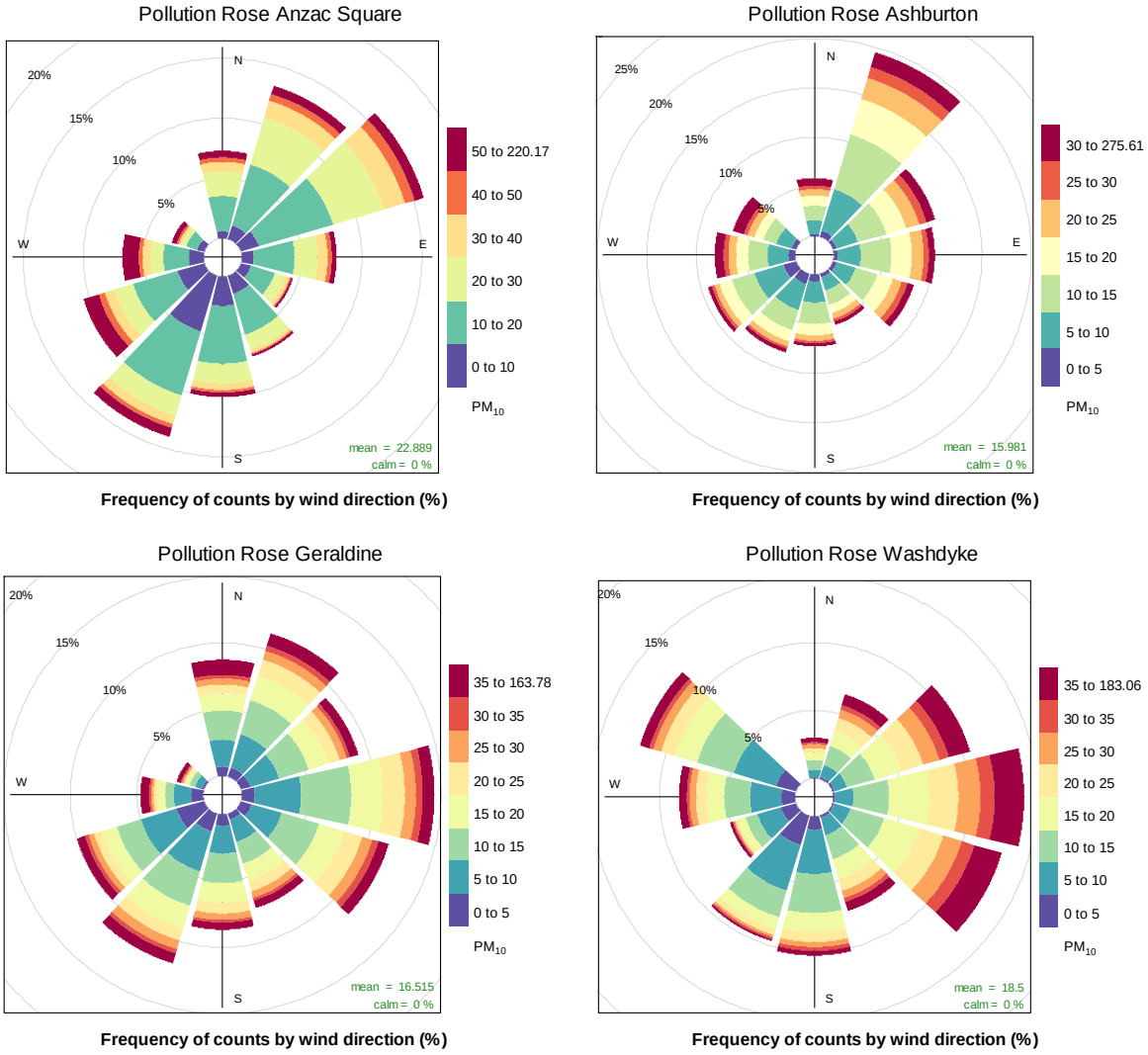

Figure 8:The effect of wind direction on the concentration of PM10 in the four regions.

Figure 3 depicts the relationship between wind direction and PM10 concentrations for different seasons in a year. The rose shows that wind blows to the north for most period of time across all seasons of the year, which is more obvious in winter where approximately 30% of the time, the wind is blowing to NNW or NNE. Moreover, during all seasons, the wind rarely blows from north west and south east and PM10 concentrations are also very low during this period. In winter PM10 values can also get quite high as about 10% of the time the wind blows to the north and PM10 concentration is above 50 µmg. In all other seasons, PM10 concentrations are between 0 and 20µmg for majority of the time. Inspection of the plots reveals a direct proportionality between the PM10 concentrations and the direction of the wind being North

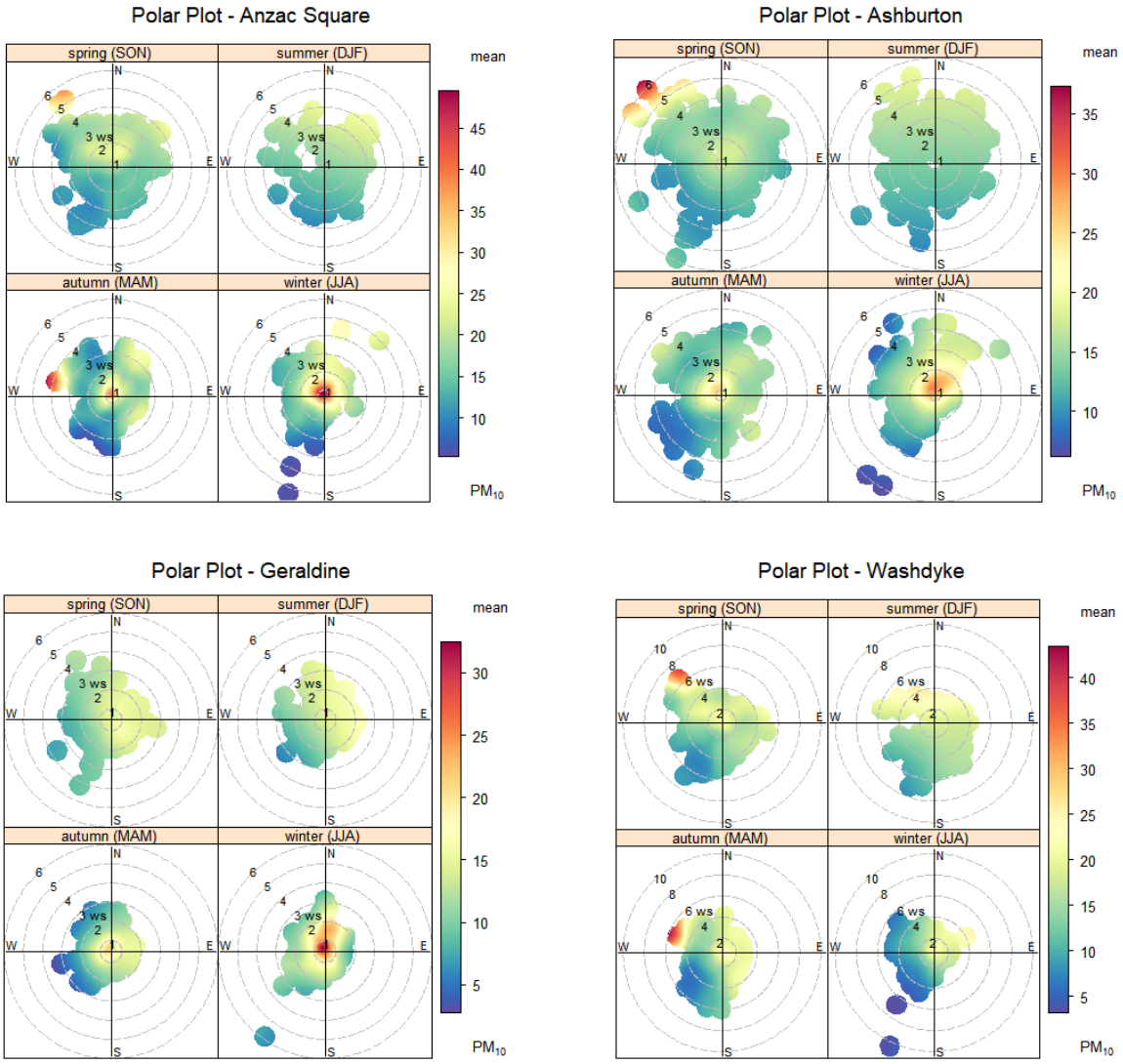

Figure 9:The effect of wind direction on the concentration of PM10 for different seasons of the year.

3.7 Significant Findings

PM10 concentrations are highest throughout the day in winter. This confirms the agreed consortium that lower temperatures enhance the concentration of PM10 in the atmosphere. The concentration hits its apex during the night-time. An increase in the usage of fossil-fuel based heaters (both industrial and household) contributes to this spike in the concentration. There are moderately regular fluctuations in the concentration during the spring and summer season, except in the city of Washdyke. In Washdyke, winters see a spike in concentration during the late evening hours. Although, the concentration of PM10 during autumn, spring and summers is not dismissible. In fact, the concentration of PM10 is almost constant throughout the years. The initial data exploration reveals the lack of CO and SO2 data. Hence, there is a case of missing values for Washdyke. Some of the values are negative which can be due to the sensor failures faults and calibration issues which can be a part of the future analysis.

4 Method

In this section, we provide all the methods and models we adopted to analyse the subject matter.

4.1 Spatial Interpolation

It is often unfeasible to collect complete geographical information of pollutant concentrations, due to the high experimental expenses. In such cases, observations from number of monitoring sites can be used to understand the distribution of air pollutants in a given study area. Most countries cannot afford the equipment to collect the required data for all sites of the country, this is where spatial Interpolation becomes beneficial.

Interpolation uses known values from discrete points in space to predict unknown values for the creation of a smooth surface. Interpolation method enables us to understand the distribution of air pollutants, such as PM10, CO and so on in different sites using few known points. There are different types of geographical Interpolation techniques. In this paper, we implement three types of interpolation techniques (Inverse Distance Weight, Ordinary Kriging and Radial Basis Function) to predict PM10 concentrations in the southern Canterbury region using data collected from four sites (Anzac Square, Washdyke, Geraldine and Ashburton). After implementing these techniques, we perform a comparison analysis to find the best performing method. We suppose that the results obtained from this paper will provide governmental or other entitled organizations with good information on the concentration of PM10 across the study region. This is beneficial as air pollution is still a concerning matter in NZ and if the concentration is above the critical threshold, immediate action can be taken.

Usually the quality of interpolation must be assessed before the interpolation methods used to interpolate set of unknown values. For this case, cross-validation is one the best preferred metrics to compare multiple interpolation methods. Cross-validation procedure works by ignoring some of the observations and using the remaining observations to predict the ignored observations in the data set using one of the interpolation methods. The process is iterated for each observation in the data until it finds set of interpolated values. Validation should be carried out before producing the final surface, where it helps in making an informed decision as to which model provides the best predictions should have (Hussain, 2013).

4.1.1 Inverse Distance Weighting (IDW)

IDW is a type of interpolation technique that is used to estimate unknown points using the set of known points. One of the main assumptions of this method is that the closest points are more associated than those which are far apart. Another feature that is considered by this method is the spatial autocorrelation. In other words, if there is much PM10 concentration in one specific area, then there is also PM10 concentration may be few miles away from the underline area.

IDW interpolation has multifaceted benefits. Most importantly, it is well known for its flexibility in terms of predicting the unknown points as it allows us to set up the interpolation in different ways. For instance, we can specify a fixed number of points and use the closest points in order to predict the unknown point. In the meantime, we can also specify the search radius and then find the closest point in the radius then predict the unknown point using the closest known values within the search radius area. Further advantage of this method is enables us to set up a barrier to block certain points from being consumed during the process of interpolation. The barrier prevents the method from finding more input values.

The most important parameter of IDW is “power”. It is a parameter that determines the smoothness of the interpolation. Higher power causes the known values to influence the interpolation and the

values become more localized. On the other hand, small value of power decreases the influence from the known points and the values are averaged out (higher peak). (Hemalatha, Wooi-Nee, Sin, & Mohammad, 2016)

Mathematically this is represented as:

$$z_p = \frac{\sum_{i=1}^{n} \frac{z_i}{d_i^p}}{\sum_{i=1}^{n} \frac{1}{d_i^p}}$$

Where p is power, d is the distance between known and unknown points and z is value of known points.

4.1.2 Radial Basis Function

RBF is a popular interpolation method that relies on functions that are radially symmetry and is a special case of spline. It is well known for being a meshless interpolation method. Put simply, this method does not require triangulation or any other sort of mesh that links the desired data sites. In addition, it's capability to handle randomly scattered data points in order to easily generalize multiple space dimensions made it popular in a variety of application domains. Hence, unlike other interpolation methods, RBF does predict values less than the minimum original value in the cross section below or greater than the maximum known value (Pengwei, et al., 2019).

RBF is an interpolation method used to produce smooth surface using large data points. However, the function gets highly negatively impacted by high variations in short distance surface. Hence, this method is recommended for data set with gently varying surface.

There is a variety of RBF methods that can be used to predict the unknown points, such as such as completely regularized splines, thin-plate splines, splines with tension, and inverse multiquadric. However, in this paper, we considered completely regularized spline, which is the function in ArcMap platform.

4.1.3 Ordinary Kriging (OK)

Ordinary Kriging is another interpolation method used to predict the value at unknown spatial points as a linear combination of neighbouring available data. It relies on the values of adjacent variables, using variogram to merge the original data and estimate region's unknown sampling point value (Yan, Yan, Xingbang, & Li, 2016). It is a weighted average interpolation method. The weights are not only distance-based (distance between the sample points and unknown location), but also in the general spatial adjustment of all the sampling points.

The method has two main phases:

1. Estimating the autocorrelation of the sample and assigning weight parameter.
2. Compute the predicted values.

Minimizing the mean square estimation errors of the partial correlation determine the weight parameter. The predicted value can estimate over the maximum or below the minimum observation point, but final surface does not pass through the sample points. OK does not support existing trends in the data, in such cases the data is differenced to remove trends (Hemalatha, Wooi-Nee, Sin, & Mohammad, 2016).

4.2 Spatio-Temporal Interpolation

Spatio-temporal analysis is a dynamic framework that is used to analyze different phenomenon collected across different space as well as time. It is an emerging area of research as more advanced computational techniques have been introduced. Spatio-temporal analysis has a wide range of applications such as economics, biology, chemistry, medicine, forestry and so on. Spatio-temporal analysis can be challenging sometimes as space has a different direction (up, down, East, North, West, South), while time has a mere single direction (forward). Hence, the practice of combining these two factors can be convoluted. Spatio-temporal analysis has several benefits over pure statistical analysis as they enable the researchers to concurrently investigate variation of patterns over time and reveal any anomaly patterns. The consideration of space-time interaction terms may also assess data grouping that may be vital in order to indicate environmental hazards or persistent errors in the process of data collection.

In this milestone, we consider two types of models to predict PM10 concentration measured in the city of Canterbury, South New Zealand. Due to computational restrictions, we only consider one month of PM10 concentration, where maximum and minimum PM10 concentration is observed. We use two spatio-temporal models to predict the PM10 concentration. While these robust models are used, the potential shortcoming of this milestone can be that we only have four stations where the PM10 concentration is measured. Hence, we suppose the shortage of sample might sabotage the prediction quality of these models.

4.2.1 Spatio-Temporal Semi Variogram:

Semi variogram is a Spatio-temporal analysis method, which is a plot of semi variance versus range that enables us to quantify the spatial-temporal correlation of an attribute. Semi variogram expresses the regionalized variable's overall rate of change along the target orientation (Fanchi, 2010). Semi variance is a measure of the degree of Spatio-temporal association between values in two different locations. Unlike spatial case, in the case of spatio-temporal case, the separating distance are pairs of spatial and temporal distance yielding a variogram surface.

The most important features of Spatio-temporal variogram are the following (Allard, Bel, Gabriel, Opitz, & Parent, 2017):

- The predictions are dependent on space and time.
- Explains the variogram regularity which is the nugget effect and smoothness that differs in space and time.

In this study, we fit two different types of models to spatio-temporal variogram model, namely separable and metric.

Extending the variogram to a two-place function for spatio-temporal random fields

$\boldsymbol{Z}(\boldsymbol{s},\boldsymbol{t})$: $\boldsymbol{Y}(\boldsymbol{h},\boldsymbol{u}) = \boldsymbol{E}[\boldsymbol{Z}(\boldsymbol{s},\boldsymbol{t}) - \boldsymbol{Z}(\boldsymbol{s}+\boldsymbol{h},\boldsymbol{t}+\boldsymbol{u})]^{\boldsymbol{2}}$ at any location (s,t), where s is space and t is time.

An empirical version of this model is as follows:

$$\hat{\gamma}(h,u) = \frac{1}{2|N_{h,u}|}\sum_{(i,j)\in N_{h,u}}(Z(S_i,t_i) - Z(S_j,t_j))^2 \quad (1)$$

While $N_{h,u} = (i,j) : h - \epsilon_s \leq || s_i - s_j|| \leq h + \epsilon_s$ and

$= (i,j): u - \epsilon_s \leq t_i - t_j|| \leq u + \epsilon_s$

Nugget effect: is a measure of variability in the samples that are closely spaced. Spatial variation can be dependent on the direction of location (anisotropic) or independent (isotropic) (Robert, Bacciu, & Kathy, 2012).

4.2.2 Separable Covariance Function

Under the assumption of isotropy and stationarity, the separating distance h between two locations is a covariance function $C(h)$. A spatio-temporal covariance function is considered as a function of a spatial and a temporal distance $C(h,t)$.

A separable covariance function is assumed to fulfill $C_s(h,u) = C_s(h)C_t(u)$. This is in general a rather strong simplification. Its variogram is given by the following function:

$$Y_{sep}(h,u) = nug_{\cdot 1h>0,u>0} + sill(\gamma_s(h) + \gamma_t(u) - \gamma_s(h)\gamma_t(u), \quad (2)$$

where y_s and y_t are spatial and temporal variograms without nugget effect and a sill of 1. The overall nugget and sill parameters are denoted by "nug" and "sill" respectively.

4.2.3 Metric covariance function

The metric kriging extends the 2-D geographical space into a 3-D spatio-temporal space. The isotropic space with consistency in all orientations can be achieved by rescaling the temporal domain to match the spatial domain. A joint covariance model C_j for spatial, temporal and spatio-temporal distances is used which is expressed as

$$C_m(h,u) = C_j(\sqrt{h^2 + (k.u)^2}), \quad (3)$$

The variogram evaluation includes the nugget effect expressed as:

$$Y_m(h,u) = \gamma_j(\sqrt{h^2 + (k.u)^2}) \quad , \quad (4)$$

Where y_j is the unknown variogram.

4.2.4 Spatio Temporal Kriging

After the construction of the semi variogram and fitting the data to the model from the given data, its viable to do predictions using spatio-temporal kriging. This can be modelled using the spatio temporal kriging equations that calculates the weights for the regions observed, the optimal spatio-temporal predictor and the variance of the predictions which specifies the accuracy of the predictions.

4.3 Data Mining and Machine Learning

4.3.1 K-means Clustering

K-means clustering is one of the most widely adopted clustering algorithms (Matthew Kyan, 2014). K-means clustering performs clustering by grouping the observations into K-prototypes. Initially, an arbitrary position of K-prototypes is assigned prior to pooling observations into K groups using nearest neighbour. The algorithm works iteratively in search of the centroid of the clusters until most of the data points in a cluster are similar, according to their Euclidean distance.

In a recent research paper, K-means clustering algorithm has been used to group data so that the result can be fed in an Artificial Neural Network to forecast the concentration of air pollutants (PM10 and PM2.5) (Fabiana Franceschi, 2018). Another paper applied K-means clustering algorithm to analyse the relationship between PM components and breast cancer. Predictive K-means was

implemented to assign predictors to a group defined by PM2.5 components profile to evaluate the impact of heterogeneity in the air pollution mixture (Aj White, 2019).

4.3.2 Multiple Linear Regression (MLR)

Multiple linear regression (MLR) is one of the most common types of regression model. MLR is a useful method to explain the relationship between a continuous dependent variable and one or more independent variables. The predictive variables or independent variables can be categorical or continuous.

A study conducted in Malaysia applied an MLR model to predict the daily PM10 concentration as a function of temperature, humidity, wind-speed, and wind direction on the data that comprises an observation from 2006 to 2010. In addition to MLR, the researchers implemented various models, such as stepwise regression (SR) and principal component regression (PCR) and the result was benchmarked based on R2. According to the result established by the models, PCR found to be the best performing model (Amina Nazif, 2017). In another PM oriented paper conducted MLR to simulate the daily variation of PM10 concentrations for the particular sea breeze cases in Split and Kastel-Sucaurac, which are based in the eastern part Adriatic Cost (2007-2009). The result obtained from this paper shows that the MLR simulation matches the PM10 concentration measurements during the selected sea breeze cases in the desired study areas (R2 =0.77and Index of argument AI=0.89) (Tanja Trošić1, 2017).

4.3.3 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a statistical method that relies on orthogonal transformation to reduce the number of dimensions. PCA takes advantage of the correlation between the variables in order to create a new linearly uncorrelated variables known us principal components. Principal components are ranked according to the amount of variations each of them can capture. The first component captures highest possible variance, and each succeeding component has highest variance under the restriction of each component must be an orthogonal to the preceding components.

PCA is a well-known method to improve quality of model prediction as well as for visualization. A paper by Wie and Jingyi applied PCA to forecast the daily concentration of PM2.5 in China. PCA was applied in order to extract the features and reduce the dimension of the data as an input. Then Support Vector Machine was applied on the reduced features, which is fine-tuned by cuckoo search to improve its generalization. The result shows that (Wie Sun, 2017). Another paper from environmental science journal presented an application of PCA to select the most relevant and highly correlated variables, then predict the concentration of PM10 using Artificial Neural Network (ANN). The method was tested by using several time series, such as solar radiation, vertical wind speed, atmospheric pleasure, PM2.5, benzene, NO and PM10 in Varanasi, India. After PCA-ANN model was applied to the PM10 data, its result was compared against MLR. The result reveals that PCA-ANN forecasted PM10 with 9.88% of Root Mean Square Error (RMSE), which is a better prediction than the MLR model as per the result provided in the paper.

5 Result

5.1 Spatial Interpolation

In order to interpolate unknown surface areas of the regions, we used Ordinary Kriging, RBF and IDW interpolation methods. However, we know that some of our interpolation methods, such as OK are highly prone to the existence of trend in the data set, thus it is important to investigate the behaviour of the data before applying the models on it. Therefore, in this sub section, we present the trend analysis as well as surface interpolation of unknown surface.

5.1.1 Trend and Data Normality Analysis

First, we perform statistical analysis to assess the quality of the data set. Subsequently, we provide the interpolation results acquired from the three different interpolation techniques.

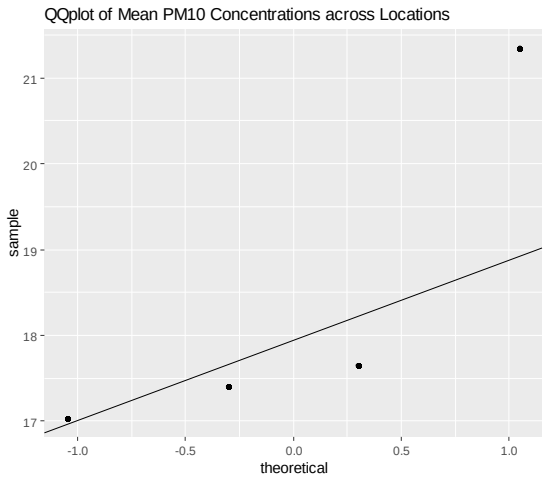

Figure 10: Quantile plot of PM10 concentration across all locations.

The Figure above shows the quantile plot of the average PM10 concentrations in the regarding study areas. From the plot, we can see that the distribution is skewed to the left due to the maximum average PM10 obtained in the region of Anzac Square. On this basis, we can understand that the data is not normally distributed.

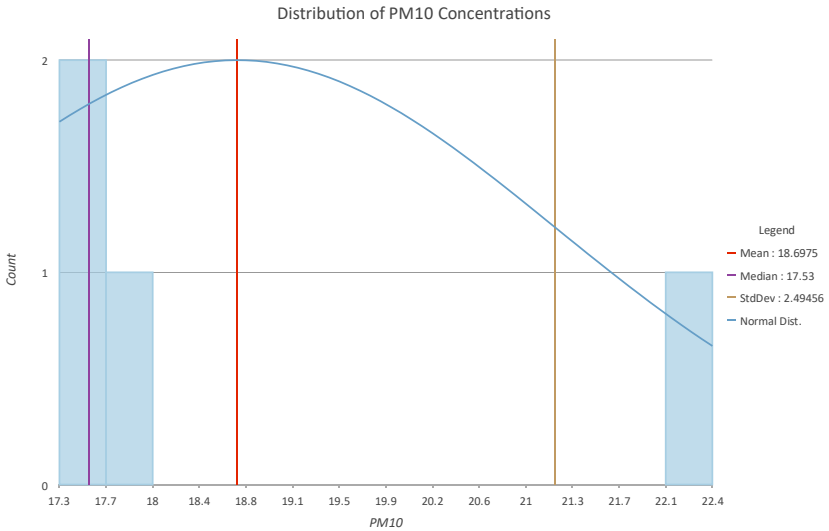

Figure 11:Histogram of PM10 concentration.

The histogram also shows that data is not normally distributed.

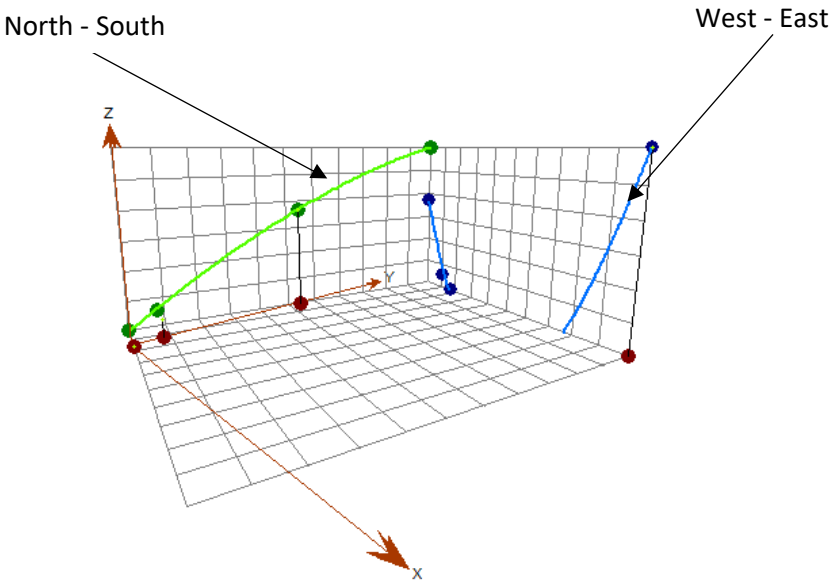

Figure 13: Trend of PM10 concentration across the four monitoring sites.

The above graph depicts the trend that existed in the data set. As we can see, there is a strong trend in the data set, which is an issue for some interpolation methods, such as ordinary kriging as it is prone to such behaviour of data.

5.1.2 Inverse Distance Weighting (IDW)

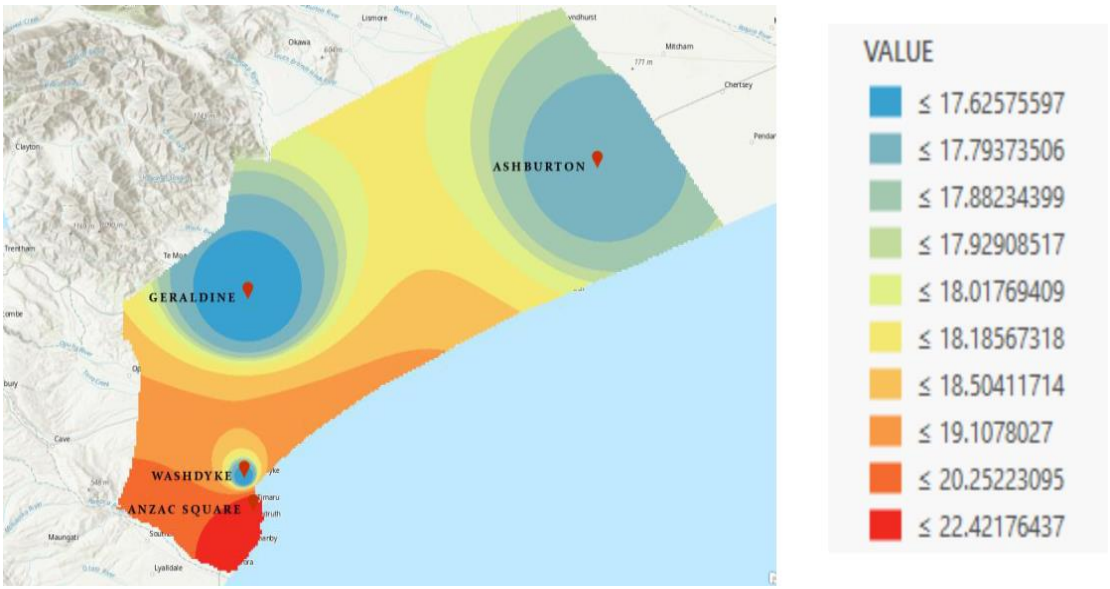

Figure 14:Interpolated surface area with IDW.

Figure 14 above shows the PM10 concentration interpolation by IDW method with power parameter value of 2. Initial experimentation showed that as power was increased the root mean squared error also increased, therefore 2 was considered as the best value. The maximum and minimum neighbour parameter we used for this interpolation is 15 and 10 respectively, along with standard searching.

IDW assumes points that are located close to each other have stronger relationship compared to those that are located far away. In our case, PM10 concentration in Geraldine and Ashburton are far apart compared to PM10 concentration in Washdyke and Anzac Square. Hence the interpolation shows major PM10 concentrations in southern parts of the study area. IDW strongly assumes that two points that are located nearby are similar, the plot above shows a significant difference between these points, which is deceiving.

Table 11: Parameter Tuning for IDW

Kernel Function	**Mean Error**	**Root-Mean-Square Error PM2.5**
Completely Regularized	0.356	2.59
Spline with Tension	0.346	2.571
Multiquadric	0.147	2.515
Inverse Multiquadric	0.393	2.604
Thin Plane Spline	210.48	249.72

5.1.3 Radial Basis Function

RBF has been used to interpolate the mentioned study area with the help of completely regularized spline and a standard search neighbourhood along with all the other default parameters.

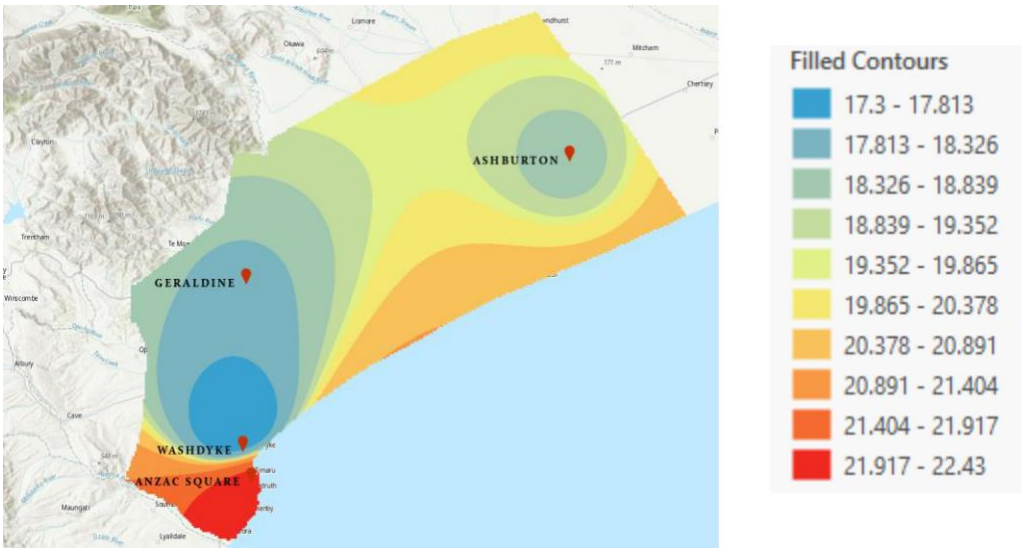

Figure 15: Interpolation surface area by Radial Basis Function.

Figure 15 above captures interpolated PM10 concentration by RBF technique in the interested area. As we can see, PM10 concentration seems to be a bit higher in the southern part of the study area (between 21.917 and 22.43 gm), which means that ambient PM10 in Anzac Square seems to be significantly higher than the other three regions. This is not a satisfying result as the difference of PM10 concentration in Anzac Square and Washdyke is quite distinctive. This is not reasonable as there

should not be this much difference between these two regions, while they are located very close to each other. This shows the highest PM10 concentration is affecting the resultant interpolation. It is known that RBF is an interpolation method that is affected by high variation of the data.

Table 12:Parameter tuning for RBF

Power	Mean Error	Root-Mean-Square Error PM2.5
2	0.715	2.785
3	0.710	2.804
4	0.663	2.786
5	0.60	2.761

Table 13:Comparison analysis between IDW and RBF.

Methods	Mean	RMSE
IDW	0.606	2.716
RBF	0.417	2.515

5.1.4 Ordinary Kriging (OK)

Here we applied OK to interpolate the unknown surface of the study area. We first apply OK with default parameter without removing the trend. Then, we remove the trend and exponential and stable parameters then compare the result.

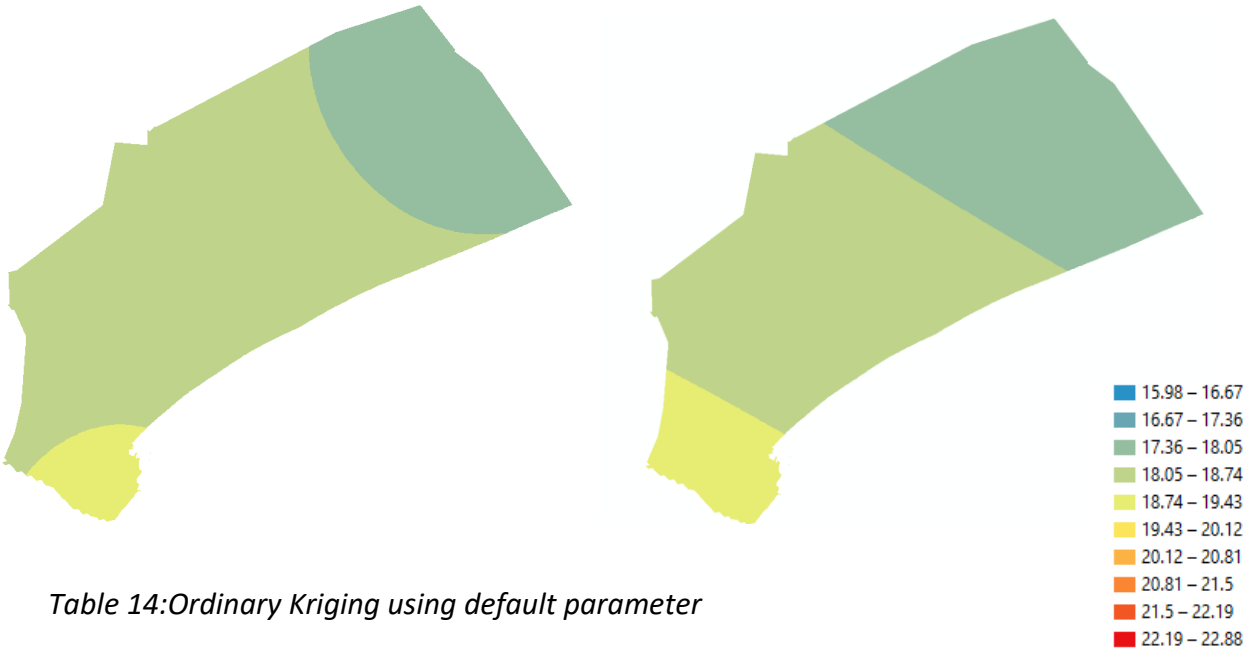

Table 14:Ordinary Kriging using default parameter

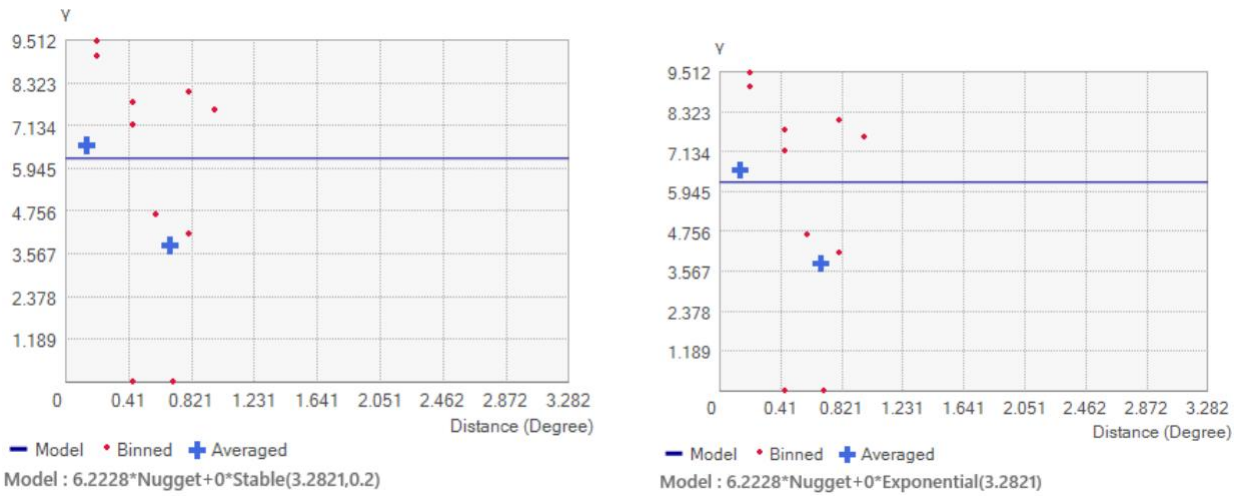

Figure 16: Semivariogram of PM10 concentration, left panel is variogram with stable and right with exponential.

The model does not fit through the points in the semi-variogram meaning that these models are not good fits for this data.

Table 15:statistical analysis of OK with exponential and stable method.

	Mean	**RMSE**	**Mean Standardized**	**Root-Mean-Square Standardized**	**Average Standard Error**
Exponential	0.102	3.508	0.013	0.977	3.606
Stable	0.102	3.508	0.013	0.977	3.606

From the above table it is quite clear that the ordinary kriging values are the same for both exponential and stable methods. This is the result of having a smaller number of data points used for interpolation, where data points correspond to only 4 different target locations. As stated, earlier kriging requires a higher number of data points to operate properly.

Ordinary Kriging after trend removal

Figure 17: Ordinary Kriging Using Exponential Method (left) and Stable method (right)

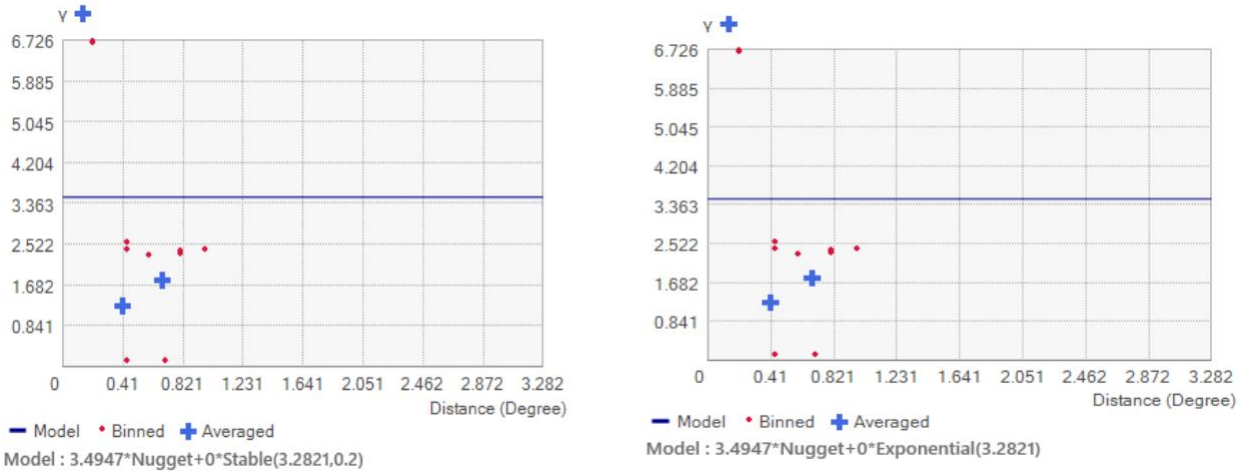

Figure 18:Semivariogram after differencing, left panel with stable and right exponential.

The semi-variograms of models after differencing also show a bad fit as the model line does not go through the points. The points seem even more further away than the non-differenced model.

Table 16:Statistical comparison of two methods of OK.

	Mean	**RMSE**	**Mean Standardized**	**Root-Mean-Square Standardized**	**Average Standard Error**
Exponential	39.85	91.78	18.46	42.51	2.15
Stable	39.85	91.78	18.46	42.51	2.15

Table 3 contains the overall error from the two OK methods. As we can see, both models are providing same error. The second OK method is worse than the first one as it's Root Mean Square Error is higher than the first one.

5.1.5 Comparison of results

Table 17: Statistical comparison of the spatial interpolation methods.

Methods	**Mean**	**Root Mean Squared Error**
IDW	0.283	3.451
RBF	0.668	3.709
OK	0.102	3.508

Table 4 shows the overall error of the three interpolation methods using cross validation. Cross-validation is a sample reuse algorithm for quantitative comparison of experimental performance of alternative interpolation methods. Validation is a statistical method that is used to perform comparisons of the modelled algorithms which are self-learning by dividing the dataset into N-folds. In this case study we have used the 10-fold cross-validation that partitions the data into 10 equal sized partitions or folds. OK appears to give the smallest error for both accuracy metrics, even though it is not producing a good surface in as shown in figure 18.

5.1.6 Significant Findings

From this analysis, key findings emerge:

For IDW only the power parameter was set to 2 as further increase yielded a higher RMSE error. For RBF completely, regularized spline produced best results, evaluated using leave one leave one out cross validation. Ordinary kriging was first fitted using default parameters using exponential and stable methods. However, semi variograms showed that the model was not a good fit to the data.

Exploratory data analysis showed that global trends were present in the data which hinder OK's ability to interpolate efficiently. Therefore, first order differencing was performed to make the data stationary. Ordinary kriging was fitted again on the differenced data using exponential and stable methods. However, the results did not show any improvements. We suppose this is caused by a small number of data points which hinder OK's ability to interpolate properly.

5.2 Spatio-temporal Interpolation

Here we apply different types of spatio-temporal interpolation methods and compare their results against each other. The main results are presented below.

5.2.1 Empirical Semi variogram:

A variogram plot shows the dissimilarity of points in space and time. A lower variogram value indicates a higher correlation between the data points. We consider 3 days' time lag; the spatial bin is 10 km and points are considered to be 20 km apart for both July 2016 data and December 2016. Various combinations of parameters were tested, and this appears to give better overall results than most.

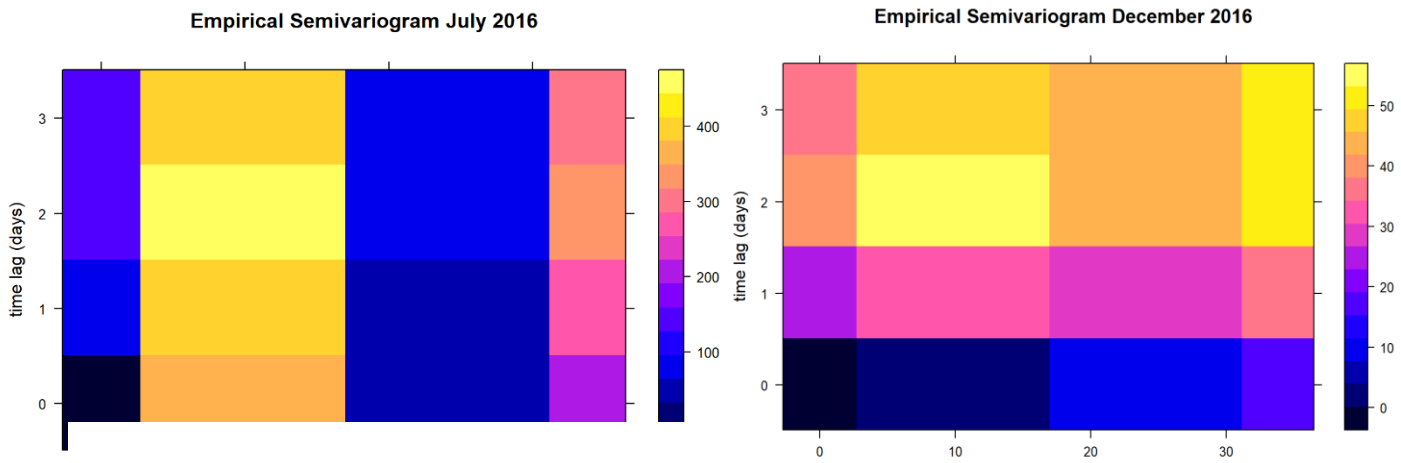

Figure 19: Semi variogram of PM10 concentration in July and December (maximum and minimum PM10 concentration of 2016).

5.2.2 Spatio-temporal Semi variogram

To proceed with spatio temporal kriging we fit separable and metric models to the variograms. To ensure that kriging variances are positive, continuous models are fitted to the empirical semi variograms obtained above. The resulting spatio-temporal semi variograms are shown below.

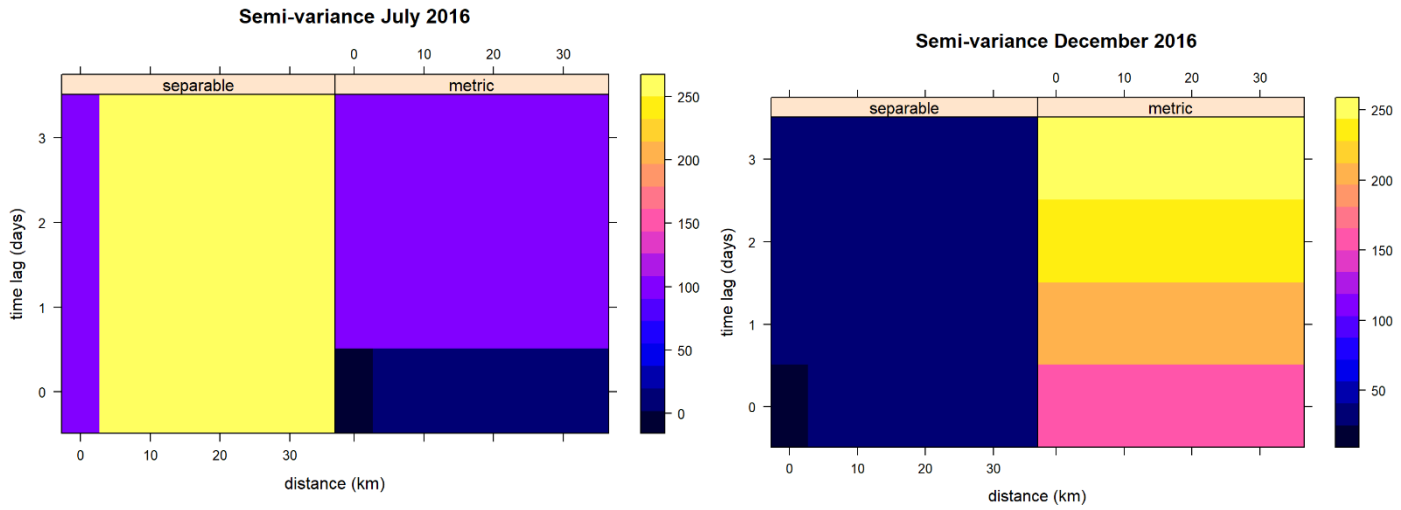

Figure 20: Semi-variance model for PM1 concentration in 2016

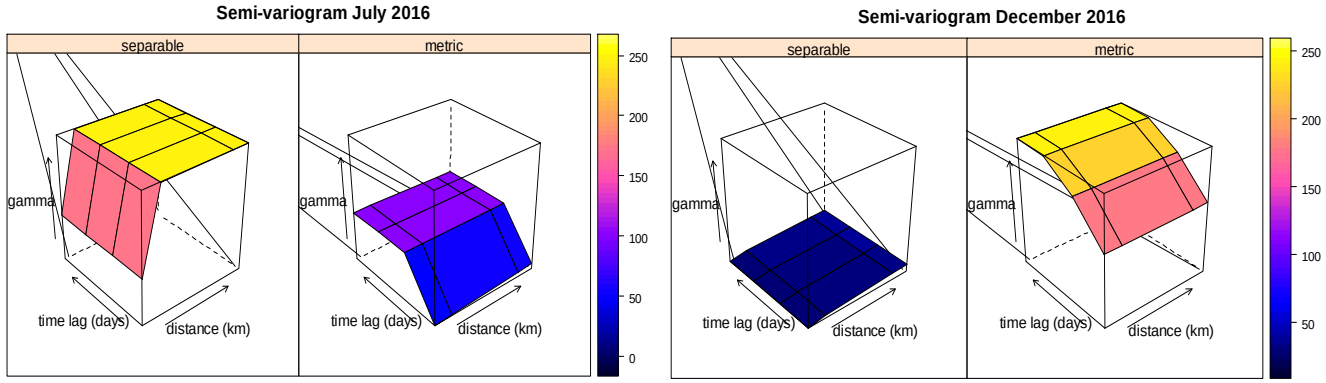

Figure 21: Semi- variogram of graph for different months of the year, left July month and right December 2016

The plots above depict that separable model showed lowest MSE error for subset of December data, emphasizing that spatio-temporal correlation is being captured by the model. On the other hand, metric is providing smaller MSE error compared to separable for the July subset of data. Therefore, for interpolation using Kriging we will consider the metric model for July 2016 data and the separable model for December 2016.

5.2.3 Spatio Temporal Kriging Prediction

Here, we perform spatio temporal interpolation using spatio-temporal prediction grid and the best semi variogram models (metric and separable). The spatial grid is between longitude values of 169 to 171 and latitude values of -43 and -45. The temporal grid was created based on the 3 lag days, for the 12th, 13th and 14th in July and December. Separate times were also considered, 6am, 7:30am and 9am in the morning.

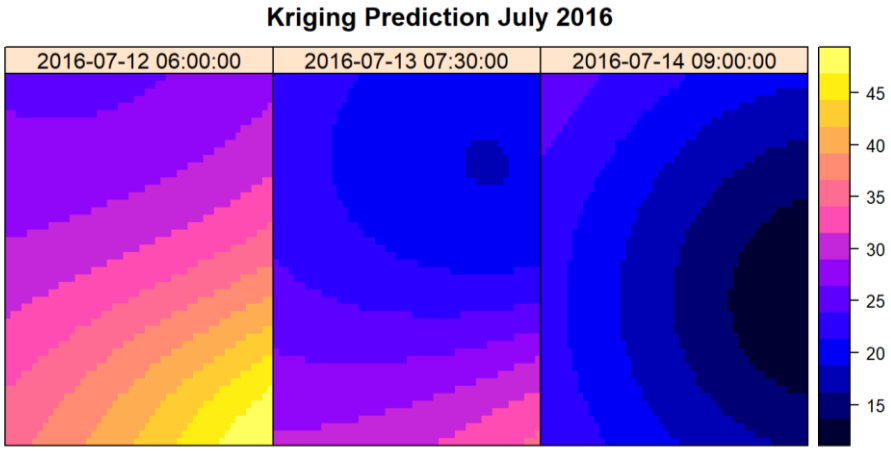

Figure 22: Daily prediction of PM10 using maximum concentration in July 2016.

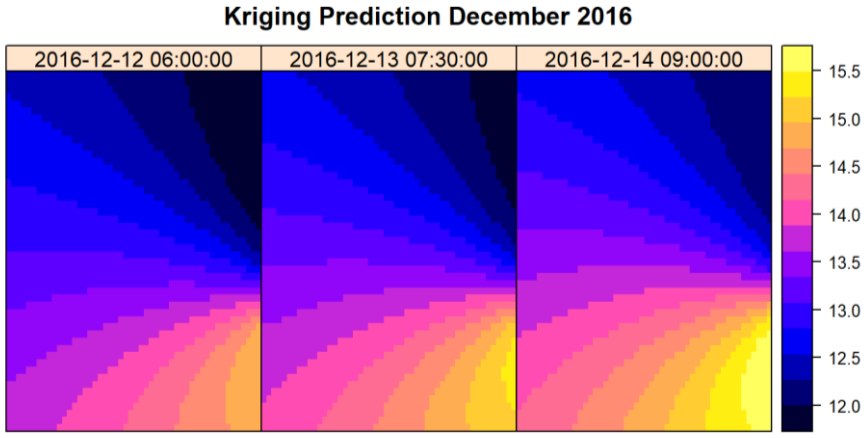

Figure 23: Prediction of PM10 using minimum in December 2016 Prediction.

On the contrary, the plot of Kriging prediction using the month when minimum PM10 concentration of the year is not showing much variation as the date increases. However, the concentration of PM10 is high in South-East of the region.

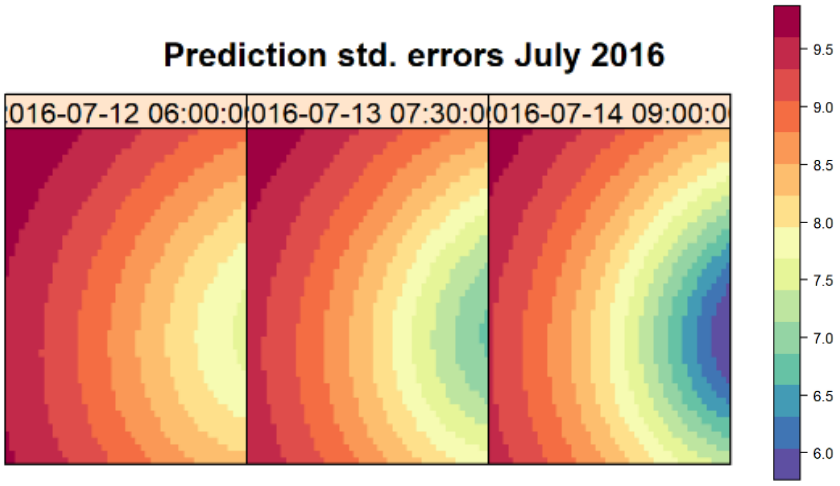

Figure 24: Standard error of kriging model prediction with maximum PM10 concentration in July.

Figure 25: Standard error of kriging model prediction with minimum PM10 concentration in December.

Figure 25 and Figure 26 show the standard error of PM10 concentration prediction in the study area of interest by the kriging model (July and December).

As we can see, the prediction standard error decreases as the time increases, this illustrates the relationship of PM10 in space and time. In general, the standard error is lower for subset of minimum PM10 concentration data (December 2016). The prediction error for July seems to be lower in the seems to predict the PM10 concentration in the eastern part of the region more precisely than the North-West and South-West part of the region

5.2.4 Significant Findings

Evaluation metrics and semi variograms showed that the separable method was performing better for both December 2016 data and metric method performed better for July 2016. Thus, we considered the respective semi variograms from separable method and metric method for Spatio Temporal Kriging. The model quality is assessed through std error plot, the error plots were almost identical across the three prediction days. The std. error was observed to be lower in December 2016 which had the lowest average PM10 concentration.

PM10 concentration varies from day to day. PM10 concentration is significantly high in 7^{th} July then it constantly decreased for the next two consecutive days. The concentration of PM 10 on 7 July 2016 seems to be very high in the East-South of the study area (>40). The plot signifies that PM10 concentration is not consistent across entire day of the year.

For the July 2016 data, there seems to be a strong spatial correlation between points when they are less than around 2 km and around 20 to 30 km apart. Temporal correlation is high for all time lags when the distance is around the 2 km mark or less. We can observe that when distance is between 2 to 19km, the correlation suddenly becomes very weak. With the December 2016 data, the correlation decreases as the time lag increases. Smaller distance between points shows higher correlation. July variogram exhibits abrupt dissimilarity in correlation in space and time, while the December variogram shows more graceful transitions into different levels of correlation. This may be caused by the high

variability in July 2016 data, as the highest PM10 concentration as recorded as 86.27 in Anzac Square and the lowest was 4.571 in Washdyke. The December 2016 data has lower variation with the highest PM10 concentration being 30.6 in Anzac and lowest 3.61 in Geraldine. The nugget effect is present in both cases (July and December).

5.3 Data Mining and Machine Learning

5.3.1 K-means

Anzac Square

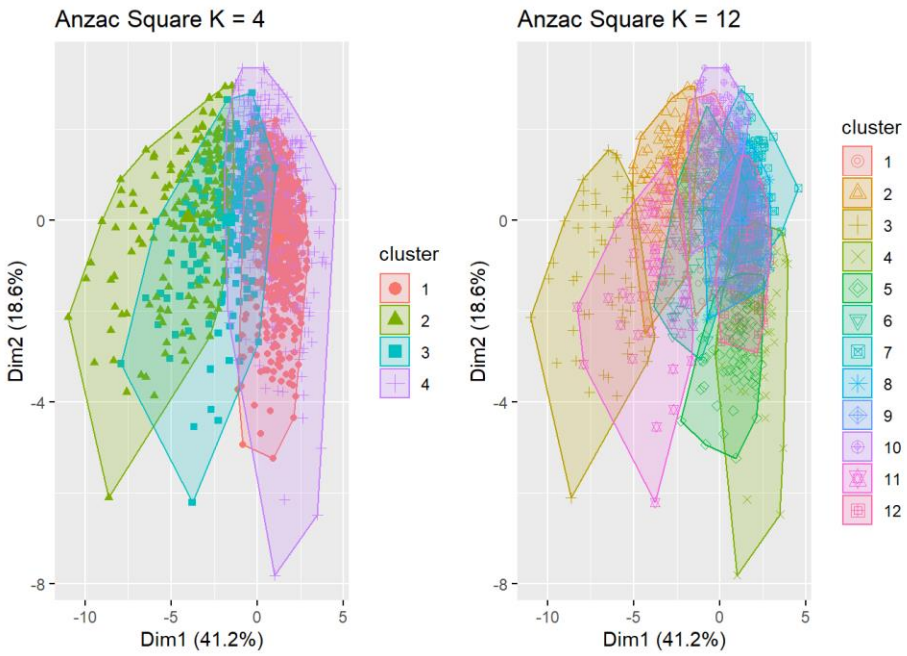

Figure 26: Season and month clusters of Anzac square region, by K-means clustering.

Ashburton

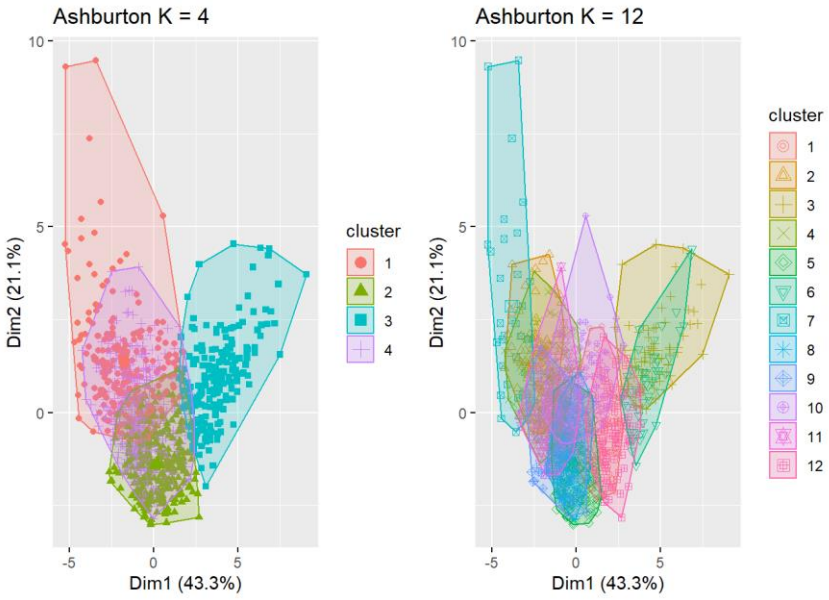

Figure 27: Season and month clusters of Ashburton region, by K-means clustering.

Geraldine

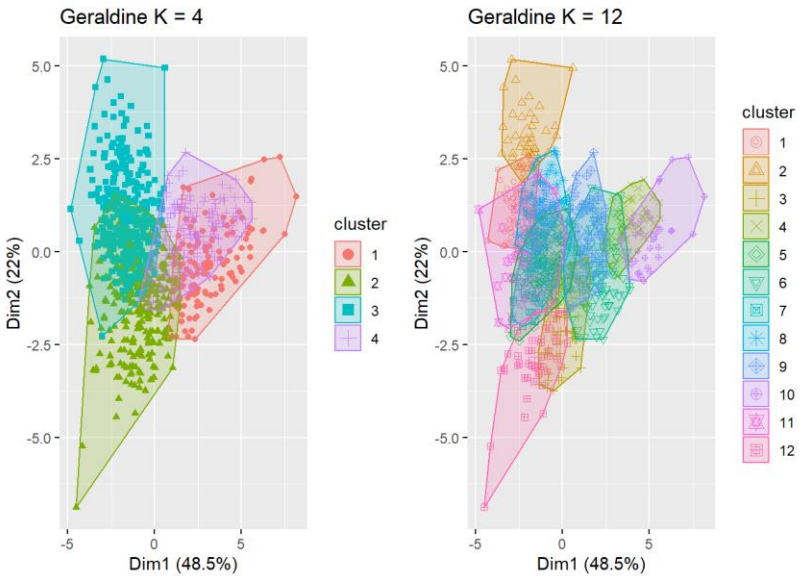

Figure 28: Season and month clusters of Geraldine, by K-means clustering.

Washdyke

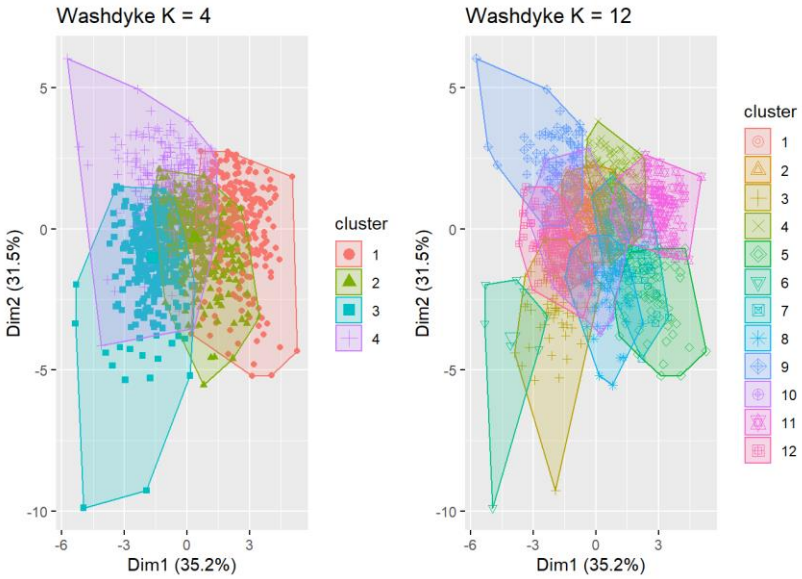

Figure 29: Season and month clusters of Washdyke region, by K-means clustering.

5.3.2 Multilinear Regression (MLR)

Anzac Square

Response: PM10

Predictors (Anzac Square): WINDMAX, TEMP2M, WS, CO, TEMPGROUND, SO2, TEMP6M, PMCOARSE, RELATIVEHUMIDITY, WD, PM2.5

There was strong correlation between PM2.5 and CO (95 correlation coefficient value), temperature at 6 meters and 2 meters, and Windspd (wind speed) and Windmax (maximum wind speed). Highly intercorrelated predictors are considered redundant as they do not provide new information to the model. In some cases, there can be more than 2 variables which are intercorrelated, this can also hinder the model's predictive capability as little or no multicollinearity is key assumption in regression analysis.

Table18: Coefficients and statistical output of predictors of PM10 prediction model for Square region.

	All Variables		After Removing Highly Inter-Correlated Variables
	Pr(>|t|)	**ANOVA Pr(>F)**	**Pr(>|t|)**
WindMax	0.819	**2.20E-16**	0.3138
Temp2m	0.951	**2.20E-16**	*Removed*
ws	0.702	**2.20E-16**	*Removed*
CO	0.288	**2.20E-16**	*Removed*
TempGround	0.927	**2.20E-16**	**0.0379**
SO2	0.144	**2.20E-16**	0.0968
Temp6m	0.95	**2.20E-16**	0.3473
PMCoarse	**<2e-16**	**2.20E-16**	**<2e-16**
RelativeHumidity	0.523	**2.20E-16**	0.9985
wd	0.769	**2.20E-16**	0.9028
PM2.5	**<2e-16**	**2.20E-16**	**<2e-16**

T test column in table 3 shows that when all variables are included in the model only two are significant (p-value < 0.05). However, in this case insignificant variables should not be dropped as t test evaluates the marginal impact of predictors, given the presence of all other predictors. Therefore, the presence of multicollinearity will affect this output. ANOVA uses F-tests for sequential model evaluation, meaning that each predictor is tested with only the intercept and no other variables. The results show that individually, all predictors are significant (P-Value < 0.05) for predicting PM10 concentrations. By removing intercorrelated variables, we observe that 3 variables become significant.

Model Summary:

Table 19: Significance of the model PM10 prediction for Anzac Square Region.

	Residual STD error	**R Squared**	**Adjusted R Squared**	**P-Value**
All Variables	0.08177	1	1	2.2e-16

The R-squared value of 1 indicates that 100% of variance in the data can be explained by the model. However, R-Squared value increases with the number of predictors and it does not account for multicollinearity in the data. Hence, the model can be misleading with large R-squared value and significant p-value. As stated earlier, a high r-squared value does not inherently mean a good fit (Dalson Britto, Jose Alexandre, & Enivaldo, 2011). To get an overall picture of goodness-of-fit, we must

consider R-squared values in combination with residual plots and VIF to validate our assumptions about residuals and multicollinearity.

Multicollinearity between predictors:

Table 20: VIF and Tolerance for Anzac Square (Underlined values violate the optimal conditions for VIF and tolerance)

	All Variables		After Removing Highly Inter-correlated variables	
Variables	**Tolerence**	**VIF**	**Tolerance**	**VIF**
WindMax	6.43E-02	1.56E+01	0.610	1.639
Temp2m	6.63E-07	1.51E+06	*Removed*	*Removed*
ws	6.30E-02	1.59E+01	*Removed*	*Removed*
CO	7.05E-02	1.42E+01	*Removed*	*Removed*
TempGround	2.54E-04	3.94E+03	0.656	1.524
SO2	7.10E-01	1.41E+00	0.720	1.389
Temp6m	6.90E-07	1.45E+06	0.612	1.634
PMCoarse	6.06E-01	1.65E+00	0.629	1.589
RelativeHumidity	6.60E-01	1.52E+00	0.737	1.356
wd	6.27E-01	1.60E+00	0.801	1.248
pm2.5	8.83E-02	1.13E+01	0.400	2.500

Ideally the tolerance should not be less than 0.1 and VIF should exceed 10. However, the table above shows that when all variables are included in the model, only SO2, PMcoarse, relativity humidity and wind direction satisfy these conditions meaning there is multicollinearity in the dataset. Multicollinearity reduces the precision of estimated coefficients which may make the P-values untrustworthy for model evaluation. This explains why only PMcoarse and PM2.5 were considered significant (helpful) in the model. After removing the highly correlated variables, we observe that VIF and Tolerance thresholds are met by all other remaining variables.

Residual Diagnostics

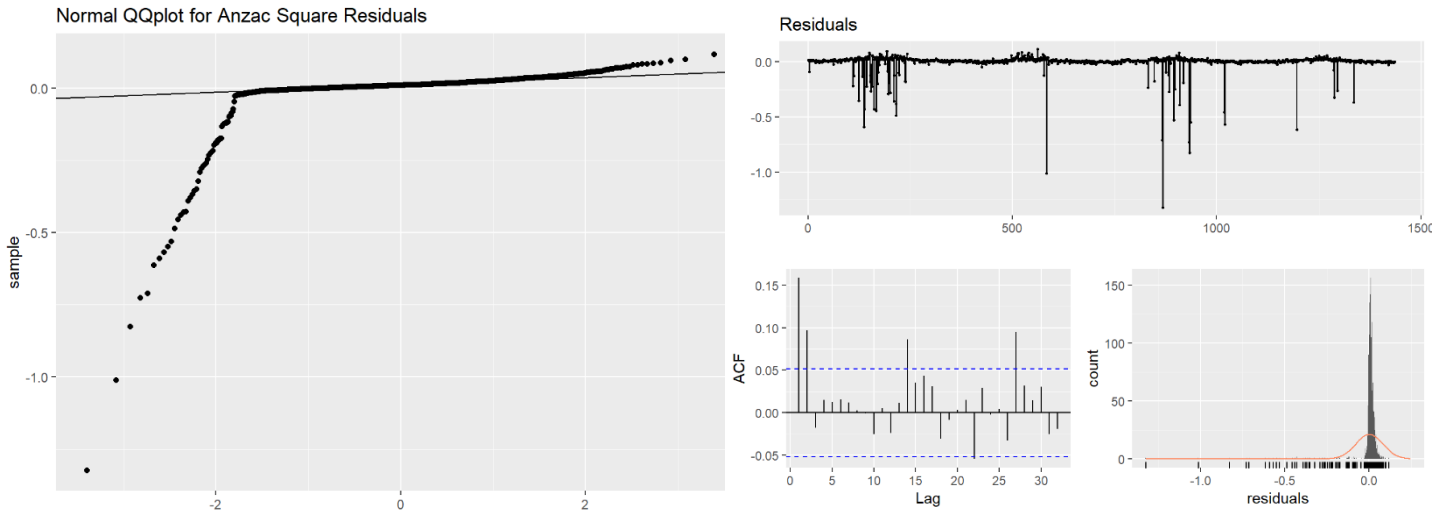

Figure 30: Residual plot of the PM10 prediction model for Anzac Square region.

One of the assumptions for linear regression is that the residuals must be normally distributed. This assumption is not met for this model as the QQ-plot of the residuals should approximately follow a diagonal straight line. Meaning that this model is not good for prediction, this also indicates that the R-squared value was biased. Further examination of the residuals shows that there is auto-correlation present in the residual of the model, which signifies the seasonal variation of PM10 which is not captured well by the model. Linear regression cannot account for seasonality. Therefore, the model is producing high value of R-square and significant p-value due to an overfitting problem (it is biased). The model is not describing the genuine relationship between the variables.

Ashburton

Response: PM10

Predictors (Ashburton): WINDMAX, TEMP2M, WS, CO, TEMPGROUND, TEMP6M, PMCOARSE, RELATIVEHUMIDITY, WD, PM2.5 (SO2 concentrations are not available in Ashburton)

In Ashburton, we observe high correlation between PM2.5 and CO, windspeed and maximum wind, temperature at 6m and temperature 2m. As stated earlier, highly inter-correlated predictors are redundant and are not helpful for prediction.

Coefficients:

Table 21: summary output of MLR of PM10 prediction in the region of Ashburton.

	All Variables	**After Removing Highly Inter-Correlated Variables**	
	Pr(>|t|)	**ANOVA Pr(>F)**	**Pr(>|t|)**
WindMax	0.93921	**2.20E-16**	*Removed*
Temp2m	0.75932	**2.20E-16**	*Removed*
ws	0.89957	**2.20E-16**	0.2409
CO	**0.00814**	**2.20E-16**	*Removed*
TempGround	0.75899	**2.20E-16**	0.0308
Temp6m	0.7592	**2.20E-16**	0.3452
PMCoarse	**2.00E-16**	**2.20E-16**	<2e-16
RelativeHumidity	0.76241	**2.20E-16**	0.4782
wd	0.22883	**2.20E-16**	0.3663
PM2.5	**2.00E-16**	**2.20E-16**	<2e-16

In this model, only CO, PMCOARSE and PM2.5 are significant (zero p-value) given the presence of all other predictors. As was the case in Anzac square this result can be misleading due multicollinearity. Sequential model evaluation using ANOVA shows that individually all variables are significant for predicting PM10 concentrations. This result is very similar to Anzac Square shown in the last section.

Model Summary:

Table 22: Model summary of MLR of PM10 (Response) prediction for Ashburton region.

Residual STD error	**R Squared**	**Adjusted R Squared**	**P-Value**
0.1681	0.9995	0.9995	2.2e-16

The model summary shows that the model can explain 99.9% of the variance in the response variable. This is slightly smaller than the R-squared value of 1 from Anzac Square. This may be caused by the

model having 1 less predictor (SO2) as increasing the number of predictors increases the R-Squared value (Dalson Britto, Jose Alexandre, & Enivaldo, 2011). The model is also giving zero p-value, meaning that the model is significant overall (it can predict PM10). However, further testing is required to evaluate the complete model.

Collinearity between predictors:

Table 23: Variance Inflation Factor of the predictor variables in the region of Ashburton (Underlined values violate the optimal conditions for VIF and tolerance).

	All Variables		After Removing Highly Inter-correlated variables	
Variables	**Tolerance**	**VIF**	**Tolerance**	**VIF**
WindMax	1.41E-02	7.09E+01	*Removed*	*Removed*
Temp2m	1.17E-08	8.57E+07	*Removed*	*Removed*
ws	1.48E-02	6.76E+01	0.670727	1.490919
CO	8.19E-02	1.22E+01	*Removed*	*Removed*
TempGround	1.67E-06	6.00E+05	0.588405	1.699509
Temp6m	1.26E-08	7.95E+07	0.476634	2.098048
PMCoarse	5.58E-01	1.79E+00	0.610536	1.637904
RelativeHumidity	4.65E-01	2.15E+00	0.573571	1.743463
wd	6.30E-01	1.59E+00	0.930274	1.074953
PM2.5	9.84E-02	1.02E+01	0.539026	1.855199

As was the case with the model fitted to Anzac Square data, only PMcoarse, Relative humidity and Wind Direction have VIF less than 10 and tolerance above 0.1. This implies the heavy presence of multicollinearity in the data and the potential inaccuracy of the R-Squared value and P-Values.

Residual Diagnostics

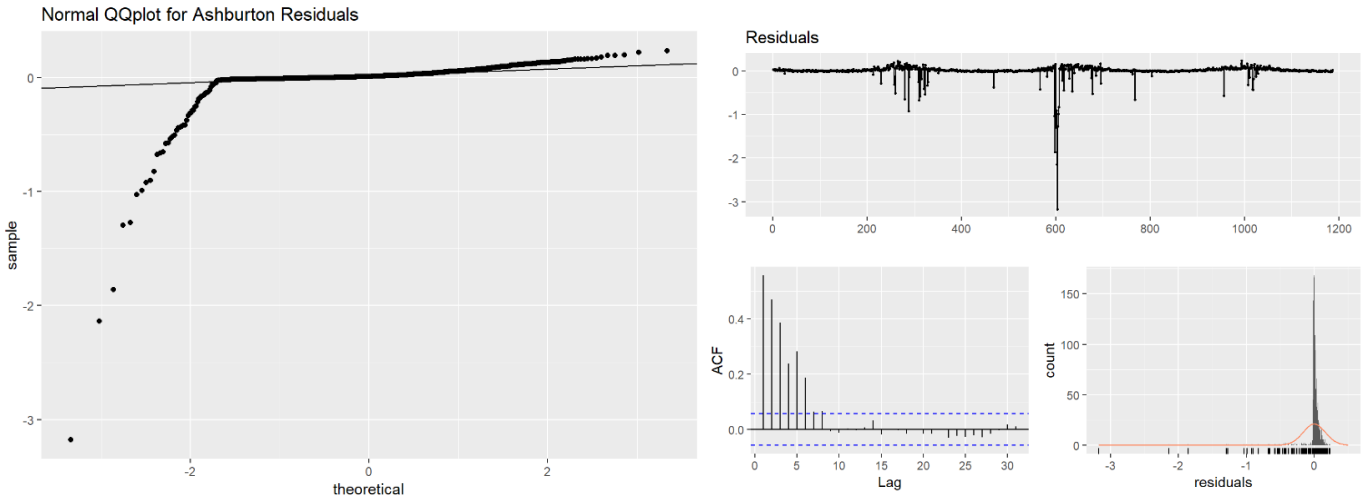

Figure 31: Quantile Plot of MLR for PM10 prediction in the region of Ashburton

We can see, the residuals are not following a straight diagonal line. Meaning that the normality assumption in Ashburton is also violated. This also suggests that the high R-squared value is biased, and this model is not suitable for predictions. The Figure above (figure 31) also illustrates the non-normality of residual. Autocorrelation plot also shows that there is strong autocorrelation between each time lag before 8. This shows that the model is not capturing the seasonal variation of PM10 in

the region of Ashburton. Overall, this suggests that the model is not good for predictions and the high R-Squared value is misleading.

Geraldine

Response:PM10

Predictors (Geraldine): WINDMAX, TEMP2M, PMCOARSE, WS, CO, TEMPGROUND, TEMP6M, WD, PM2.5

Coefficients:

Table 24: Summary output of MLR for PM10 prediction in the region of Geraldine.

	All Variables	**After Removing Highly**	**Inter-Correlated Variables**
	Pr(>|t|)	**ANOVA**	**Pr(>|t|)**
		Pr(>F)	
WindMax	0.4059	**2.00E-16**	*Removed*
Temp2m	0.3777	**2.00E-16**	*Removed*
PMcoarse	**2.00E-16**	**2.00E-16**	**<2e-16**
ws	0.5786	**2.00E-16**	0.9232
CO	0.6994	**2.00E-16**	*Removed*
TempGround	0.6531	**2.00E-16**	0.7581
Temp6m	NA	**0.02139**	**0.0249**
wd	**0.0241**	**2.00E-16**	**0.0319**
PM2.5	**2.00E-16**	**2.00E-16**	**<2e-16**

As was the case with previous locations PM2.5 and PMCoarse are significant in predicting PM10 concentration, additionally, in Geraldine wind direction is also significant. Although, we need to be cautious with these findings due to multicollinearity and autocorrelated data. Notice that TEMP6M is defined as NA, this is because two or more of our independent variables are perfectly collinear. The results from ANOVA show similar findings, individually all predictors in Geraldine are helpful (significant) for predicting PM10 concentrations. After removing the highly correlated variables we observe two additional variables become significant.

Model Summary:

Table 25: Overall model summary of MLR for PM10 prediction in the region of Geraldine.

Residual STD error	**R Squared**	**Adjusted R Squared**	**P-Value**
0.1757	0.9996	0.9996	2.2e-16

Compared to Anzac Square, Geraldine is missing SO2 and Relative humidity. We observe that the R-squared value is slightly smaller which may be caused by a lesser number of predictors. However, such a high R-Squared value is suspicious, suggesting a bias model and meaning that Geraldine data may also contain multicollinearity. As was the case with other locations, CO and PM10, temperature at 2 meters and 6 meters, wind speed and max wind speed are highly correlated. The coefficients for

Temp6m were defined as NA in the model and as we can see Temp6m and Temp2m are perfectly correlated.

Collinearity between predictors:

Table 26: VIF of MLR in the region of Geraldine (Underlined values violate the optimal conditions for VIF and tolerance).

	All Variables		**After Removing Highly Inter-Correlated Variables**	
Variables	TOLERANCE	VIF	TOLERANCE	VIF
WindMax	1.24E-01	8.07E+00	Removed	Removed
Temp2m	4.44E-16	2.25E+15	Removed	Removed
PMcoarse	4.63E-01	2.16E+00	0.492547	2.030263
ws	1.31E-01	7.65E+00	0.769784	1.299065
CO	5.94E-02	1.68E+01	Removed	Removed
TempGround	3.66E-14	2.73E+13	0.607573	1.645893
Temp6m	5.55E-16	1.80E+15	0.436483	2.291038
wd	7.94E-01	1.26E+00	0.88342	1.131965
PM2.5	7.82E-02	1.28E+01	0.502017	1.991965

Originally, PMcoarse, WindMax, Windspeed and Wind Direction have VIF less than 10 and tolerance above 0.1. This implies the heavy presence of multicollinearity in the data and the potential inaccuracy of the R-Squared value and P-Values. After removing the highly intercorrelated variables, VIF and Tolerance conditions are met for all variables.

Residual Diagnostics

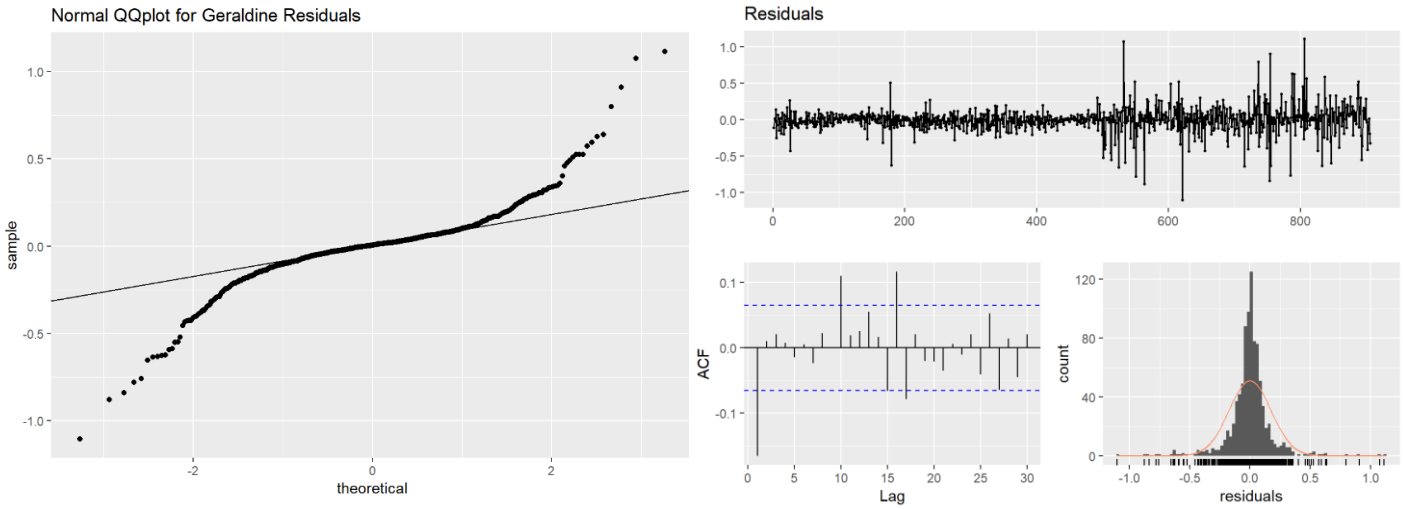

Figure 32: Quantile plot of residual from MLR in the region of Geraldine.

The Normal QQ-plot for residuals shows closer resemblance of residuals to normal distribution. However, the points still don't follow a straight line, the residuals are approximately normal. There is also an improvement in the other residual plots as there is less autocorrelation in the data. The histogram also shows closer resemblance to normal distribution. Although, there is some autocorrelation which seems to be caused by seasonal effects, it is comparatively less than Anzac Square and Ashburton this may be due to large number of missing values in the dataset.

Washdyke

Respone:PM10

Predictors (Washdyke): WINDMAX, TEMP2M, PMCOARSE, WS, TEMPGROUND, TEMP6M, WD, PM2.5

Coefficients:

Table 27: Summary output of MLR for the prediction of PM10 in the region of Washdyke.

	All Variables		**After Removing Highly Inter-Correlated Variables**
	Pr(>|t|)	ANOVA Pr(>F)	Pr(>|t|)
WindMax	0.09531	2.20E-16	Removed
Temp2m	0.01702	2.20E-16	Removed
PMcoarse	2.00E-16	2.20E-16	< 2e-16
ws	0.08943	2.72E-17	0.002885
TempGround	0.01692	2.20E-16	0.522996
Temp6m	0.01702	0.2333	0.779107
wd	0.00203	2.20E-16	0.005299
PM2.5	2.00E-16	2.20E-16	< 2e-16

In Washdyke CO, SO2 and relative humidity are missing, for previous locations we saw that these variables showed some inter-correlations. In the summary output above we can see that apart from wind max and windspeed all variables are significant. This is contrary to other locations where only 2 or three variables were significant. This is not entirely unexpected as there are lesser number of predictors which contribute to multicollinearity. In previous locations we observed that wind max and windspeed were highly correlated, in the output above they are both non-significant (not helpful in model according to t test). This is because t test asses each variable with respect to all other variables and as both are highly correlated, they are deemed insignificant in the model. The sequential ANOVA test shows that individually windspeed and wind max are helpful in predicting PM10 concentrations. Overall this is similar to all other locations, however temperature at 6m is not significant.

Model Summary:

Table 28: Model summary for prediction of PM10 in the region of Washdyke.

Residual STD error	**R Squared**	**Adjusted R Squared**	**P-Value**
0.4311	0.9978	0.9978	2.2e-16

The model summary shows that the model can explain the 99.7% of the variance in the response variable. However, as with previous locations further analysis is required to verify this. Since there less variables in Washdyke, we observe that there are less number of correlated variables. As was the case with previous locations, wind speed and wind max, Temp6m and Temp2m are highly correlated.

Collinearity between predictors:

Table 29: VIF of predictors in the region of Washdyke (Underlined values violate the optimal conditions for VIF and tolerance).

	All Variables		**After Removing Highly Inter-Correlated Variables**	
Variables	Tolerance	VIF	Tolerance	VIF
WindMax	1.47E-02	6.78E+01	*Removed*	*Removed*
Temp2m	7.94E-09	1.26E+08	*Removed*	*Removed*
PMcoarse	5.41E-01	1.85E+00	0.566084	1.766522
ws	1.51E-02	6.62E+01	0.73003	1.369806

TempGround	1.67E-06	6.01E+05	0.682898	1.464348
Temp6m	8.50E-09	1.18E+08	0.660945	1.512985
wd	7.46E-01	1.34E+00	0.738983	1.353212
PM2.5	5.65E-01	*1.77E+00*	0.574613	1.740303

Although residual analysis showed better results than Anzac Square and Washdyke, we can see that Washdyke data is also subject to heavy multicollinearity. This means that our p-values used in hypothesis testing (t-test, f-test) may not be valid and R-squared value is misleading.

Residual Diagnostics

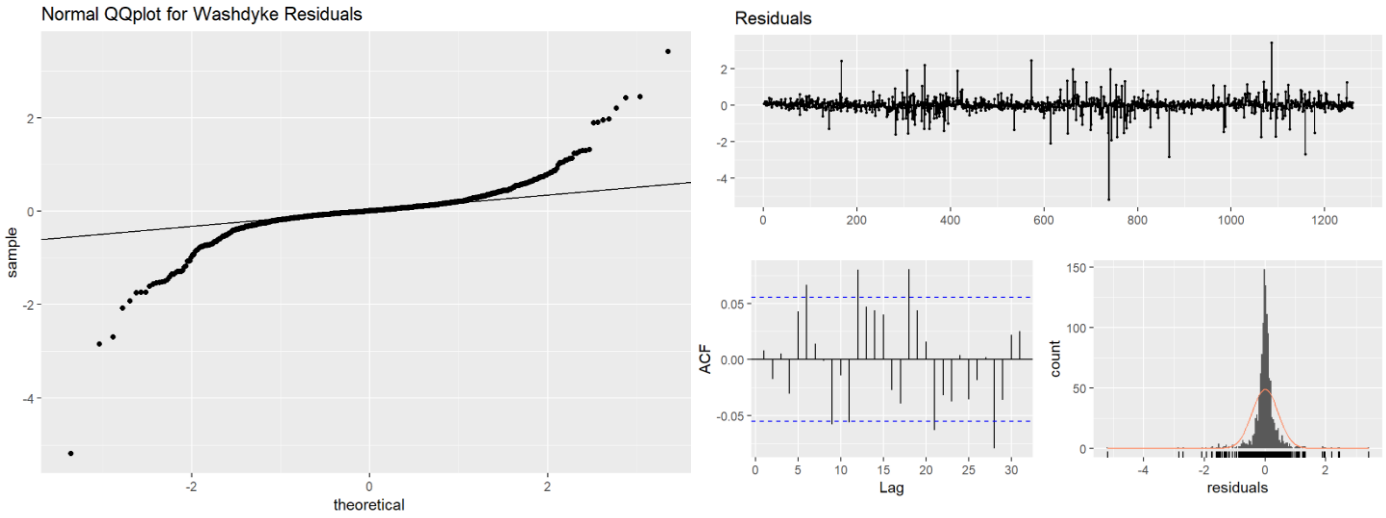

Figure 33: Quantile plot of PM10 prediction in the region of Washdyke.

The Normal QQ-plot for residuals shows closer resemblance of residuals to normal distribution. However, the points still don't follow a straight line. There seems to be autocorrelations present in the data due to the seasonal effects. Similar to other locations, data collected from Washdyke also violates to residual assumptions for the model. Although residual diagnostics seem better, it is also essential to check for multicollinearity to assess the validity of our prediction.

5.3.3 Principal Component Analysis (PCA)

PCA transforms highly correlated variables into the principal component, inherently this dimension reduction technique, reduces the multicollinearity in the data without losing crucial information.

Anzac Square

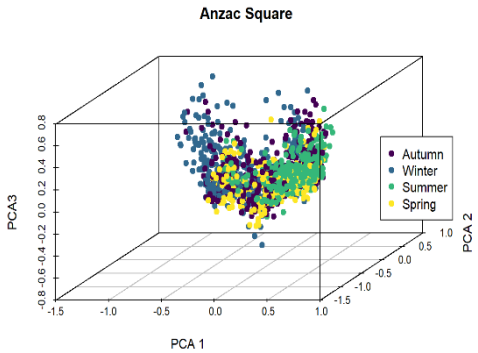

Figure 34: 3-D PCA plot in the region of Anzac Square region.

The first three principal components for Anzac Square data account for 78% variation. The plot above shows each observation in the Anzac Square data set plotted into the 3-dimensional subspace constructed by PCA. We can clearly see some separation between the different seasons. Winter and summer are well separated while spring and autumn being transition seasons are more spread.

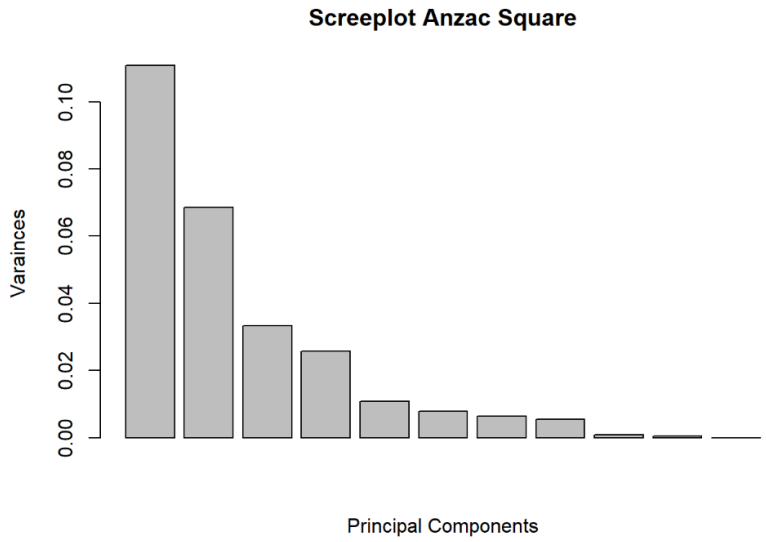

Figure 35: Screeplot of PCA for the region of Anzac Square region

The above graph is for Anzac Square region that represents the scree plot that tells us the number of Principal Components required to retain the adequate representation of the original variables. It is quite clear that to retain 85% of the explained variance in the data we require the first four PC's.

Loadings for Anzac Square:

Table 30: Loadings of each predictor variables on the PCA.

	PC1	**PC2**	**PC3**	**PC4**
WindMax	0.095009	-0.28858	-0.1899	-0.34534
Temp2m	0.316976	-0.35632	0.333084	0.317787
Ws	0.125626	-0.26891	-0.16203	-0.39358
CO	-0.29599	0.285316	0.344369	-0.06715
TempGround	0.204211	-0.07375	-0.18682	0.117947
SO2	-0.0336	0.03068	0.354322	-0.05727
Temp6m	0.314646	-0.36564	0.361459	0.32211
PMCoarse	0.028979	0.026636	0.463495	-0.13959
RelativeHumidity	0.05616	0.339251	-0.30367	0.638409
Wd	-0.76973	-0.55666	-0.0607	0.266668
PM2.5	-0.2209	0.265143	0.326079	-0.03505

The above table represents the loading values for each variable with respect to the first four PC's that explains the effect of each variable on the respective Principal Components.PC1 is affected majorly by wind-direction, PC2 is affected majorly by wind direction , PC3 is affected majorly by PMCoarse and PC4 is affected majorly by Relative Humidity.

Ashburton

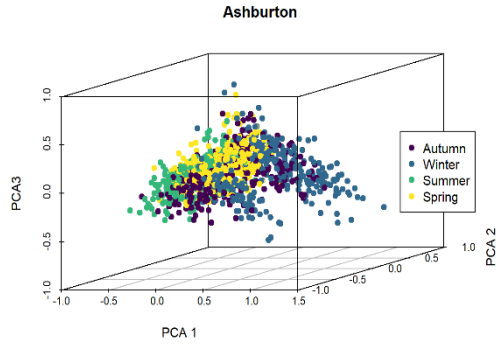

Figure 36:Seasoanl clusters for Ashburton Regions.

The first three principal components for Ashburton data account for 82% variation. Here, we can also see separation between the different seasons. Clusters formed by summer and winter are separate, but autumn and spring are more spread out

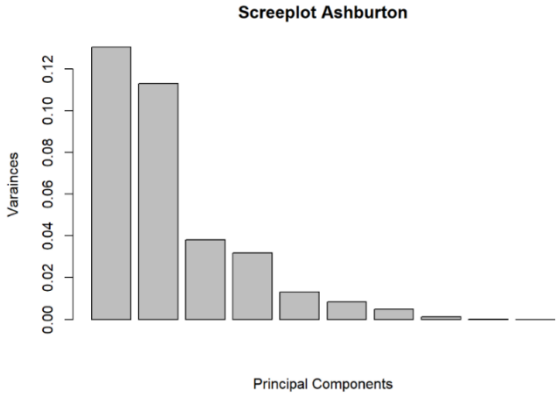

Figure 37: Screeplot of PCA for Ashburton

The above graph represents the scree plot that tells us the number of Principal Components required to retain the adequate representation of the original variables. It is quite clear that to retain 85% of the explained variance in the data we require the first four PC's.

Loadings for Ashburton

Table 31: Loadings of each variables on PCA.

	PC1	**PC2**	**PC3**	**PC4**
WindMax	-0.25297	0.14924	0.312502	-0.45543
Temp2m	-0.4683	0.082768	-0.32376	0.242566
ws	-0.25503	0.119722	0.316691	-0.47282
CO	0.380004	0.016646	-0.32034	-0.18679
TempGround	-0.27352	-0.12069	0.301476	0.281488
Temp6m	-0.47046	0.10195	-0.37982	0.228139
PMCoarse	-0.09602	0.039073	-0.20508	-0.08743
RelativeHumidity	0.21618	-0.3194	0.41466	0.479342
wd	0.222724	0.908804	0.163235	0.295905
PM2.5	0.324886	-0.04291	-0.34469	-0.13797

The above table represents the loading values for each variable with respect to the first four PC's that explains the effect of each variable on the respective Principal Components.PC1 is affected majorly by Temp6m, PC2 is affected majorly by wind-direction, PC3 is affected majorly by RelativeHumidity and PC4 is affected majorly by WindMax and windspeed.

Geraldine

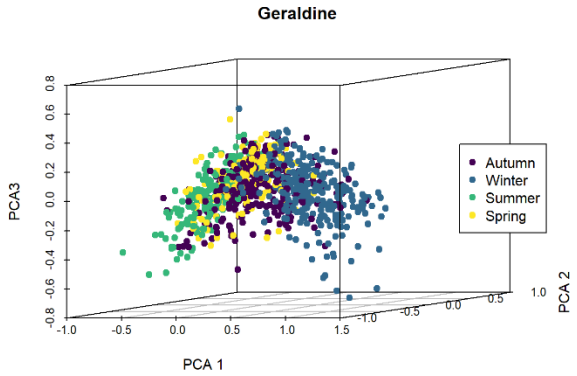

Figure 38: Seasonal clusters for in the regions of Geraldine.

The first three principal components for Geraldine data account for 88% variation. This plot also shows clear separation between the different seasons in Geraldine.

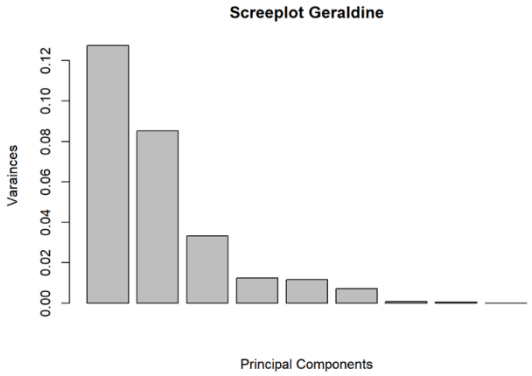

Figure 39:Screeplot of PCA for Geraldine region.

The above graph if for Geraldine region that represents the scree plot that tells us the number of Principal Components required to retain the adequate representation of the original variables. As for Ashburton we can retain 85% of the explained variance in the data from the first three PC's.

Loadings for Geraldine

Table 32: Loadings of each variables on PCA for the region of Geraldine.

	PC1	**PC2**	**PC3**
WindMax	-0.10593	-0.15438	0.231302
Temp2m	-0.49738	-0.20878	-0.24235
PMcoarse	-0.27442	0.014151	-0.59377
ws	-0.082	-0.14597	0.212406
CO	0.337124	0.134733	-0.35061
TempGround	-0.32672	0.004048	0.181193
Temp6m	-0.46281	-0.21527	-0.27613
wd	0.357597	-0.90502	-0.10392
PM2.5	0.311288	0.165832	-0.49728

The above table represents the loading values for each variable with respect to the first four PC's that explains the effect of each variable on the respective Principal Components.PC1 is affected majorly by temp2m, PC2 is affected majorly by wind-direction, PC3 is affected majorly by PMCoarse.

Washdyke

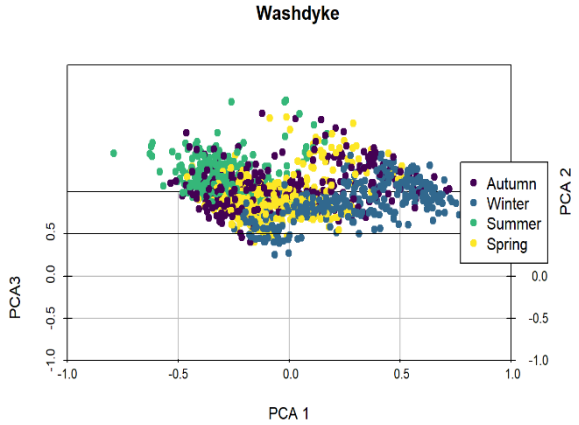

Figure 24: Seasonal cluster for the region of Washdke.

The first three principal components for Washdyke data account for 87% variation. By plotting the data into the three-dimensional subspace using the first three PCAs we can see clear separation between the four differed seasons.

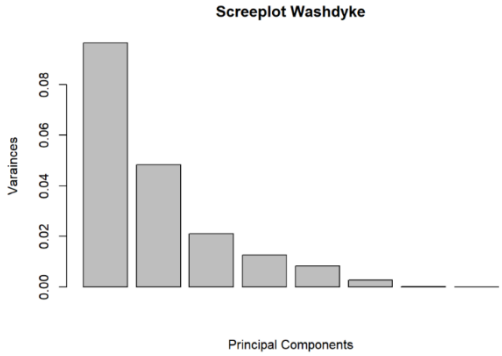

Figure 40: Screeplot of PCA for the region of Washdyke.

As for the other target regions the above scree plot that tells us the number of Principal Components required to retain the adequate representation of the original variables. It is quite clear that to retain 85% of the explained variance in the data we require the first three PC's.

Loadings for Washdyke.

Table 33: Loadings of each variables for the region of Washdyke.

	PC1	**PC2**	**PC3**
WindMax	-0.07506	0.262612	-0.50991
Temp2m	-0.40826	0.486154	0.220489
PMcoarse	-0.06394	0.050769	0.305248
ws	-0.06916	0.233768	-0.49804
TempGround	-0.29276	0.082501	-0.10365
Temp6m	-0.3983	0.497434	0.237874
wd	0.757404	0.610923	0.114228
PM2.5	0.02831	-0.09995	0.519404

The above table represents the loading values for each variable with respect to the first four PC's that explains the effect of each variable on the respective Principal Components.PC1 is affected majorly by wind-direction, PC2 is affected majorly by wind-direction , PC3 is affected majorly by PM2.5.

5.3.4 MLR- PCA

Anzac Square

Table 34: PCA-MLR model summary for the region of Anzac Square.

Residual STD error	R Squared	Adjusted R Squared	P-Value
0.04939	0.8584	0.858	2.2e-16

Coefficient Table

Table 35: MLR-PCA using the first four principal components.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.191492	0.001303	146.93	<2e-16	***
PC1	-0.15942	0.003916	-40.71	<2e-16	***
PC2	0.223251	0.004981	44.83	<2e-16	***
PC3	0.497702	0.007141	69.69	<2e-16	***
PC4	-0.09977	0.008128	-12.27	<2e-16	***

ANOVA table

Table 36: ANOVA of MLR-PCA with the first four principal components.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
PC1	1	4.0423	4.0423	1657.16	2.20E-16	***
PC2	1	4.9012	4.9012	2009.26	2.20E-16	***
PC3	1	11.8485	11.8485	4857.32	2.20E-16	***
PC4	1	0.3675	0.3675	150.65	2.20E-16	***
Residuals	1431	3.4906	0.0024			

Linear regression on PCA for Anzac Square provides notably different results than without PCA. We observe that the R Squared value is reduced to 0.8584 and all principal components (PC1 – PC4) are deemed significant by the t-test. As PCA combines closely correlated variables, multicollinearity is purged from the data. Therefore, all PC's are significant with respect to all other PC's. Sequential analysis using ANOVA also shows that individually, all PC's are helpful in predicting PM10 concentrations.

Residual Diagnostics:

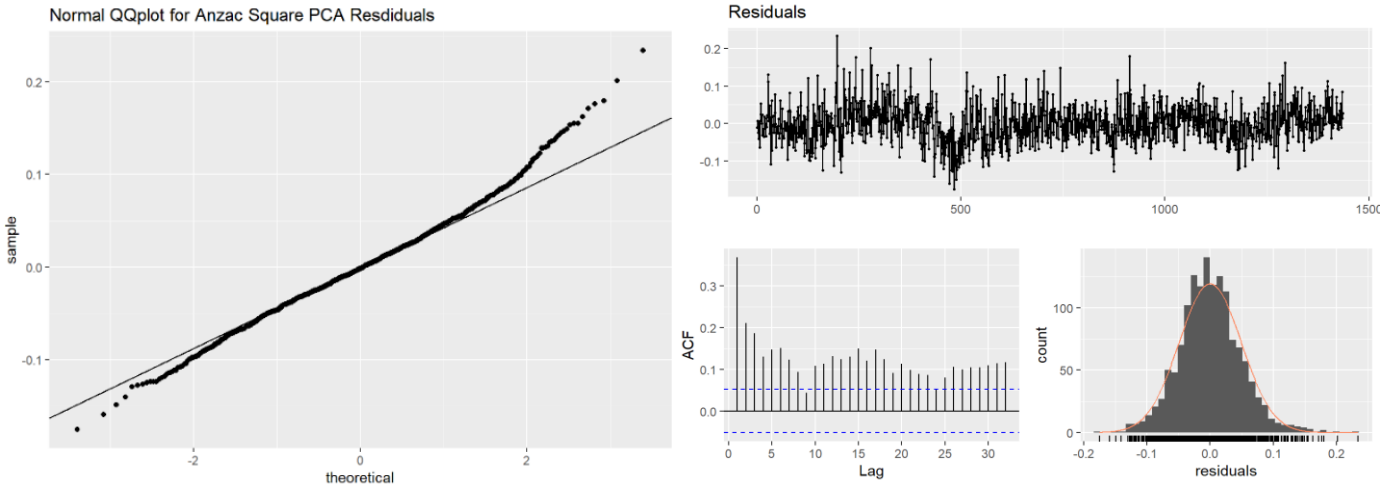

Figure 41: Residual plot of MLR-PCA for Anzac Square region.

Normal QQ-plot after PCA approximately resembles normal distribution. This is much better compared to residual plots from Non-PCA linear regression. *VIF and tolerance were recorded as 1 for all PC's indicating no presence of multicollinearity in Anzac Square data after Principal Component analysis.*The residuals from Anzac Square shows signs for the presence of autocorrelation, we can observe periodic trophs and peaks which represent seasonal effects of the time series. The histogram shows that the data is approximately normal. Although, residual assumptions are still not met for Anzac Square after PCA, we can see that eliminating multicollinearity has improved the residual plots. However, to reduce autocorrelation differencing is required. Overall, we conclude the R-square and the p-values to be more accurate than regression with PCA. All factors including pollutant concentrations and metrological conditions are contributing for prediction of PM10 concentrations.

Ashburton

Table 37: PCA-MLR model summary for the region of Ashburton region.

Residual STD error	R Squared	Adjusted R Squared	P-Value
0.09102	0.5804	0.579	2.2e-16

Coefficient Table

Table 38: MLR-PCA using the first four principal components for Ashburton region.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.234613	0.002642	88.81	<2e-16	***
PC1	0.147383	0.007321	20.132	<2e-16	***
PC2	0.004416	0.007862	0.562	0.574	
PC3	-0.44437	0.013537	-32.827	<2e-16	***
PC4	-0.18245	0.014804	-12.325	<2e-16	***

ANOVA Table

Table 39: ANOVA of MLR-PCA with the first four principal components.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
PC1	1	3.3574	3.3574	405.2997	<2e-16	***
PC2	1	0.0026	0.0026	0.3156	0.5744	
PC3	1	8.9267	8.9267	1077.602	<2e-16	***
PC4	1	1.2583	1.2583	151.8988	<2e-16	***
Residuals	1182	9.7915	0.0083			

Linear regression on PCA for Ashburton gives a low R-Squared value than Anzac Square. PC2 is deemed insignificant in the model by t-test and f-test using ANOVA. Wind direction in Ashburton has the highest effect on PC2 (0.90), Wind direction in Ashburton had very small correlation with PM10 concentration.

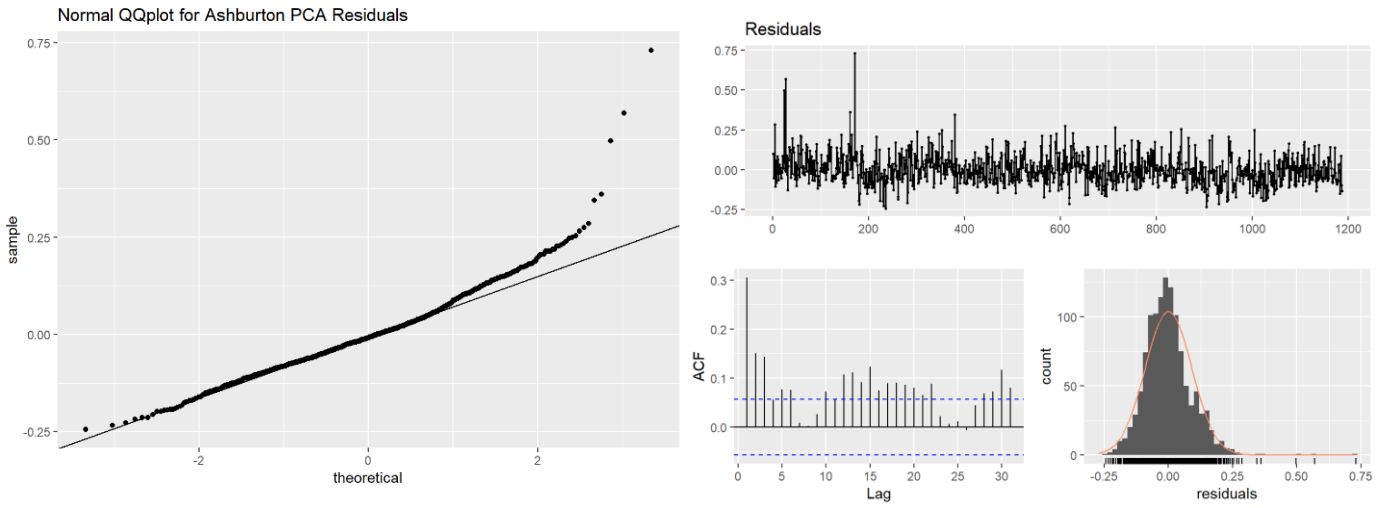

Figure 42: QQ-plot for Ashburton Residuals after PCA

As was the case with Anzac Square and Ashburton, Normal QQ-plot for Ashburton after PCA loosely resembles normal distribution. However, the distribution seems to be skewed to the right. *VIF and tolerance were recorded as 1 for all PC's indicating no presence of multicollinearity in Ashburton data after Principal Component analysis.* The residual plots show signs of autocorrelation, due to seasonal effects, meaning that residuals are not random. The histogram confirms the analysis from QQ-plots as the data seems approximately normal and skewed to the right.

Geraldine

Table 40: PCA-MLR model summary for the region of Geraldine region.

Residual STD error	R Squared	Adjusted R Squared	P-Value
0.04782	0.9073	0.907	2.2e-16

Coefficient Table

Table 41: MLR-PCA using the first four principal components for Geraldine region.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.267998	0.001589	168.7	<2e-16	***
PC1	0.144168	0.004452	32.38	<2e-16	***
PC2	0.154961	0.005442	28.47	<2e-16	***
PC3	-0.72702	0.00871	-83.47	<2e-16	***

ANOVA Table

Table 42: ANOVA of MLR-PCA with the first four principal components for Geraldine region.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
PC1	1	2.3976	2.3976	1048.65	2.20E-16	***
PC2	1	1.8537	1.8537	810.76	2.20E-16	***
PC3	1	15.931	15.931	6967.75	2.20E-16	***
Residuals	902	2.0623	0.0023			

Linear Regression on PCA gives a high R-squared value (0.9073) for Geraldine compared to Anzac Square and Ashburton which means 90% of the variance is explained by this model. From the coefficients and ANOVA table we can conclude that all the Principal Components are significant in the model as the p-value is less than 5% level of significance.

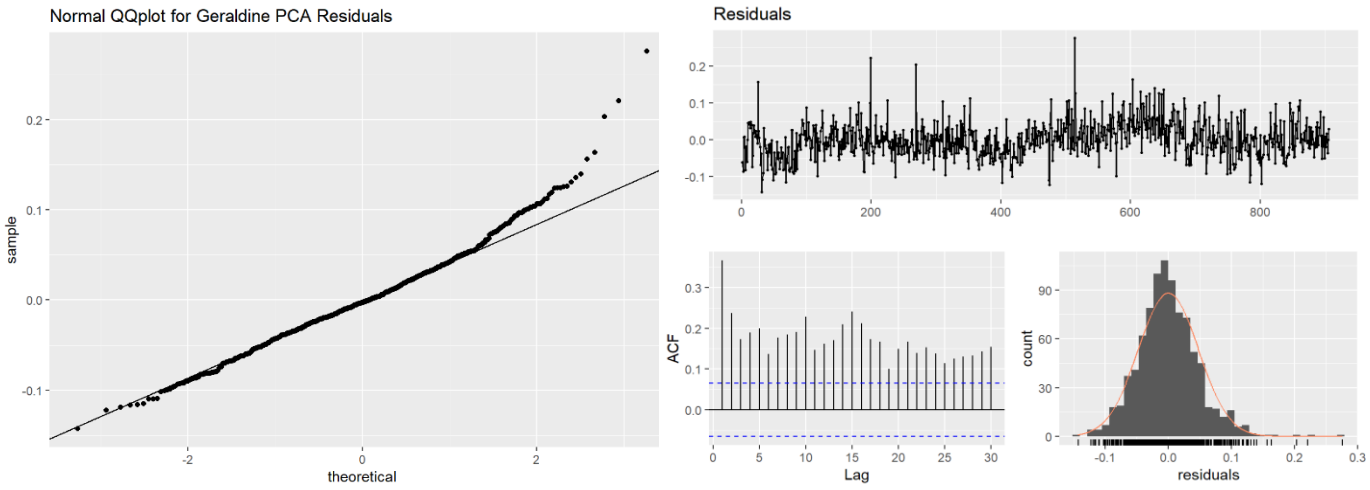

Figure 43: QQ-plot for Geraldine residuals after PCA

The Normal QQ-plot for Geraldine also resembles normal distribution as the points approximately follow a straight. *VIF and tolerance were recorded as 1 for all PC's indicating no presence of multicollinearity in Geraldine data after Principal Component analysis.* The ACF plots shows presence of autocorrelation due to seasonal effects and the histogram resembles Normal distribution. Although multicollinearity is eliminated, the high R-square value does not mean that the model is good as residuals are still not random due to autocorrelation.

Washdyke

Table 43: PCA-MLR model summary for the region of Washdyke region.

Residual STD error	R Squared	Adjusted R Squared	P-Value
0.04939	0.5053	0.5041	2.2e-16

Coefficients Table

Table 44: MLR-PCA using the first four principal components for Washdyke region.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	0.167994	0.001918	87.574	<2e-16	***
PC1	-0.05147	0.006175	-8.335	<2e-16	***
PC2	0.012364	0.008723	1.417	0.157	
PC3	0.460004	0.013218	34.802	<2e-16	***

ANOVA Table

Table 45: ANOVA of MLR-PCA with the first four principal components for Washdyke region.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
PC1	1	0.3221	0.3221	69.4662	<2e-16	***
PC2	1	0.0093	0.0093	2.0089	0.1566	
PC3	1	5.6158	5.6158	1211.189	<2e-16	***
Residuals	1256	5.8236	0.0046			

Linear Regression on PCA gives a low R-squared value (0.5053) as compared to other regions which means 50.53% of the variance is explained by this model. From the coefficients and ANOVA table we can conclude that PC2 is not a significant predictor in the model as the p-value is more than the 5% level of significance.

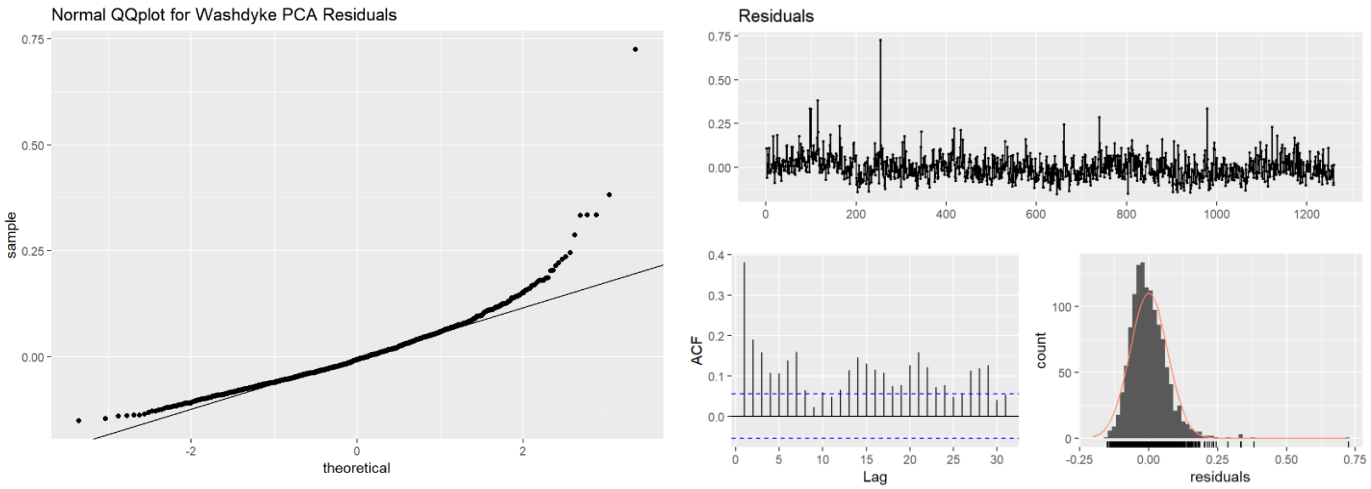

Figure 44: Q-plot for Washdyke residuals after PCA

As was the case with all other locations, Normal QQ-plot for Washdyke after PCA loosely resembles normal distribution. *VIF and tolerance were recorded as 1 for all PC's indicating no presence of multicollinearity in Washdyke data after Principal Component analysis.* However, the residual plots show that residuals are not random and are subject to autocorrelation. The Normal distribution assumption is met; however, the residuals are not random. Indicating that the model is still not appropriate for accurate predictions.

5.3.5 Significant Findings

Original predictors (pollutant concentrations and meteorological factors) were subject to heavy multicollinearity, which violates essential assumptions for linear regression. PCA was used to reduce the number of dimensions and remove multicollinearity in order to improve the input variables for MLR. The VIF and tolerance values for predictors after PCA were 1 for all locations, indicating the elimination of multicollinearity. We observed improvement of MLR models with PCA by comparing the results against the MLR with original predictor variables. On this basis, we notice that PCA improved the prediction quality of MLR for each region. Additionally, we also observe that wind direction is an important factor for predicting PM10 concentrations as it had highest loadings in some PC's for all locations. The model prediction quality was assessed through multifaceted statistical methods, such as quantile plots, p-values and R-squared of the overall model. However, for all models built on PCA-MLR, the residual plots showed that they are not satisfying the residual randomness assumption because PM10 concentrations adhere to seasonal effects and MLR is not a preferred model to capture seasonal variation present in the dataset. The model quality can be enhanced further by withdrawing the seasonal variation in the data.

6 Discussion

The time series analysis with different time resolution details the variation of PM10 concentration across the four regions. The high-resolution plot shows that Anzac Square has the highest PM10 concentration, followed by Washdyke. However, the variation of PM10 concentration in Ashburton is showing some unusual spikes occurring in 2016 (figure 3). Our result from the monthly resolution (figure 4) demonstrates that PM10 concentration is high during the winter and low during the summertime, with an exception of Washdyke. We also notice that the concentration of PM10 is following an upward trend, magnifying an increasing level of PM10 concentration in the regarding study area over the past four years. Another promising finding was that the concentration of PM10 hits its peak at night across all regions, apart from the region of Washdyke (figure 5). Lower resolution data (average hour of each day) shows that there is high air ambient level during the night-time in Ashburton, Geraldine and Anzac, while it is antithetical for the region of Washdyke. On the other hand, the concentration of PM10 is high during the weekday across all sites (>40 μg), in contrary the concentration of PM10 in Washdyke reaches its peak during the weekday, afternoon (>40μg). The data is sufficiently large and pre-processed cautiously; therefore, we believe this analysis is quite accurate.

The trend analysis show that there is strong downward trend as we go from North to South (figure 12), indicating the lowest average PM10 concentration in Geraldine (North) and highest in Anzac and Washdyke (South). Based on the RMSE comparison analysis (table 3), IDW produces the best interpolated surface as it has smallest prediction error. From the graph (figure 12), as we travel from north to south, PM10 concentration level increases, which is expected to see (>22.42 μg). Removing the trend from the data does not seem to improve OK's interpolation surface area. However, it was unexpected that it is yielding relatively smaller RMSE while in fact it is not producing good interpolated surface. This is highly likely to be due to the small sample size as differencing the data

reduces the sample size even further to 3. Although the model is not good for visualizing the sample data interpolated surface, it is robust in terms of connecting points, which is why it is still producing relatively smaller RMSE (3.508). Furthermore, the spatio-temporal interpolation result show that there is high variation in maximum data, whereas there is low variation in minimum. We extrapolate that this is due to the high consumption of home heating burning fuels during the wintertime thus the maximum PM10 concentration in July is 45 μg, while the minimum is 15 μg.

The result from K-means clustering shows that there is strong seasonal overlap between the clusters, emphasizing the transition period from one season or one month to another. This is acceptable as the concentration of PM10 does not change significantly as the season or moth changes. Also, the linear regression result reveals that there is strong collinearity among the predictors, hindering the model from producing precise prediction of PM10. Even though all the variables are significant, there is still strong seasonal variation which is not well explained by the model. This is simply because multilinear regression model is not suitable for seasonal data as it assumes a straight-line relationship between the predictor and response variable. Therefore, the model is still producing a significant p-value, but still violating assumption of normally distributed residual. Surprisingly, PCA is also not improving the result ascertaining the seasonal variation of PM10 which is not captured by the model.

7 Conclusion

Taking everything into consideration, our result authenticates that most of the cities we looked at in this paper follow a very distinctive seasonal pattern, not including Washdyke. We assume this is occurring because Washdyke is an industrial site, as a result, the smoke emission of the factories tends to impact the concentration of PM10 severely. Hence, the concentration of PM10 tends to follow an arbitrary variation throughout the year. Correspondingly, PM10 concentration is significantly high during the night-time, during both weekend as well as weekday for Anzac, Geraldine and Ashburton. On contrary, the concentration the subject matter in the region of Washdyke is high during the daytime, which might be the time when most of the factories are operated.

Despite the sample size limitation, this analysis is valuable in light of interpolating the unknown locations. IDW is providing a close resemblance of the expected result, whereas OK produces a poor interpolated surface, but lower prediction error. In future work, it is important to observer a few more data points so as to obtain a better interpolated surface area and compare OK against other spatial interpolation methods. We expect more data points might also improve the spatio-temporal interpolation.

K-means seems to provide a best seasonal as well as monthly clusters. This can even be further enhanced through another unsupervised machine learning algorithm, such as self-organized map (SOM). The model is suitable for the purpose of clustering and visualizing such high dimensional data. MLR does not seem to be a preferable model in terms of providing for prediction of PM10 concentration in the regarding study area. This is because MLR simply assumes a linear relationship between the predictors and response variables. PCA is certainly improving the model in terms of capturing the linear relationship between the relevant variables (significant p-value and close to normally distributed residual), but the autocorrelation plot shed a light on the seasonal variation of PM10 which is not well explained by the model. This is an exciting desired future work, where another complex models such as Spiking Neural Network (SNN), Singular Spectrum Analysis (SSA) or other models that incorporate seasonal aspect of the data into their model factor can be implemented.

While we apply some of the dynamic models to analyse the data, this paper faces a potential limitation. As we saw in the data exploration section, there are a number of missing values existed in the data set. Simple imputing algorithms failed to predict the missing values. Therefore, we remove all features with >50% of missing values. This might cause loss of prominent information. Therefore, some advanced seasonal models such as, SSA, ARIMA or ETS can be employed to more accurately predict the missing values.

8 Bibliography

Aj White, J. K. (2019). Air Pollution, Clustering of Particulate Matter Components and Breast Cancer. *Cancer Epidemiology Biomarkers & Prevention*.

Allard, D., Bel, L., Gabriel, E., Opitz, T., & Parent, E. (2017). An introduction to geostatistical analysis.

Amina Nazif, N. M. (2017). Regression and multivariate models for predicting particulate matter concentration level. *Environmental Science and Pollution Research*, 283-289.

Dalson Britto, F. F., Jose Alexandre, S., & Enivaldo, R. (2011). What is R2 all about? *Leviathan – Cadernos de Pesquisa Política*, 60-68.

EPA. (2018). *Particulate Matter (PM) Pollution*. Retrieved from Https://Www.epa.gov/Pm-Pollution/Particulate-Matter-Pm-Basics#PM

Fabiana Franceschi, M. C. (2018). Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá, Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering. *Atmospheric Pollution Research*, 912-922.

Fanchi, J. R. (2010). Semivariogram. *Integrated Reservoir Asset Management*.

Hemalatha, K., Wooi-Nee, T., Sin, L., & Mohammad, F. A. (2016). Interpolation of low resolution Digital Elevation Models: A comparison. *2016 8th Computer Science and Electronic Engineering (CEEC)*, 71-76.

Macharia, P. M., Giorgi, E., Noor, A. M., Waqo, E., Kiptui, R., & Snow, E. A. (2018). Spatio-temporal analysis of Plasmodium falciparum prevalence to understand the past and chart the future of malaria control in Kenya. *Malaria Journal*.

Matthew Kyan, P. M. (2014). Unsupervised Learning.

Matthew, Q., Jane, L., & Guangquan, L. (2017). Time-varying relationships between land use and crime: A spatio-temporal analysis of small-area seasonal property crime trends.

Pengwei, Q., Peizhong, L., Yanjun, C., Wenxia, W., Sucai, Y., Mei, L., & Tongbin, C. (2019). Comparison of common spatial interpolation methods for analyzing pollutant spatial distributions at contaminated sites. *Environ Geochem Health*.

Robert, E. K., Bacciu, V., & Kathy, G. (2012). Spatial variability of wildland fuel characteristics in northern Rocky Mountain ecosystems. *US Department of Agriculture, Forest Service, Rocky Mountain Research Station*.

StatsNZ. (2018). *PM10 Concentration. New Zealand Goverment.*

Tanja Trošić1, F. A. (2017). Multiple Linear Regression (MLR) model simulation of hourly PM10 concentrations during sea breeze events in the Split area. *Nase More.*, 77-85.

WHO. (2016). *Ambient Air Pollution. A Global Assessment of Exposure and Burden of Disease.*

Wie Sun, J. S. (2017). Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. *INTERNATIONAL JOURNAL OF ENVIRONMENTAL SCIENCES* , 144-152.

Yan, T., Yan, Y., Xingbang, H., & Li, H. (2016). Performance analysis of different kriging interpolation methods based on air quality index in Wuhan. *Sixth International Conference on Intelligent Control and Information Processing*.

Appendix A

I believe this project is a small scale of real-world scenario, where I have gained several knowledge and experience in a variety of areas. In this project, several things went well, and some didn't' as expected. For instance, one of the main things that I did well was removing the features with huge missing values from the data. SO2 and CO had several missing values with huge values, which is one of the challenging parts of the project and I didn't know what to deal with it. I tried to impute them using some imputing algorithms. However, my methods failed to produce accurate prediction of missing values as they were not seasonal models, and we know that concentration of CO and SO2 is seasonally variated features. Therefore, I removed both features from the region of Washdyke as >50% of the observation is missing. Therefore, being decisive and able to remove these features was advantageous in terms of improving the interpretability as well as saving time. On the other hand, one thing that I wish I did better is that if I used a seasonal forecasting models such as ARIMA, ETS or SSA to impute the missing values it would be beneficial to understand the relationship between CO and PM10 CO is a highly correlated feature with PM10. CO is directly related to combustion of carbon-containing fuels, such as gasoline, natural gas, oil, coal, and wood. It would be interesting to see the relationship between PM10 and CO in the region of Washdyke.

Appendix B

Process Log:

This section lists the key steps we took to perform analysis for each section of this project.

8.1 Build A Study Area Map In ArcMap

1. Start ArcMap and add shape file **studyarea.shp and spatial_data.csv**
2. Add and XYZ theme using **x field=Longitude, y field = Latitude** and **Z field=PM10.**
3. Set the layout view on and add a data frame
4. In the new data frame, add a **World National Geographic World Basemap**
5. Add another **Basemap** for the shapefile made
6. Add another **Basemap** on the new dataframe just created
7. Magnify the New Zealand map using Zoom In magnifying glass
8. Right click on the new data frame and open **Extent Indicators** tab.
9. Click on the **Layers** data frame from **Other Data Frames** section and insert it into **Show extent Indicator for these data frames**
10. Then you see the **Layers** data frame on the **Show extent Indicator for these data frames, c**lick on **Frame**
11. Select red Color and 2.0 Point Border Line as well as 100 Round value
12. Then you will see red circles rounding the study are of this project on New Zealand map.
13. Then save the layout file by clicking **File**, **Export File** and save it as a **Study_Area_map.PNG**
14. Once you saved the image, open it in word and customize it, Labelling the data points and connection line between the New Zealand map and the map with shape file in it.

8.2 Data Cleaning:

8.2.1 Missing Values

1. First, we obtain our data from: http://data.ecan.govt.nz/Catalogue/Method?MethodId=94. The time period is within 30/12/2014 - 01/01/2019 and the Selected sites are Ashburton, Geraldine, Anzac Square and Washdyke.
2. Import the 4 datasets into R and rename the columns into simpler names for easier referencing in the code.
3. Omit all rows with missing values in Anzac Square and Ashburton using the "na.omit" function.
4. For Geraldine first remove the relative humidity column as 14166 values were missing and then omit all rows with missing values using "na.omit" function.
5. For Washdyke remove the CO and SO2 columns and then omit all rows with missing values.

8.2.2 Missing Values

1. Apart from the temperature variables Temp2m, Temp6m and TempGround. Use the "dframe[dframe < 0] <- NA" command to set all negative values as missing in other columns.
2. Use "na.omit" function to omit all rows with negative values.

8.2.3 Preparing Datasets

1. Make the date as a POSIXct object using the "as.POSIXct" command.
2. Create new CSV files for the 4 cleaned datasets.

8.3 Data Exploration:

1. Open new R script, import the "openair", "dplyr" packages.
2. Using the "timePlot" function create full resolution time series plots for pm10 concentrations using the 4 cleaned datasets. Using the same command create monthly resolution plots by setting avg.time = "month"
3. Using the "timeVariation" command create diurnal variation plots for all locations and each season. Set hemisphere = "southern" to match plots with NZ season times.
4. Create seasonal polar plots with "polarPlot" command set hemisphere = "southern".
5. Create heatmaps using "trendlevel" function, with type = "weekend".

8.4 Spatial Interpolation:

8.4.1 Create dataset and study map

1. Create new dataset with 4 rows and 4 columns as follows: Column 1 should contain the names of all selected sites, column 2 and 3 should have the longitude latitude coordinates for each location. Column 4 should have the overall mean PM10 concentrations for all locations. Save this file as a CSV
2. To create the study map, Visit http://www.diva-gis.org/gdata, for country select: New Zealand and for subject select Administrative areas. Download the Shape File.
3. Open Arcgis pro, create a blank project and click new map.
4. Click add data and import the downloaded shape file.
5. Click on the edit tab and use the reshape tool to morph the map into the desired study area.
6. Save the edited shape file.

8.4.2 Interpolation

1. Open arcmap and click file, add data, add xy data.

2. Import the csv file and select x as longitude and y as latitude and z as the PM10 concentrations. Use the GCS_WGS_1984 coordinate system and click ok.
3. Import the modified shape file representing the study area.
4. Click customize toolbars and geostatistical analyst geostatistical wizard.
5. Click inverse distance weighting and select x as longitude and y as latitude and z as mean PM10 concentrations, set power = 5 and finish.
6. Enter geostatistical wizard again, click radial basis function and select the appropriate xyz fields. Select the multiquadric kernel function and click okay.
7. Using the geostatistical wizard select kriging/cokriging, kriging type as ordinary and semivariogram model as stable click finish.
8. Repeat the same steps as 7 but select exponential and semivariogram model.
9. Right click each interpolation layer go to properties, extent and set the extent to "the rectangular extent of the study area". Click apply.
10. Right click layers and goto properties, dataframe and set clip options to "clip to shape".
11. Click specify shape and set outline of features as study area. Click apply.

8.5 Spatio Temporal Interpolation:

8.5.1 Preparing Dataset

1. Open a new R script Import the "dplyr", "sp", "tidyr", "spacetime", "SpatioTemporal", "openair" and "gstat" packages.
2. Import the cleaned datasets for all locations.
3. Average the datasets by day using the timeAverage function, set avg.time = "day".
4. Separate the date of each location into year, month and day columns using the separate function from dplyr.
5. Split the dataset for each location into two subsets of July 2016 and December 2016 using the filter function.
6. Using cbind, create new dataframe called data using the split datasets with columns representing each location and rows representing PM10 observations for those particular locations.
7. Create another dataframe called locations with columns representing station name, longitude, latitude for each of the four sites.
8. Set coordinates to the WGS84 format in the locations dataframe using the proj4string function,
9. Create a spatiotemporal object using the "data" and "locations" data frame for each split (July, December)

8.5.2 Spatiotemporal Kriging

1. Feed the STFDF object to the variogram function to create an empirical semi variogram. set tlags from 0 days to 3 days.
2. Using the vgmST function fit separable and metric models to the empirical semi variograms.
3. Asses the fit of models by comparing predicted errors using attr function.
4. Create a spatial grid ranging from longitude 169 – 171 and -43, -45.
5. Create temporal grid using as.POSIXct spanning 3 days.
6. Use the KrigeST function to perform and kriging prediction and plot using stplot.

8.6 Data Mining:

8.6.1 K-Means

1. Open new r script import the “ggplot2”, “openair”, “olsrr” and “factoextra” packages.
2. Import clean dataset at daily resolution and scale all variables between 0 and 1.
3. For each location perform k means clustering with the “kmeans” function by setting k=4 and and k=12.
4. Plot clusters using the “fviz_cluster” function.

8.6.2 Linear Regression

1. Import clean dataset at daily resolution.
2. Fit linear regression model using the lm function with PM10 as response and all other variables as predictors.
3. Obtain model summary and perform ANOVA test using “summary()” and “anova()” functions.
4. Find the highly correlated variables using the findCorrelation function with cutoff value of 0.7 and remove the selected variables from the dataset.
5. Fit linear regression model again with PM10 as response and all other remaining variables as predictors.
6. Check residuals using the checkresiduals function and check for multicollinearity using the ols_vif_tol function.

8.6.3 Principal Component Analysis

1. Using the scaled dataset, calculate principal components using the prcomp function.
2. Make the screeplot by squaring standard deviations outputted by the prcomp function.
3. Select top variables which explain up to 85% variation in the data
4. Perform regression using PM10 as the response and selected PC’s as the predictors.
5. Perform ANOVA and multicollinearity test using ANOVA and ols_vif_tol function respectively.