ESTIMATION OF NEAR GROUND PARTICULATE MATTER IN URBAN AREAS

A THESIS SUBMITTED TO AUCKLAND UNIVERSITY OF TECHNOLOGY IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Supervisors

Assoc. Prof. Jacqueline Walley

Prof. Philip Sallis

Prof. Stephen MacDonell

April 2021

By

Sara Zandi

School of Engineering, Computer and Mathematical Sciences

Table of Contents

List of Abbreviations & Acronyms

Adaptive Moment Estimation (ADAM)

Air Quality Guidelines (AQGs)

Air Quality Index (AQI)

Artificial Neural Networks (ANNs)

Auckland City Council (AC)

Augmented Dickey-Fuller (ADF)

Auto Correlation Function (ACF)

Auto Regression (AR)

Autoregressive Integrated Moving Average (ARIMA)

Carbon Dioxide (CO_2)

Carbon Monoxide (CO)

Continuous Ranked Probability Score (CRPS)

Cumulative Rank Difference (CRD)

Degrees of Freedom (DF)

Effective Degrees of Freedom (EFD)

First-Order Auto-Regressive (AR (1))

Generalized Additive Mixed Models (GAMMs)

Generalized Additive Models (GAMs)

Generalized Cross Validation (GCV)

Generalized Linear Mixed Model (GLMM)

Generalized Linear Model (GLM)

Generalized Linear Models (GLMs)

Index of Agreement (IA)

Inverse Distance Weighting (IDW)

Kruskal–Wallis Test (K-W)

Kwiatkowski–Phillips–Schmidt–Shin (KPSS)

Leave-One-Out Cross-Validation (LOOCV)

Locally Estimated Scatterplot Smoothing (LOESS)

Locally Weighted Smoothers (LOWESS)

Mann–Kendall (MK)

Mean Absolute Percentage Error (MAPE)

Mean Squared Prediction Error (MSPE)

Methane (CH_4)

Ministry for the Environment (MfE)

Moran's Index (IM)

Moving Average (MA)

Multi-layer Perceptron (MLP)

National Emissions Inventory (NEI)

National Environmental Standards for Air Quality (NESAQ)

Nitrogen Oxides (NOX)

Nitrous Dioxide (N_2O)

Non-constant Variance Score (NCV)

Nonmethane Hydrocarbons (NMHC)

Normalized Difference Vegetation Index (NDVI)

Number of Missing Data (NA)

Optical Particle Counter (OPC)

Ordinary Kriging (OK)

Ordinary Least Squares (OLS)

Partial Autocorrelation Function (PACF)

Particle Number Count (PNC)

Particle-Bound Polycyclic Aromatic Hydrocarbons (PB-PAH)

Particulate Matter (PM)

Planetary Boundary Layer (PBL)

Positive Matrix Factorization (PMF)

Predictive Cross-Validation (PCV)

Principal Component Analysis (PCA)

Recurrent ANN (RC ANN)

Recurrent Neural Network (RNN)

Relative Humidity (RH)

Root mean squared error (RMSE)

Seasonal and Trend Decomposition using Loess (STL)

Seasonal ARIMA (SARIMA)

Self-Organizing Map (SOM)

Simple Exponential Smoothing (SES)

Simple Kriging (SK)

Space-Time Index (STI)

Spatio-Temporal (S-T)

Spearman’s Rho (SMR)

Sulfur Dioxide (SO_2)

Temperature (Temp)

The National Environmental Standards for Ambient Air Quality (AQNES)

Tobler’s First Law (TFL)

Total Organic Carbon (TOC)

Total Suspended Particulate (TSP)

Universal Kriging (UK)

Wind Direction (WD)

Wind Speed (WS)

World Health Organization (WHO)

List of Figures

List of Tables

Attestation of Authorship

"I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person (except where explicitly defined in the acknowledgements), nor material which to a substantial extent has been submitted for the award of any other degree or diploma of a university or other institution of higher learning."

Sara Zandi

Dedication

I dedicate this research to late Professor Philip Sallis without his able leadership, this thesis would not have been possible, and I shall eternally be grateful to him for his guidance.

Acknowledgment

I am deeply indebted to my late supervisor Professor Philip Sallis for his faith in me and for the wonderful opportunity he had given to me to peruse my research in field of Geo-Computation after I completed my MPhil degree under his supervision. He always has been there providing his heartfelt, continues support and has given me invaluable advice, inspiration, and suggestions in my quest for knowledge. Philip became my teacher, an inspiration, and my role model in my life journey.

I would like to express my sincere gratitude to my primary supervisor, Associate Professor Jacqueline Whalley for her patience, motivation, and immense knowledge and helping me to realize the power of critical reasoning. Jacqui was always a source of motivation and a great inspiration as a female in STEM. Her guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better supervisory team for my PhD study. I would like to express my gratitude to Professor Stephen MacDonell for his valuable advice and guidance.

I appreciate the School of Engineering, Computer and Mathematical Sciences (SECMS) for providing me the scholarship to conduct my PhD program. My gratitude to Dr. Victor *Miranda* Soberanis, the statistical lecture at SECMS, AUT and author of VGAM statistical package for lending me his knowledge and being there anytime I needed his advice. I would like to thank Dr Perry Davy from the Auckland City Council for providing me with Auckland data which was used in my thesis and his suggestions.

My sincere appreciation goes to my dear friend Dr Kathy Garden for her encouraging words and continuous support.

My acknowledgement would be incomplete without thanking the best source of my strength, my family. The blessings of my late father, the love of my siblings, and my mother's endless love and support to follow my dream of getting this degree. I would like to acknowledge my gratitude to my dear husband Dr Akbar Ghobakhlou who has patiently supported me and provided his valuable knowledge. Last but not least, I thank my precious gift of life, my son Rayan, who taught me how to see life differently with his pure love and his innocent heart. You rock my life with your extra chromosome son!

Abstract

Air quality and its effect on human health is an area of increased research and interest over the last twenty years. As the world's population increases understanding the effects of human activity on the environment and air quality becomes even more important. The health effects of PM_{10}, the default priority pollutant for New Zealand, was quantified by Kuschel, et al. (2012). They estimated that in 2006, over 600 premature deaths in Auckland were related to air quality. They also reported a cost of over $2 billion to the Auckland region in 2006 as a result of exposure to ambient $PM_{10.}$ Despite this high economic and social cost studies in the literature, both grey and white, related to prediction of Auckland's PM_{10} are sparse. Most of the PM_{10} models in the literature are highly dependent on the input data used. Each model uses different inputs making it hard to compare the effectiveness and evaluate the generalisability of these models. The data used is largely opportunistic – use what we have – rather than informed. Moreover, for many regions including Auckland access to data such as detailed emission inventories, land use, and demographic distributions is not always possible. Hence, the methods in the literature have a limited practical use.

This thesis aims to answer questions related to Auckland's site-specific PM_{10} concentrations, including PM_{10} trends, relative contribution of meteorological sources, and one day ahead prediction of PM_{10} concentration. Semi-empirical, statistical, and geo-statistical methods are explored. Attempts to tackle the challenges of modeling a nonlinear system by using Artificial Neural Networks (ANNs), Long short-term memory (LSTM), and Random Forest (RF) methods are reported. These models are parsimonious and make use of routinely available meteorological data collected from the six fully operational monitoring stations in Auckland during 2011-2016.

It was found that Auckland's PM_{10} has complex seasonal patterns and that PM_{10} concentration trends are very localized and cannot be fully explain by land usage (rural vs urban). Using GAM and GAMM models, not previously used for Auckland, a clear difference was found between the effects of temporal aspects of anthropogenic sources and atmospheric conditions on PM_{10}. For modelling with linear statistics, the main challenge encountered was to find a way to characterise the spatio-temporal dependence structure. The inability to accurately and fully define this structure limits the usefulness of these linear approaches. In contrast it was found that machine learning (MLP, LSTM) and ensemble methods RF were able to account for this underlying structure and for the dynamism of the process. Of all the methods explored the RF model was found to be the most accurate and therefore the most promising avenue for future work.

Chapter 1 INTRODUCTION

1.1 Background, Motivation, and Scope

Atmospheric Particulate Matter (PM) is a gaseous suspension of solid or liquid particles in the earth's atmosphere. It is common to classify PM based on their size into two main modes: The fine particles have a diameter of 2.5 µm ($PM_{2.5}$) or less, and coarse particles a diameter of 10 µm (PM_{10}) or less. Local meteorology, local emission sources and pollution micro-environments as well as characteristics such as land use and building heights impact micro-scale chemical and physical reactions that affect the pollution quantity, size, and composition (Galatioto et al., 2014). PM is known to have significant health effects, especially for infants and the elderly. High concentrations of PM can have environmental impacts such as reduced visibility and acid rain. While high air pollution levels in developing countries hits the headlines, the problem in other places where it has not hit the news headlines, has not disappeared. The problem of urban air pollution in many first world countries is increasing due to rapid population growth, urbanization, and industrialization. In New Zealand, human pressure including burning wood or coal for home heating, fossil fuel consumption by industry and motor vehicle emissions result in a notable rise the PM_{10} levels, therefore PM_{10} is the major air pollutant currently being monitored (PCE, 2015). In a review by EPAQS (2000) it was concluded that based on present evidence the appropriate measure on which to base for air quality standard in United Kingdom is by measuring PM_{10}. In New Zealand, the significant association of PM_{10} with severe health effects and its frequent breaches of national standards and international guidelines makes it a pollutant of most concern (MfE, 2011b).

Due to the adverse health effects of PM exposure, in many cities around the world the ambient concentrations of these pollutants are closely monitored and assessed. In 2016 (the most recent report to this date), an estimated 27 premature deaths, 14 hospital admissions and 31,800 restricted activity days per 100,000 people were attributed to human-made PM_{10} in New Zealand (MfE & Stats NZ, 2018). During 2014-2016, the national short-term standard guideline of 50 µg/m^3 for PM_{10} (24-hour average between midnights) provided by the World Health Organization (WHO) was exceeded by 30 of 51 monitored airsheds (MfE & Stats NZ, 2018). Exceedances of the short-term standard of PM_{10} are primarily a winter problem in most countries including New Zealand (MfE, 2012). Modeling the industrial air discharges in Auckland is not easy due to complex terrain of Auckland along with land-sea breeze interactions and the calm or light wind cycles. The California Puff (CALPUFF) model, an advanced dispersion model, can overcome these limitations (Gimson et al., 2010) but it is a complex model and not practical in situations in which limited monitoring occurs such as is the case for Auckland..

The influence of meteorological factors on particulate pollution levels are well known in literature (Salmond et al., 2016). Therefore, identifying local atmospheric conditions that affect PM_{10} concentrations level, both in general and quantitative manner, are necessary. From a general point of view, the analysis of atmospheric influences gives insight into underlying processes that impact to the observed meteorological variability of PM_{10} (Appelhans, 2010).There are number of studies that have investigated general meteorology controls on PM_{10} concentration in Wellington and cities in the South Island of New Zealand where PM_{10} pollution is more commonly a problem (Appelhans, 2010; Appelhans et al., 2013; Fiddes et al., 2016; Pezza & Mitchell, 2016). For Auckland however, one case study predicting PM_{10} concentration based on a single air quality station using wind speed and wind direction within a short period of time has been published (M. Elangasinghe, 2014). A recent technical report by Talbot & Crimmins (2020) performed trend analysis and investigated the effect of wind speed and wind direction on PM_{10}, $PM_{2.5}$ and NO_2 concentrations in the Auckland airshed. Apart from these as of April 2021, to this thesis author's knowledge, no further examination of meteorological relationships with local concentrations of PM_{10} on larger spatial and temporal scales has been reported for Auckland. To understand the physical nature of meteorological controls of air pollution, all space and time scales need to be considered (Appelhans et al., 2013). In part therefore, this research aims to provide a quantitative analysis of atmospheric influences on PM_{10} concentration in Auckland's airshed within a broad range of spatial and temporal scales including local to regional and daily to inter-annual scale.

The perspective of this thesis is atmospheric science and hence other air pollution related areas, such as PM_{10} emissions sources and health effects are briefly reviewed where applicable. Furthermore, meteorological investigations are used to identify the influence of climate on PM_{10} concentration only and therefore, potential transformations of pollutants in the atmosphere and their chemical and physical reactions are not investigated.

1.2 Objectives

Efforts have been made to analyze and model Auckland PM_{10} concentration. However, the previous section has also highlighted the knowledge gaps in quantitative analysis of meteorological impacts on larger spatial and temporal scales. This thesis aims to answer questions related to Auckland's site-specific PM_{10} concentration, including PM_{10} trends, relative contribution of meteorological sources, and one day ahead prediction of PM_{10} concentration. The objective of this thesis is to develop models built from *pragmatic cost-effective* data from existing low cost and low resolution (sparse) sensor networks as access to the high-resolution networks collecting different sets of predictors is not always possible. Exploring what type of parsimonious models are best for predicting and modeling PM_{10} in Auckland with its unique topology, island climate and geographical location. Specifically, we are interested in the effectiveness of traditional geostatistical approaches and how they compare with machine learning approaches. Semi-empirical, statistical, and geo-statistical methods are explored in an attempt to estimate one day ahead PM_{10} concentration primarily based on routinely available meteorological data collected from the six monitoring stations in Auckland during 2011-2016.

Unlike the theory-driven deterministic models, statistical models use a data-driven approach to modeling air quality. Variation analysis of observed measurements performed by statistical approaches are mainly quantitative extensions of the underlying analysis. Nevertheless, they are an essential progression to provide a basis for assessment of trends which need a quantitative reference. Therefore, Time Series Analysis (TSA) as a necessary step in modeling $PM_{10,}$ is another objective of this study aiming to detect the nature of the phenomenon characterized by the structure of PM_{10} observations over the Auckland airshed

One well known but still to be addressed challenge in using sensor data, and in particular meteorological data such as rainfall, is quality in terms of missing data and incorrect measurements. Thus, another aim of this work is to evaluate the capability of these models to cope with such data issues.

1.3 Research Questions

To meet the objectives of this research the following research questions were framed:

Q1. Data Exploration, Quality, and Imputation:

- Q1.1. What is the quality of the meteorological and PM data available for the Auckland airshed?
- Q1.2. What is the spatial/temporal relationship between Auckland's PM_{10} site measurements and meteorological over the entire region in different time scales (daily, monthly, and yearly)?
- Q1.3. Can missing rainfall data be imputed from satellite-based sources for the Auckland airshed? **and** How accurate are these satellite rainfall measurements?
- Q1.4. How is PM_{10} concentration influenced by Aerosol Optical Thickness (AOT)?
- Q1.5. Do seasonality trends exist in Auckland's PM_{10}? And if these trends exist,

  a) What is the nature of seasonal patterns in Auckland's PM_{10}?
  b) Which seasonality detection method should be used to account for any observed seasonal trends in Auckland's PM_{10}?

Q2. Prediction Models:

- Q2.1. In the absence of comprehensive emission inventories or information on potential sources of emission affecting a particular site, to what degree can the daily concentration of PM_{10} in Auckland airshed be explained by site-specific predictors variables?

Q2.2. Can computationally simple semi-empirical models such as GAMs and GAMM$_S$ be used to model Auckland's daily PM$_{10}$ concentrations reliably?

Q2.3. How can we, from a high-level perspective, use descriptive (marginal) models to characterize spatio-temporal dependence structures for PM$_{10}$ modeling?

Q2.4. How accurate are machine learning methods for predicting next day PM$_{10}$ concentration?

Q2.5. How successful are non-linear statistical and ensemble approaches in predicting next day PM$_{10}$ concentration?

Q2.6. To what extent does data quality become a significant factor in determining the performance of PM$_{10}$ concentration models?

1.4 Methodology

The thesis takes an exploratory and quasi-experimental methodological approach. A breadth of PM modeling approaches is explored in order to determine which types of models, if any, are best for understanding and predicting PM in Auckland, New Zealand. The methods explored need to perform with the data that is available. In Auckland, the PM monitoring sites are sparse and due to lack of funding not monitored and maintained as much as would be ideal. This means that the models need to be robust to missing data or alternative pragmatic and parsimonious measures, such as publicly available satellite imagery are needed. As part of this approach the data quality and trends within the data need to be explored and as appropriate data cleaning needs to take place. Once the data has been established, and fully understood, then different types of models are explored and evaluated. Details of the methods used are presented in the relevant Chapters.

To compare and evaluate the accuracy and adequacy of the presented model, we used the results of those models available in literature. It should be noted that there are very limited studies on some of the models used in this thesis namely TBAT, (S-T) GAM, and (S-T) GAMM, and geostatistical S-T models for PM$_{10}$ modeling reported in literature. Consequently, the comparison of the models in this thesis in terms of accuracy and adequacy is limited. The bench-mark model that could be found for evaluating adequacy and accuracy was the CALPUF (Scire et al., 2000) s model. CALPUF is an advanced physical model with an extensive and complex set of inputs (Holmes & Morawska, 2006). CALPUF EPA recommended model for air quality modelling (Barclay & Scire, 2011) so can be considered to be a benchmark for PM$_{10}$ model performance. In a study for modeling PM$_{10}$ using CALPUF, two metrics namely IOA and FB were used for the purpose of evaluation of the model (Lee et al., 2014). The authors stated that their model showed 'good' agreement based on the 'typical' values of 0.284 to 0.850 for IOA, and 0.043 to 0.821 for FB. For want of a better measure, in this research a "moderate" performing model is one that had an IOA between 0.284 and 0.850. This is a very large range of values but is the performance reported for CALPUF and accepted model.

1.5 Contributions

Considering all the above discussed, this study contributions can be summarised as below:

- Identify the strengths and limitations of the existing state-of-the-art PM models through the related literature review. This thesis provides an evaluation of the strengths and limitations of existing PM models that are reported in the literature between **2000** and **2020.** While others have attempted this, they have not included the grey literature related to New Zealand's PM modeling efforts. Other works tend to focus on the regions within the Northern Hemisphere.
- Provide a quantitative analysis of the available atmospheric influences on PM_{10} concentration in Auckland's airshed across a wide range of spatial and temporal scales, from site-specific to regional and daily to inter-annual to establish a quantitative relation between various atmospheric conditions and PM_{10} concentration. This quantitative analysis was either not applied in Auckland PM models that are reported in the literature between **2000** and **2020** or were limited to only a few atmospheric variables such as wind direction and speed gathered from a single station over a short (few months) duration.
- Evaluate nonparametric local smoothing methods as a means for identifying the predictors influencing the short-term variation of PM_{10} over the Auckland region.
- Explore the potential of spatio-temporal statistical and geostatistical methodologies for modeling of PM_{10} concentrations over the Auckland urban area. These semi-empirical modeling approaches are rarely undertaken for New Zealand. For the Auckland air shed in particular the studies are limited over space and time.
- Undertake scientific inference to determine the importance of covariates on PM_{10} concentration in the presence of spatio-temporal dependence; and estimating the future value of the PM_{10} concentration at a specified predication grid, along with the uncertainty of that estimation.
- Examine the feasibility of effectively modeling regional-scale urban air pollution using machine learning approaches such as: artificial neural networks (ANNs/MLP), Long Short-Term Memory (LSTM) and Random Forest (RF) as an ensemble method. This examination allows for the comparison of the modeling capacity of such models for short-term prediction of PM_{10} over the study area. Another contribution of this thesis is to introduces an optimisation framework for tuning the LSTM model parameters to obtain a fair trade off between accuracy and generalization of the LSTM model. This includes tuning

the epochs, batch size and number of neurons parameters and running the diagnostics test on selected parameters. Such tuning process is lacked in literature for LSTM models on air pollution as these parameters were selected without tuning process.

- Investigate data mining techniques for identifying the most influential meteorological and temporal (day of the week, month) parameters on PM_{10} concentration and analyze how the MLP based model performs with different combinations of input parameter. Additionally, this investigation allows for an examination of the robustness of the MLP PM_{10} model to input data errors.

1.6 Thesis Structure

The second Chapter provides a general introduction to Particulate Matter (PM) and recent literature related to PM and factors which influence PM concentrations. However, it is important to note that because of the diverse and complex nature of the methods used in this thesis the literature reviews related to specific methods and their use in the modeling of air pollutants and/or PM is presented in each Chapter.

The aims of this research will be delivered in the rest of this thesis, organised as follows:

Chapter 2 introduces atmospheric aerosols and in particular PM. The chemical composition, size distribution and the natural and anthropogenic sources of PM are described. Subsequently an examination of the spatial distribution of PM on an international and national level is presented. Finally, the latest research on influencing factors on PM concentration such as topographical and climate characteristics are outlined and discussed in this Chapter, paving the way for a site-specific comparison in Chapter 3.

Chapter 3 describes the data utilised in this thesis along with an exploration and analysis of the data. A comprehensive assessment of Auckland PM_{10} concentrations variation and associated trends through analysis of five years (2011-2016) of data is provided.

Chapter 4 provides a general introduction to time series analysis followed by the results of an approach to time series analysis that will form the basis for much of what is done in Chapters 5, 6, and 7 which look at statistical models, and spatio-temporal statistical and geostatistical methods, respectively. The concept of dependence and stationarity is introduced, and site-specific stationarity test are provided in Section 4.3, prior to the reporting of time series analysis and modeling approaches. A review on methods for seasonally adjusting a time series followed by the results of performing site-specific seasonal decomposition analysis experiments are provided in Section 4.6. Time series analysis for such complex seasonalities are performed using dynamic harmonic regression with multiple seasonal periods and adopts the use of a relatively new method, TBAT. Details of these method, and related experiments, and results are provided.

Chapter 5 examines the use of nonparametric local smoothing methods based on Generalized Additive Models (GAMs) and Generalized Additive Mixed Models (GAMMs) to model the influence of meteorology on the distribution of daily average of PM_{10} concentration. These models are used as alternative analytic approaches to the advanced Auto Regressive (AR) models discussed in Chapter 4 to avoid specification of a parametric form for seasonal trends allowing a more robust model against model misspecification. Section 5.2 provides a literature review on the use of GAM for modeling and predicting of air pollution including PM_{10}, then the experimental method and smoothing parameter selection are discussed. Finally, the results of applying GAM to the Auckland airshed data are presented. Section 5.3 provides a literature review on GAMMs followed by a description of the experimental GAMM method used in this research. The main drawback with GAMMs – the assumption of independency between the observations of response – for use in Auckland is highlighted and discussed. Finally, site-specific experimental results are provided and discussed in Section 5.4. A conclusion and suggestions for future work using nonparametric local smoothing methods are provided in Section 5.5.

Chapter 6 models and maps PM_{10} concentration by applying spatio-temporal statistical approaches namely Regression (Trend-Surface) Estimation, Spatio-Temporal (ST) GLM and ST-GAM. Exploratory analysis of spatio-temporal data and space-time modeling is described in Section 6.2. In Section 6.3, spatio-temporal PM_{10} prediction is obtained using a statistical regression model assuming that the “trend” terms can take all the spatio-temporal dependencies into account. The regression model that attempts to account for spatial and temporal trends is described followed by the results and a discussion of the findings. ST-GLM and ST-GAM are applied to explore the spatio-temporal patterns of the PM_{10} concentration integrating with secondary information at different spatial resolutions and temporal aggregations. Parameter inference for models using Ordinary Least Square (OLS) along with spatio-temporal analysis of residuals are performed by computing and visualizing the empirical spatio-temporal semivariogram of the residuals. Section 6.4 provides the conclusion of this Chapter.

Chapter 7 introduces a spatio-temporal geostatistical model which allows space-time predictions in 24 hours temporal resolutions. The aim is to model and analyze spatio-temporal point referenced data, where PM_{10} measured over time at several spatial locations, which vary continuously over the study region. The model performance is analyesd and evaluated against both random and block test sets used in Chapter 6.

Chapter 8 in this Chapter machine learning approaches namely MLP and LSTM techniques and an ensemble method (RF) approach are applied, and their modeling capability are compared and evaluated. In Section 8.2, the input selection methods (forward selection and backward elimination) and PCA was used to find the best sets that can describe the PM_{10} concentration as the model inputs. There site-specific MLPs were obtained and performance of input selection on MLP were analyzed and the related model was selected as base model (Section 8.3). The best set of site-specific input sets are then used in LSTM (Section 8.4) and RF (Section 8.5). Since PM_{10} distribution are not even, K mean clustering was applied to ensure PM_{10} data are evenly distributed in train, test, and validation sets. Mathematical formulation and literature review of each model is provided followed by experiment methodology and results and discussion. Conclusion of the Chapter, limitation and future studies is provided in Section 8.6.

Chapter 9 discusses the model application and challenges in PM_{10} modeling. Research questions raised in Introduction Chapter are answered and discussion on limitations and future works are provided.

Chapter 2 ATMOSPHERIC AEROSOLS

The aim of this Chapter is to provide an up-to-date general view of air quality. This Chapter is organised in five sub-sections. The first section introduces a brief history of air pollution episodes and major related events in Northern Hemisphere. The second section briefly describes different types of atmospherics particulate matters, their chemical composition and life expectancy. Section 2 also provides details of the instrumentation and measurement devices used to measure PM concentration and the recommended monitoring methods in New Zealand. The third section provides contextual information on air pollution at an international level. The fourth part describes the state of New Zealand ambient air (from the beginning of the monitoring up to present). Section five provides information on the legal and institutional arrangements for managing the environment in New Zealand as they relate to air quality. Efforts have been made to incorporate the available information up until mid-2017., Historical information, from the original sources, was hard to find and the author had to search earlier literature reviews for some of the information. Most of the information in this section was retrieved from regional councils and government department's online archives.

2.1 Introduction

Air pollution is not a new phenomenon. Urban air pollution was referred as *gravioris caeli* (heavy heaven) or *infamis aer* (infamous air) by the ancient Romans (Hughes, 1993). The urban outdoor air pollution in large cities has emerged over the course of the millennia as civilisations became more organised. Historically, coal burning was the main source of PM and was measured as black smoke (Harrison, 2020). In the 9th century the problem of urban outdoor air pollution was first recognised in England. King Edward I established a community for redemption of pollution in 1285 (Markham, 1994). However, by the mid-13th century, forest resources depletion forced London to shift to coal as the main source of energy, launching 700 years of pollution (Brimblecombe, 1999). In the 19th century, thousands of people died in London as a result of severe air pollution episode. Air pollution remained as a significant unaddressed challenge until the mid-twentieth century. Table 2.1 summarises the air pollution episodes and key historical milestones related to the history of air pollution highlighting the concerns and compliance strategies taken to action to address these issues.

In the second half of the twentieth century, black smoke emissions from coal burning were reduced in developed countries. Road traffic exhaust emissions and secondary pollutants such as ammonium salts from agricultural activities and secondary organic carbon became the dominant sources of pollution (Harrison, 2020). Unbalanced urban development and the substantial growth of mobility and road traffic are the main factors attributed to the degradation of urban air pollution in the early twenty first century.

Moreover, pollution in urban areas has been shown to mainly consist of industrial and automobile emissions still in current years (Manoli et al. 2002).

Table 2.1: Early history of severe air pollution, milestones, and compliance strategies (Morrison, 2016; Smith, 2017).

Year	**Event**	
1157	"Unendurable" air pollution from wood smoke	Henry II's wife Elanor of Aquataine to flee Tutbury Castle.
c.1590	Coal smoke	Queen Elizabeth "greatly grieved and annoyed".
1880	"killer" fogs reported in London due to a January inversion.	
1892	London smog	1,000 deaths
1898	Pittsburgh air pollution	People leaving the city. Appointment of Committee on Smoke Abatement by the chamber. The Engineer's Society of Allegheny County rejected to collaborate emphasizing on Legislation and not engineering.
1909	Glasgow, Scotland, winter inversions and smoke growths	1,000 deaths in "Old Reeky" city.
1930	Industrial Meuse River Valley, Belgium, 3-day weather inversion	63 deaths and 6,000 made ill
1939	St. Louis smog episode	Lanterns were used in daylight for one week. The St. Louis Post Dispatch started a campaign due to the smog episode.
1948	Donora, Pennsylvania smog incident. London killer fog	22 deaths, 600 hospitalized, 1,000s ill. 600 deaths
1948	The New York Times urged women to participate an anti-pollution demonstration in New York city.	
1949	First US conference on air pollution sponsored by Public Health Service.	
1950	Mexico, Poza Rica, an oil refinery gas fumes caused killer smog.	22 deaths, hundreds hospitalized
1952	Dec4-Dec8: worst of the London "killer fogs."	4000 people deaths. Vehicles used lights in daylight. Busses run only with a guide walking ahead. By Dec. 8, all transportations except for the subway were prevented.
1953	Smog incident in New York	between 170 and 260 deaths.
1954	Los Angeles Heavy smog conditions	Schools closed for most of October. 2000 auto accidents in a single day
1955	New York City hosted the First International Air Pollution Congress.	
1956	London killer smog	1,000 deaths Clean Air Act was passed by British Parliament.
1960	Two-year Public Health Service research on car emitted air pollution was funded by US Congress.	
1962	London smog	750 deaths
1963	New York City, Major smog event.	

1966	New York City, Major smog event during November 23–26.
1970	Replacement of lead additives with catalytic converters was urged by General Motors president Edward Cole as he promised “pollution free” cars by 1980.
1970	April 22- First Earth Day: Environmental concerns were raised by a national political member. Demonstration by Millions of Americans for clean air and water and preservation of nature.
2017	March 28 — An executive order was signed by Donald Trump dismissing air pollution and greenhouse gas regulations.

Ambient air quality denotes the near ground level outdoor air quality, that is away from direct sources of pollution. High concentrations pollutants that affect human health and/or the environment causes poor ambient air quality. An important factor influencing particles deposition in the respiratory tract affecting human health is the size of particle. Fine and ultrafine particles usually penetrate in lung whereas coarse particles accumulate mostly in the nose and throat. Fine and ultrafine particles are generally considered to be more toxic and appear in greater numbers with larger surface area than coarse particles of the same mass (Mahapatra et al., 2018). The unanimous scientific consensus on air quality degradation as a major environmental and health hazard, caused significant research efforts in the field of air pollution. Airborne PM exposure has been associated with increased mortality as well as adverse effects on respiratory, cardiovascular, and neurological conditions (Sun & Zhu, 2019). High concentration of PM can have both persistent and short-term effects on human health from respiratory irritation and cardiovascular disease, to cancer and premature death (Manisalidis et al., 2020). The findings of the Air Pollution and Health: A European Approach 2 (APHEA2) project on short-term effects of ambient particles on mortality showed a 0.8% increase in the daily deaths of the elderly was associated with a 10 µg/m^3 increase in PM$_{10}$ (0.7–0.9%) (Aga et al., 2003). A study by Sun & Zhu (2019) examined air pollution related health studies between 1992 and 2018 and concluded that the most common health related effect of outdoor air pollution was mortality rate.

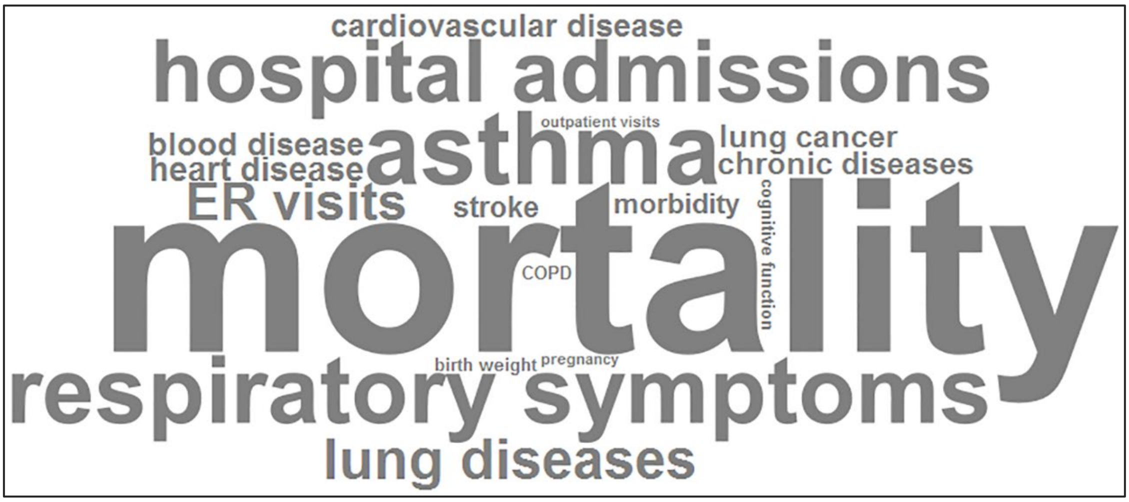


Figure 2.1: Outdoor air pollution related health consequences presented in word cloud. The higher frequency of the outcome is presented in bigger font size (Sun & Zhu, 2019) permission granted from PlosOne journal .

According to World Health Organisation (WHO) (2016) low- and middle-income countries account for nearly 90% of air pollution related deaths, with approximately two out of three taking place in WHO's Western Pacific and South-East Asia regions. In 2019, seven million premature deaths, massive loss of crops, and declines in biodiversity across Europe, North America and East Asia were reported due to poor global air quality (David et al., 2020). All impacts are caused by the accumulation of airborne particulate matter, both chemical and physicals making analysis and monitoring of these characteristics crucial. Local studies have reported various health effects of PM on New Zealanders (Emission Impossible Ltd, 2019; Fukuda et al., 2011; Hales et al., 2012; Wilton, 2005). In 2010, the Health and Air Pollution in New Zealand (HAPiNZ) study was established to investigate the environmental, health, social and economic costs of air pollution from all sources in New Zealand (Travis, 2012). The study revealed that anthropogenic PM_{10} pollution caused nearly two million restricted activity days. The HAPiNZ study approximated the total cost of air pollution in terms of restricted activity days and hospital admissions was nearly \$1.3 billion per year. The HAPiNZ study also estimated that the most significant contributor to more than 1600 premature deaths was PM_{10} emissions (Kuschel & Mahon, 2010). In New Zealand, anthropogenic PM_{10} was associated with 277 premature deaths (27.2 per 100,000 people), 236 cardiac hospitalisations (5.0 per 100,000 people), 440 respiratory hospitalisations (9.4 per 100,000 people), and 1.49 million restricted activity days (31,839 per 100,000 people) in 2016 (latest report to date). The South Island had the highest number of modelled cases of premature deaths, hospital admissions for cardiac and respiratory problems, and restricted activity days per 100,000 people. The PM_{10} associated health effects (per 100,000) were decreased compared to 2006. This decrease however is likely due to increase in population in lower PM_{10} areas rather than a reduction in PM_{10} concentration (MfE & Stats NZ, 2018).

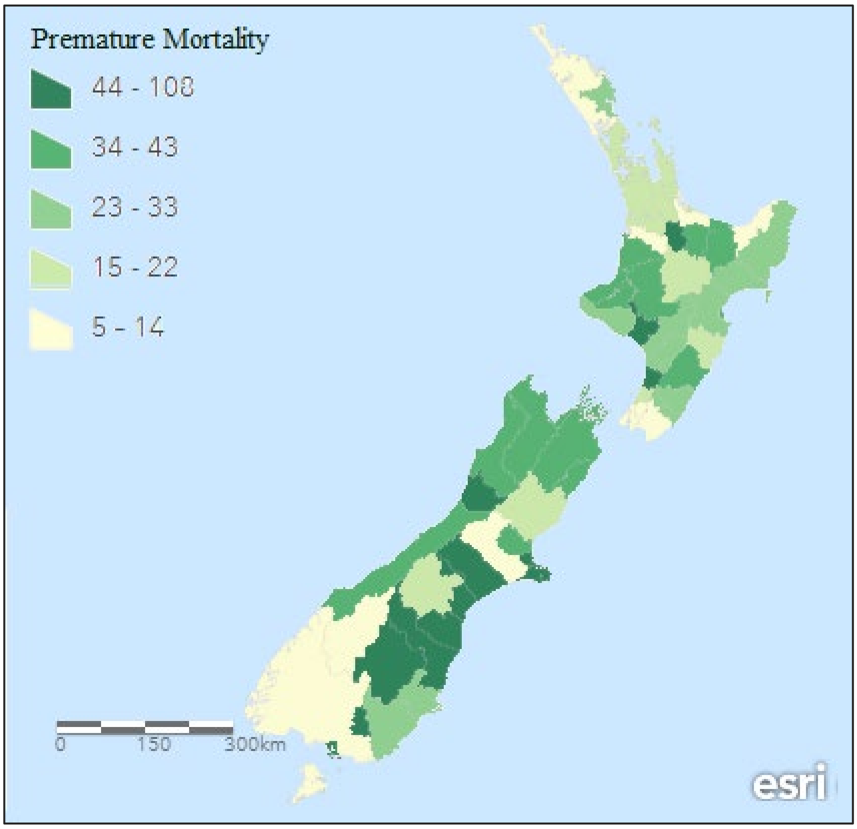


Figure 2.2: Estimated number of premature mortalities per 100,000 persons, by territorial authorities, in 2016 Adopted from (MfE & Stats NZ, 2018).

The report discussed that between 2006 and 2016 there was a negative health effects per 100,000 persons which was possibly due to more people living in lower PM areas, than an actual decrease in PM exposure (MfE & Stats NZ, 2018). The presence of high PM in the air also effects climate change as it changes the amount of incoming solar radiation and Outgoing Longwave Radiation (OLR) maintained in the earth's system. Particle properties such as composition and size have significant impacts on visibility degradation. Brown haze is a visual indication of degradation in air quality (Salmond et al., 2015). In Auckland, brown haze builds up during the calm and cold winter mornings and nights (Senaratne & Shooter, 2004). According to Salmond et al. (2015), 88 haze events were observed intermittently in Auckland Central, 43 of them were categorised as severe. Visibility degradation due to brown haze was a major concern in Christchurch, New Zealand where PM has been shown to contribute to a significant amount of light extinction (Wilton 2003). According to Ancelet (2012), the contribution of PM from motor vehicle emissions and secondary particles in poor visibility was more significant than the contribution of PM emitted from biomass burning. Atmospheric aerosols affect the ecosystem and vegetation. The effect of dust on ecosystem depends on the nature of the environment and the rate of dust transmission from the air to vegetation surfaces. This rate of dust transmission depends on the dust's properties (Burkhardt & Grantz, 2016). In this Chapter, the focus is primarily on PM_{10} which is PM with a diameter between 2.5 µm and 10µm. PM_{10} is the data that is explored, experimented with, and modeling in this thesis.

2.2 Atmospheric Aerosols Background

Atmospheric Aerosols are relatively small solid/liquid or mixed particles that are suspended in the atmosphere. Primary aerosols are emitted directly to the atmosphere whereas secondary aerosols are created from precursor gases (Fadnavis, 2020). Fossil fuel combustion are the main sources precursor gases followed by volatile organic compounds (VOCs) biogenic emissions and fires. Volcanic eruptions can also produce primary and secondary aerosols at the ground level as well as in the stratosphere (Boulon et al., 2011). Primary aerosols can consist of both inorganic (sea spray, mineral dust, and volcanoes) and organic (caused by anthropogenic sources) components. Anthropogenic sources include combustion engines, biomass burning for households and industry energy production, industrial activities such as building and mining, traffic related activities such as pavement destruction by vehicles and braking, and wearing of tyres (Silva & Mendes, 2011).

2.2.1 Particulate Matter (PM)

The term PM refers to any airborne material in the form of particles and includes pollutants which contains various mixture of solid and liquid particles. Airborne particles are complex and diverse in their: physical properties (see section 2.3.1), chemical composition (see section 2.3.2), and their mechanism of formation or origin (see section 2.3.3) and by what is measured by a particular sampling technique (see section 2.3.4).

2.2.2 Physical Properties of Atmospheric Aerosol Particles: Size Distribution

A principal common feature is particulate discrete units ranging in size aerodynamic diameters of several nanometers to about 100 µm in diameter. In general, particles are categorized into two modes based on their size. Particles with an aerodynamic diameter 2.5 µm < d < 10 µm (PM_{10}) are coarse mode whereas a fine mode is made up of particles with aerodynamic diameters < 2.5 µm (generally denoted as $PM_{2.5}$) (Travis, 2012) (Figure 2.3). The size of aerosol particles changes in the atmosphere through the processes of growth or removal. Coagulation and condensation increase the particle size whereas evaporation and removal processes cause reduction of PM size (Lu & Ren, 2014).

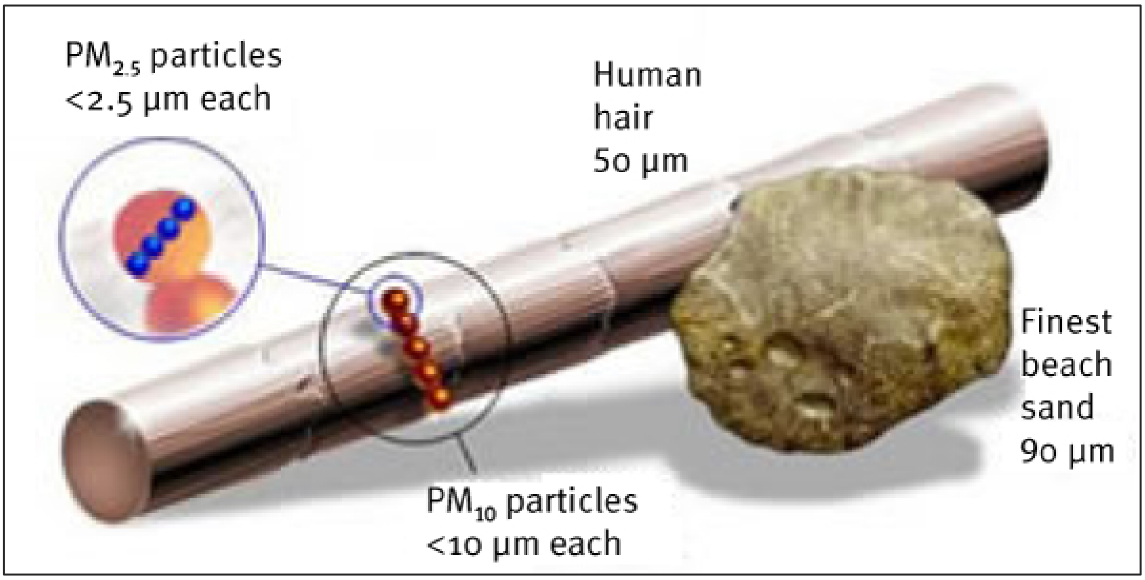


Figure 2.3: Particle sizes (MfE, 2009).

Fine mode particles can be subcategorised based on their size and formation process (Travis, 2012):

Nucleation Mode corresponds to particles with diameters ranging from 0.005 and 10 nm, formed through nucleation processes. The nucleation process takes place when the number of particles is not sufficient to scavenge the molecule of interest and the liquid or solid phase for a molecule is more energetically feasible. The nucleation process can occur during condensation of hot gases or when the gas phase reactions create lower volatility or species that are highly hygroscopic. In each of these cases clusters of molecules lead to the production of a new particle, in a process known as gas-to-particle conversion (Seinfeld & Pandis, 1998).

Aitken Mode made up of particles (with diameter between 10 nm < d < 100 nm) originating from vapor nucleation, condensation, or coagulation of particles. Particles grow through condensation to form a larger secondary particle when pre-existing particles make contact with reactive vapors which condense onto the particle surface. Coagulation results in a shift in the aerosol-size distribution toward larger particle sizes and increases particle mass effectively (McMurry et al., 2004) . The Aitken nuclei mode and coarse mode lifetime is largely determined by rain washout (Chatea & Praneshab, 2004).

Accumulation Mode particles with diameter between 0.1 µm and 1 µm that are result of primary emissions and formed mainly by the coagulation of smaller particles or the condensation of organics from the gas phase or the vapor constituents.

Ultrafine Particles: refers to particles in the Aitken and nucleation modes. Although the number of ultrafine particles is high, they have the smallest mass and volume.

In urban/industrial particle growth at high relative humidity and interactions of aerosols with clouds, are the main causes of aerosol size variation. Inorganic primary particles in coarse mode have short lifetimes, usually only a few days, in atmosphere due to gravitational settling. On the other hand, secondary aerosols, in fine mode, stay in the atmosphere longer, from days to weeks , and as a result can be transported over large distances (Griffin, 2013). During episodes of long-range transport of dust, particles from Sahara Desert can be transported as far as the United States and the Amazon (Chin et al., 2007). The findings of a study by Masri et al., (2015) showed a sharp decrease in $PM_{2.5}$ level compared to PM_{10} over time, suggesting that the significance of PM_{10} should be taken into account in future traffic-related air pollution policies. The concentrations of the ambient air PM are measured and recorded in term of the mass of PM in one cubic meter of air, mainly using the microgram per cubic meter ($\mu g/ m^{-3}$) units.

2.2.3 Chemical Composition and Sources of Atmospheric Aerosol Particles

This thesis does not focus on PM_{10} chemical composition; however, it is crucial to recognize that PM_{10} is a complex mixture of chemical compounds with great dependency on the atmospheric conditions. Airborne particles include both major and minor components. Trace metals such as copper and lead are categorized as minor components and present at very low levels. The use of trace metals in industrial products or as impurities or additives in fuels can cause this rather small concentration to the atmosphere. The major components usually include (Laongsri, 2013):

Sulphate arises from atmospheric oxidation of SO_2. A small primary component arises from sea salt or mineral matter such as Gypsum.

Nitrate normally present as NH_4NO_3, which results from the neutralisation of HNO_3 vapor by NH_3, or as sodium nitrate (NaNO3), due to displacement of hydrogen chloride from NaCl by HNO_3 vapor.

Ammonium generally present in the form of ammonium sulphate ($(NH_4)_2SO_4$) or NH_4NO_3.

Sodium and chloride found in sea salt.

Elemental carbon high temperature combustion of fossil and biomass fuels forms the black, graphitic carbon. The organic carbon from organic compounds can form from automotive or industrial sources as primary source. The secondary organic carbon results from the volatile organic compounds oxidation (VOCs).

Mineral components Aluminum, silicon, iron, and calcium elements can be found in rock and soil as crustal materials. These generally appear in coarse dusts emitted from wind-driven entrainment processes, mining, construction, and demolition processes.

Water Water soluble components of PM, mainly Ammonium sulphate ($(NH_4)_2SO_4$), NH_4NO_3 and Sodium Chloride (NaCl), absorb water from air at high relative humidity, transforming from crystalline solids into liquid solution droplets. In a hysteresis effect, particles will still maintain bound water indicating a significant component of the mass, even after drying at 40–50% relative humidity.

Various studies on source apportionment of ambient PM have been conducted around the world. Kim et al. (2004) studied the fine particles source apportionment in Washington, DC using the thermal optical reflectance (TOR) (Han et al., 2007) method. The authors identified 10 sources namely (SO4 2-) rich secondary aerosol I, gasoline vehicle, (SO4 2-) secondary aerosol II, nitrate-rich secondary aerosol, (SO4 2-) rich secondary aerosol III, incinerator, aged sea salt, airborne soil, diesel emissions, and oil combustion. Gugamsetty et al. (2012) conducted source apportionment study in New Taipei City, Taiwan on PM, PM_{10}, $PM_{2.5}$ and $PM_{0.1}$ were collected simultaneously, using a dichotomous sampler and the Positive Matrix Factorization (PMF) (Lee et al., 1999) method. The authors found contribution of both anthropogenic and natural source processes in their study area. In the study by Marsi et.al., (2015) the PM chemical compositions and sources apportionment were investigated in Boston over nine years of simultaneous collection of 2000 samples of fine and coarse particles. The results suggested a significant difference in elemental compositions of coarse and fine particles, showing different sources and mechanisms of formation, as well as distinct annual trends and seasonal variation. The study showed that between 50 and 75% of Al, K, Br, and Ba elements (the major components of crustal and road dust) measured in PM_{10} were caused by the coarse mode. Over 75% of the elements found in the coarse mode included the crustal and road dust elements Ca, Si, Ti, Fe, and Mn as well as Cl (sea salt) (Figure 2.4).

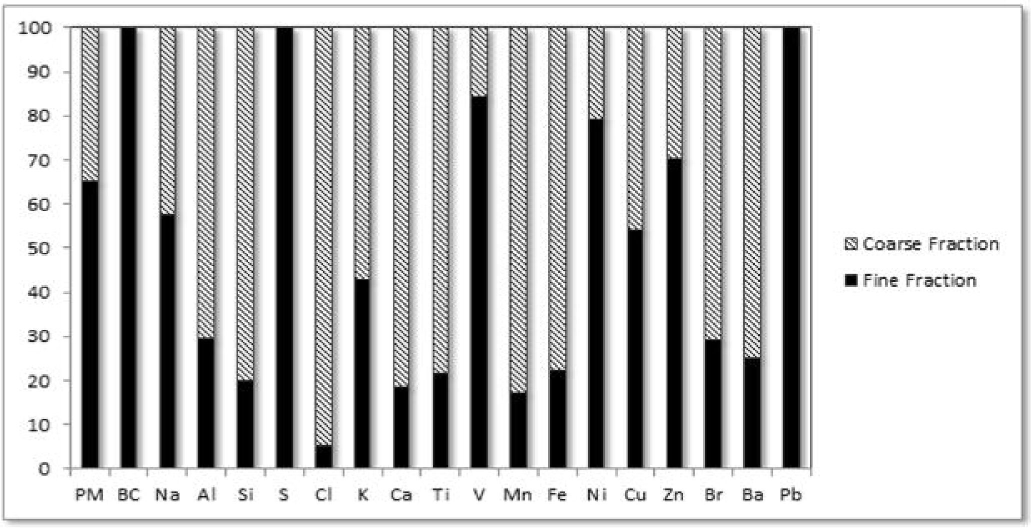


Figure 2.4: Relative contributions of coarse and fine PM to total PM_{10} mass *(Masri et al., 2015)* with permission.

PMF identified six sources for $PM_{2.5}$ including regional pollution (48%), motor vehicles (21%), sea salt, crustal/road dust, oil combustion, and wood burning (19%). The three source types reported for PM_{10} included: crustal/road dust (62%), motor vehicles (22%), and sea salt (16%) (Figure 2.5).

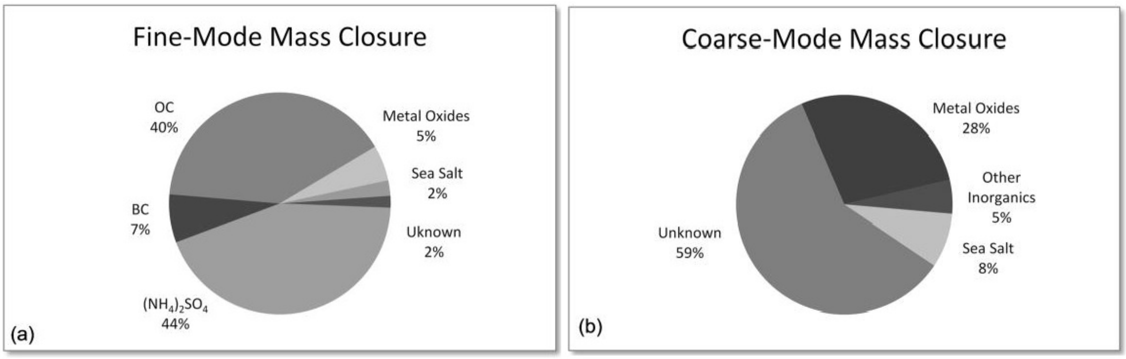


Figure 2.5: Results of mass closure for fine and coarse particles adopted from (Masri et al., 2015) with permission.

Dust Deserts or semi-arid areas produce tons of dust particles which are dispersed into the atmosphere. Dust particles are generally categorized as PM_{10} as coarse mode mass concentrations are higher than those in the fine mode for soil dust aerosols over desert surfaces. The west coast of North Africa is a well recognised source of dust (León and Legrand, 2003; Prasad and Singh, 2007). The strong vertical thermal turbulence, caused by warming of the surface during the day, is usually followed by cycles of nocturnal stability. This cycle prevents the deposition of suspended particles to atmospheric height so that particle stays at these altitudes for weeks or even months distances, creating 'dry smog' events (EEA, 2012). In 1998 (from late April to early May) an Asian dust event resulted in 65 µg/m^3 of PM_{10} concentrations over the entire area, whereas normal levels would be 20 µg/m^3 (U.S. EPA, 2002). A study by (Suzuki & Taylor, 2003) reported the continental effect of the Asian dust event in Chilliwack, British Colombia, during this period recording high PM_{10} concentrations of 120 µg/m3 and 44 µg/m3 concentrations of $PM_{2.5}$. In a study on African dust in the Mediterranean region seasonality of African dust episodes over the whole Mediterranean basin were summarised. The study revealed the African dust PM_{10} patterns are almost comparable with the increasing PM_{10} values from north to south and from west to east throughout the basin (Querol et al., 2019). In Auckland dust is a contributor to overall PM and tends to be a result of construction and road traffic (Davy et al., 2017). The composition of Auckland's PM is discussed in detail in Section 2.5.1)

Sea Salt The main source of salt aerosols in the atmosphere are the world's oceans. Sea-spray aerosol is a mixture of inorganic sea salt and organic matter. In coastal regions such as Eastern Mediterranean ,sea salt could contribute up to 80% of the annual mean particulate mass (Gerasopoulos et al., 2006). NaCl with traces of magnesium (Mg) and sulphate (SO^{2-} 4) are the main component of sea salt. Sea spray can contribute to raise of PM_{10} concentrations in the air as sea-spray particle size varies from less than one micrometer to a few micrometers (EEA, 2012). As Auckland is located on two harbours and most areas are close to the sea, sea spray is a contributor to PM_{10}. In a recent source apportionment study ((Davy et al., 2017)), PM concentrations dominated by sea salt in Auckland showed a downward trend for all sites in Auckland that can be result of larger inter-decadal cycle related to Southern Hemisphere circulation patterns or a more permanent trend.

Volcanic Emissions Precursor gases, water insoluble dust and ashes are emitted in the atmosphere during volcanic activities. These volcanic activities are mostly found in certain areas in Iceland and in the Mediterranean area especially Italy and Greece, where sudden eruption of volcanoes can potentially produce short-term peaks in PM_{10} levels within Europe (EEA, 2012). During April and May 2010 (MACC, 2010) the plume of the Eyjafjallajoekull glacier eruption in Iceland reached altitudes of six to seven kilometers. The suspended ash increased the 24 hours mean concentration of PM_{10} to 1230 µg/m in Vík, a small town located 38 km south east of the volcano. On the day after the volcanic activity ceased, the level of PM_{10} concentration was 25 times the health limit (Thorsteinsson et al., 2012). During eruption of Soufriere Hills volcano, located on Caribbean Island of Monserrat, the volcanic ashes travelled 80 km at the south east of the volcano and reached the town of Pointe-a-Pitre in 2010 (Molinié et al., 2014). The study showed 11 hours after the major eruption, the mean hourly PM_{10} mass concentration increased to 271µg/m3 causing a partial dome collapse in the crater. Auckland is located in a volcanic field of over 50 volcanoes. The most recent eruption was Rangitoto which occurred around 600 years ago. Because the field is not extinct there is the potential for a significant PM10 event linked to volcanic activity but during the period of this study there were no notable events.

Fires Pollutants emitted during a wildfire include atmospheric PM and gaseous compounds, such as carbon dioxide (CO2), carbon monoxide (CO), methane (CH4), nonmethane hydrocarbons (NMHC), nitrogen oxides (NOX), and nitrous dioxide (N2O). Results of the 3 years of air quality monitoring in the agglomeration of Porto Littoral showed forest fires contributed (35%, 8%, and 18%) to PM10 pollution episodes from 2001 to 2003, respectively (Miranda et al., 2008). In 2017, wildfires caused episodes of high PM concentration in the Iberian Peninsula. The measured daily mean PM10 concentrations averaged over a region covering the Iberian Peninsula, France, Benelux and the south of Great Britain before and after the episode were 17 µg/m3. This average daily mean was increased 26 µg/m3 during the fire episode (EEA, 2020). Fires are events that affect local Auckland PM measurements, most notable in recent years was the Sky City Fire (an International Convention Centre). This event caused PM to exceed the National Environmental Standard for Air Quality for the first time in a decade.

Most aerosol particles are unstable and can change by growing or might disperse from the atmosphere and sink on the surface. The aerosol particle lifetime varies from days in the troposphere to a year or more in the stratosphere (Griffin, 2013) and is determined by the efficiency of the removal mechanism. Wet, dry, and occult depositions are the three main mechanisms for particle deposition onto vegetative surfaces (Grantz et al., 2003). Wet deposition is mostly a function of PM concentration and precipitation rate. Dry deposition is the deposition of particles by convective transport, dispersion, and adhesion to the Earth's surface. Dry decomposition acts as a sink for aerosol particles and therefore is more related to local air quality than a global scale (Janhäll, 2015).

2.2.4 Concentration Measurement Methods

In this section, special attention will be given to PM concentration measurement methods and equipment. PM concentration measurements are key to the standardization and regulation of emission thresholds. Several methods and instruments for measuring different characteristics of particulate matter are outlined below:

2.2.4.1 Gravimetric Method

The weight of filters before and after the sampling period are used to determine particle mass concentration. All available PM granulometric fractions (nucleation, accumulation, and coarse modes) are collected by the filters in a resolution time of 15 minutes or more. The filters are packed under controlled conditions of temperature and relative humidity. Cascade Impactor is one gravimetric instrument used for measuring PM mass (Amaral et al., 2015) in addition to size distribution methods (Giechaskiel et al., 2014).

2.2.4.2 Optical Methods

Optical instruments can be used for real-time monitoring of PM_{10} concentrations. These instruments use the principles of scattering, absorption, and light extinction to conduct a measurement (Amaral et al., 2015). A light beam is used to light the PM and to scatter this light in all directions. Part of this light is transformed into other forms of energy (absorption) at the same time. The light extinction is calculated by adding the degree of scattering and absorption. The Optical Particle Counter (OPC) instrument uses a diode laser as a light source, to light a sample of particles from every angle. A photodetector is then used to measure the light that scattered from the particles. Particles are then counted and measured simultaneously based on the intensity of the flash (Giechaskiel et al., 2014).

2.2.4.3 Microbalance Methods

Over the surface of an oscillatory microbalance element, microbalances use resonance frequency variation to find the PM properties. Tapered Element Oscillation Microbalance (TEOM®) and Quartz Crystal Microbalance (QCM) are the two common instruments that use the microbalance method. In TEOM, PM mass is measured based on the variation of resonance frequency of a tapered quartz wand, caused by particle buildup in a sampling filter. In QCM, the piezoelectric property of the quartz crystal is used for measuring PM mass. Resonance frequency changes when there is a small addition of mass to the quartz crystals surface. In QCMs particles are deposited by electrostatic precipitation in a fine quartz crystal resonator Microbalance. (Giechaskiel et al., 2014). Figure 2.6 illustrates PM measurement methods and instruments.

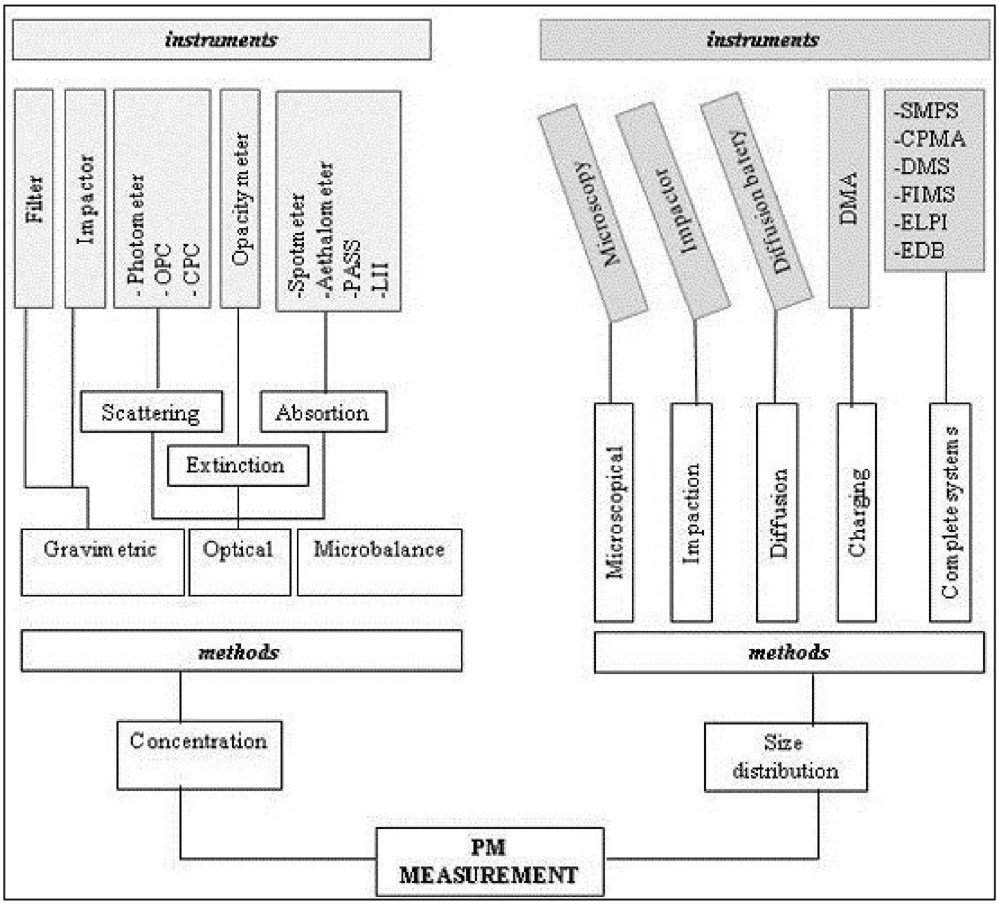


Figure 2.6: PM measurement methods and instruments (Amaral et al., 2015) with permission.

In New Zealand, the recommended monitoring methods to establish compliance with the 2002 Ambient Air Quality Guidelines (AAGQ) were reviewed (MfE, 2002) and a new PM_{10} monitoring method ,US 40 CFR Part 50, is recommended. It is also stated that in case of using TEOM® for monitoring PM_{10} and $PM_{2.5}$, "a*nother recommended monitoring method should be co-located at the site for at least one year to calculate an appropriate adjustment factor"* MFE (2002, p. 32.)

2.3 Air Particulate Matter Management at International Level

The World Health Organization (WHO) guideline for a 24-hour average PM_{10} is 50 μg/m^3, with three exceedances accepted per year (WHO, 2017). Any exceedances must be reported. The standards adopted by different countries may vary. However, most countries implement the target standard of 50 μg/m^3 but differ on the number of permitted exceedance days. While some countries agreed that natural events should be excluded from the count of exceedances of because they are outside the control of the region,, New Zealand, however does not provide such exceptional event exclusions (MfE, 2011a).

The ambient air quality directive (EU, 2008) places thresholds for both short term (24 hour) and long-term (annual) PM_{10} concentrations. The EU 24-hour PM_{10} threshold (50 μg/m^3) is often exceeded in

Europe. The EU annual PM_{10} threshold is 40 µg/m^3. WHO sets stricter Air Quality Guidelines (AQGs) aiming to reach the lowest concentrations possible. A comparison of the two standards is provided in Table 2.2.

Table 2.2: Air quality standards for protecting human health from PM_{10} (EEA, 2020).

Averaging period	**Standard type: PM_{10} conc.**	**Comments**
1 day	EU limit value: 50 µg/m^3	max exceedance of 35 days per year
	WHO AQG: 50 µg/m^3	9th percentile (3 days per year)
Calendar year	Limit value: 40 µg/m^3	
	WHO AQG: 20 µg/m^3	

Coal and biomass combustion in households, commercial, and institutional buildings are reported to be the most important contributors to total PM emissions in the EU. A study of PM_{10} concentrations showed in 2014 the PM_{10} daily threshold was widely exceeded in Bulgaria, Italy, Poland, Slovakia and the Balkan region as well as a number of urban areas across Europe, including in the Nordic countries. About 95% of the these exceedances happened in urban or suburban areas (Chlebowska-Styś et al., 2017). According to Guardian (2017) for the first time in more than 130 years, Britain powered itself without coal for an entire day in April 2017. According to EEA (2020), 20 Member States and six other reporting countries reported PM_{10} concentrations above the EU daily threshold during 2018 (Figure 2.7).

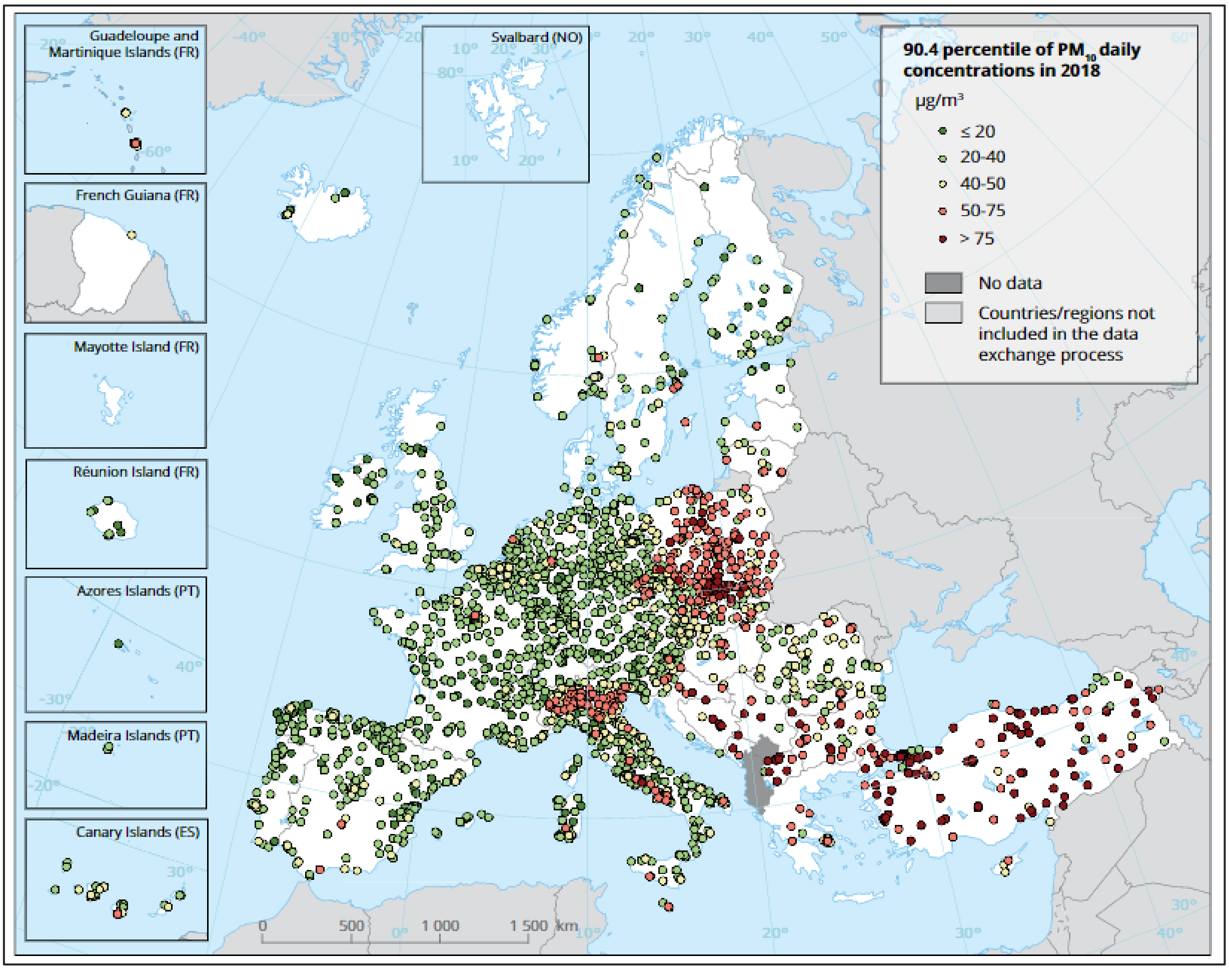
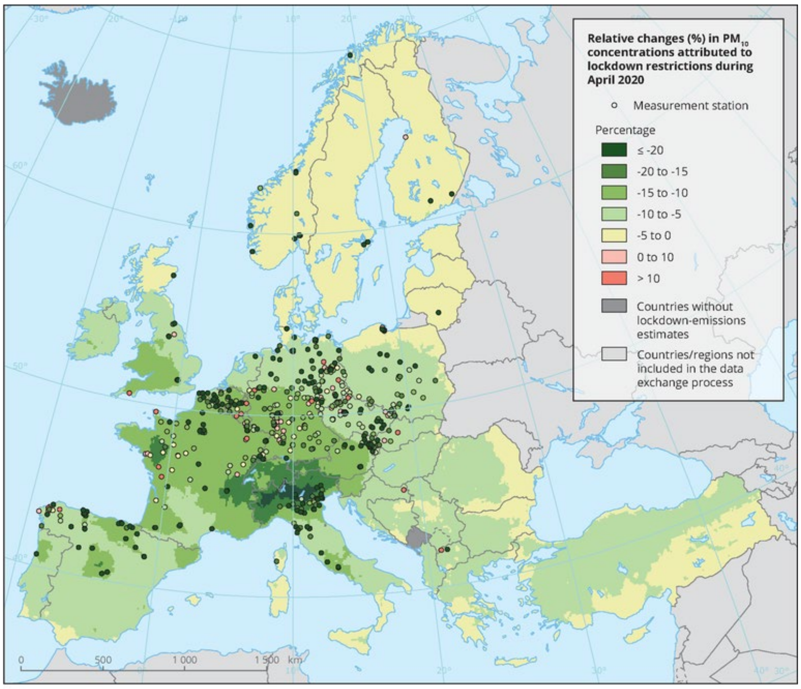


Figure 2.7: Concentrations of PM_{10}, 2018- daily limit value (EEA, 2020).

In 2020 PM_{10} concentrations across Europe has changed (Figure 2.8) due to COVID-19 lockdowns (EEA, 2020).


Relative changes (%) in PM_{10} concentrations attributed to lockdown restrictions during April 2020
○ Measurement station
Percentage
≤ -20
-20 to -15
-15 to -10
-10 to -5
-5 to 0
0 to 10
> 10
Countries without lockdown-emissions estimates
Countries/regions not included in the data exchange process
0
500
1 000
1 500 km

Reference data: ©ESRI

Figure 2.8: Relative changes in PM_{10} concentration across Europe due to COVID-19 lockdown conditions (EEA, 2020).

In US, the National Emissions Inventory (NEI)'s PM data, covering all 50 states and their counties for the 2010 to 2019 period, was analyzed. The results showed that estimated primary PM_{10} emissions from anthropogenetic sources had reduced 17% at a national level between 2010 and 2019. In 2014, it was estimated that primary PM_{10} emissions from miscellaneous and natural sources and fugitive dust accounted for 87% of total primary PM_{10} emissions (EPA, 2018).

2.4 Air Particulate Matter Research in New Zealand

In New Zealand, the dominant and largely uncontaminated westerly winds, disperse air pollutants before they can become too concentrated. During calmer wintertime conditions, however, poor air quality can become a major concern in some cities (MfE, 2009).

The historical monitoring of particles in New Zealand, was based on total suspended particulate (TSP) measurements, which involves all suspended particles in the air (MfE, 2002). During the 1990s monitoring methods were launched to capture the PM_{10} size fraction. A smoke monitoring methodology was also used as an alternative in most areas but was overlooked by late 1990s as it was biased towards elemental carbon measurements.

New Zealand studies have shown that the relative contribution of different sources to the total PM_{10} mass differs between sites. In Hastings (one of the two major urban areas in Hawke's Bay, North Island) domestic home heating, marine aerosol, motor vehicles, sulphate, and soil were found to contribute to the PM_{10} concentrations (Wilton, Appelhans, Baynes, & Zawar Reza, 2009). Unusually, in their report the authors included outdoor burning of domestic waste in domestic heating sources. In more rural areas of New Zealand, the burning of domestic waste is fairly common. In Hastings, biomass burning, and domestic heating were reported to be the main contributors to PM_{10} concentrations. The concentrations of PM_{10} were notably higher during the colder months of April–October when household hearing is required. The total background or "natural" PM_{10} contribution (soil and sea spray) for Hastings during winter was estimated to be in the range of 13-15% of the total PM_{10}.

In Christchurch, the largest city in Canterbury (population in 2015 was 367,800) , domestic fuel burning and temperature inversion frequently contribute to high wintertime smoke levels (Salomon & Smithson, 2015). In the larger Canterbury region, spatial variations of PM_{10} concentrations were studied in the small towns of Rangiora (population 19,250) and Kaiapoi (population 11,847) (Hamilton et al., 2004). The study found that older residential neighborhoods in both towns had extremely elevated PM_{10} concentrations. During settled anticyclonic conditions, drainage movements during the night carried PM_{10} plumes from Rangiora to Kaiapoi, triggering rises in PM_{10} concentrations in Kaiapoi. Alexandra, also in the South Island, located in an inland basin in central Otago faces extremely poor air quality during the winter. A stable boundary layer forms in high pollution days for 18 hours from 17:00 to 11:00 during the wintertime.

Rotorua in the central north island also experiences high concentrations of PM_{10} that regularly exceed the National Environmental Standards for Air Quality (NESAQ) (Fisher et al., 2008). The high PM_{10} concentration in Rotorua is attributed to the use of domestic wood burners (60%) during wintertime (BOPRC, 2014).

Davy (2007) studied PM pollution sources in different areas within the Greater Wellington region using multi-element analyses. In Upper Hutt, road dust, soil, and sea salt (coarse) and sulphate, motor vehicles, wood burning, and sea salt (fine) sources were identified as PM sources. The dominant source during winter was biomass burning closely followed by motor vehicle emission. During the summer, secondary sulphate was found to be a significant contributor to the fine particle mass. In the industrial area of Seaview, sea salt, road dust, soil, and zinc (coarse) and sulphate, soil (fine) sources were identified. A local galvanizing operation in the region contributed to the zinc sources. The most significant contributor to fine particle concentrations was motor vehicle emissions.

A source apportionment study from anthropogenic activities was conducted for towns of Napier (a coastal city in the Hawkes Bay region of the North Island). Hastings and Havelock (suburbs of Hastings) were the locations that the PM_{10} data was collected from for 2010 and 2015 (Wilton, 2015). For all three towns, the main source of PM_{10} particulates was domestic heating ranging from 88% in Napier to 98%

in Havelock North. In Napier and Hastings, the other main contributions were from transport and shipping was and industrial activities. The second contributor to PM_{10} in Havelock North was transport which was accounted for only 2% of the PM_{10}.

PM_{10} elemental composition of samples collected in the three largest (by population) New Zealand cities of Christchurch, Hamilton, and Auckland for the winters of 2000 and 2001 were inspected (Senaratne, 2003). Sea spray, suspended soil/road dust, domestic emissions, diesel, and petrol emissions were identified as the main sources of PM_{10} mass. Marine and crustal elements were dominant in the total elements mass across all four seasons. However, Elemental Carbon (EC) dominated all elements in the winter. In winter, similar contributions from both domestic and vehicle emissions were found in Christchurch. In contrast, in Hamilton vehicle emissions had a greater contribution when compared to domestic emissions even in winter. During winter, in Auckland suspended soil, road dust, and vehicle emissions had larger contributions than domestic emissions.

Different conclusions have been drawn as to the composition of organic pollutants during the winter in Auckland (Krivácsy, Blazsó, and Shooter, 2006; Wang, Kawamura, and Shooter, 2006). One hundred organic compounds from PM_{10} samples were collected in 2001 and characterized by Krivácsy et al. (2006). They found that organic species accounted for 21–45% of PM_{10} during winter. They suggested that the most dominant source of carbonaceous PM was domestic wood combustion. Dehydroabietic acid, a tracer for biomass burning (Simoneit, 2002), was found to be the most abundant organic compound. The authors did not detect hopanes or steranes, indicators of vehicular emissions (Cass, 1998). During the winter of 2004, 69 organic species in PM_{10} were collected and characterized (Wang, Kawamura, Shooter, 2006). Wang et al. found that both tracers for biomass burning plumes (levoglucosan and dehydroabietic acid) were the richest compounds of their PM_{10} organic species. Low concentrations of Hopanes were detected (0.02–0.05% of the total organic carbon (TOC)). Although a great abundance of biomass burning tracers were found, the authors decided that the most significant contributors to PM_{10} concentration in Auckland were motor vehicles.

Studies of PM_{10} source apportionment have also been conducted in Nelson, South Island (Ancelet et al., 2014) and in Tokoroa, the fifth-largest town in the Waikato region of the North Island (Ancelet & Davy, 2014). Figure 2.9 summarises the results of these source apportionment analyses in wintertime from urban areas around New Zealand. It can be observed that in both the North Island and the South Island, biomass burning (wood-burners) are identified as the dominant wintertime source for PM_{10} ranging from 50% in Auckland to 89% in Alexandra (Ancelet & Davy, 2014).

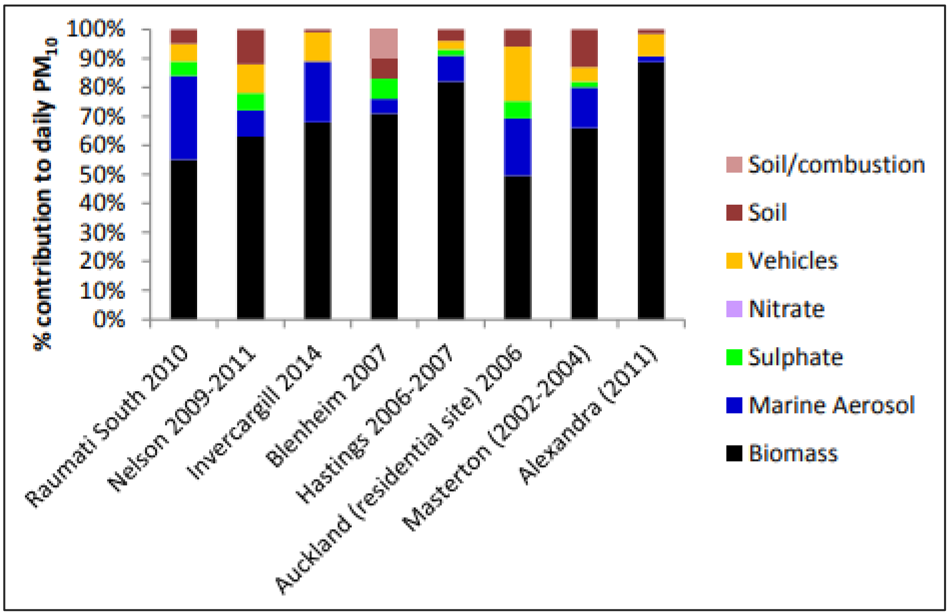


Figure 2.9: Comparison of source apportionment analyses from urban areas around New Zealand (Ancelet & Davy, 2014).

The concentrations of different carbonaceous and ionic components in PM_{10} during the winter in Auckland and Christchurch was studied using principal components analysis (PCA). The most significant contributor to Auckland's PM_{10} concentrations was marine aerosol followed by traffic emissions. Ambient concentrations of carbonaceous materials in Christchurch was found to be significantly higher compared to Auckland due to Christchurch's residential use of wood and coal burners in winter (Wang et al., 2005). Adverse effects from air quality can be also aggravated by land use. In New Zealand, the population is very unevenly distributed therefore the pressures from land use is more likely to be felt regionally than nationally. In New Zealand 75% of population live in the North Island (half in upper North Island), and the remaining 25% live in the South Island. Population increase has been particularly intense in the Auckland area, creating concerns on man-made contribution toward air pollutants.

2.4.1 Auckland Regional Air Quality

Some areas in Auckland are subject to the formation of localized micro-climates that can raise PM_{10} concentrations significantly for few days (Ancelet et al., 2012). For Auckland water-soluble inorganic species were characterized in three separate studies (Wang & Shooter, 2001, 2002, 2005). Wang and Shooter (2001) characterized eight soluble ions in PM_{10} and found significant seasonal variations in Na, C, and SO_4 concentrations with highest concentrations happening during the summer due to an increased marine influence. In a study of fine/coarse and day/night variations for eight water soluble ions during Auckland's winter, Wang and Shooter (2002) found that PM_{10} concentrations did not show a noticeable day/night change, but SO_4 and NO_3 concentrations were elevated during the day and night, respectively. Sea salt ions were found in coarse samples whereas fine samples were enriched with non-sea salt ions. The primary marine aerosol generation and source regions identified by (Davy et al., 2011) were below Australia in the Southern Ocean and to the northeast of Auckland in the Pacific Ocean. Figure 2.11 shows source apportionment analyses of PM_{10} at the Takapuna monitoring site located in Auckland, New Zealand. The Takapuna site is considered by Auckland Council to be representative of the entire Auckland airshed (PCE, 2015).

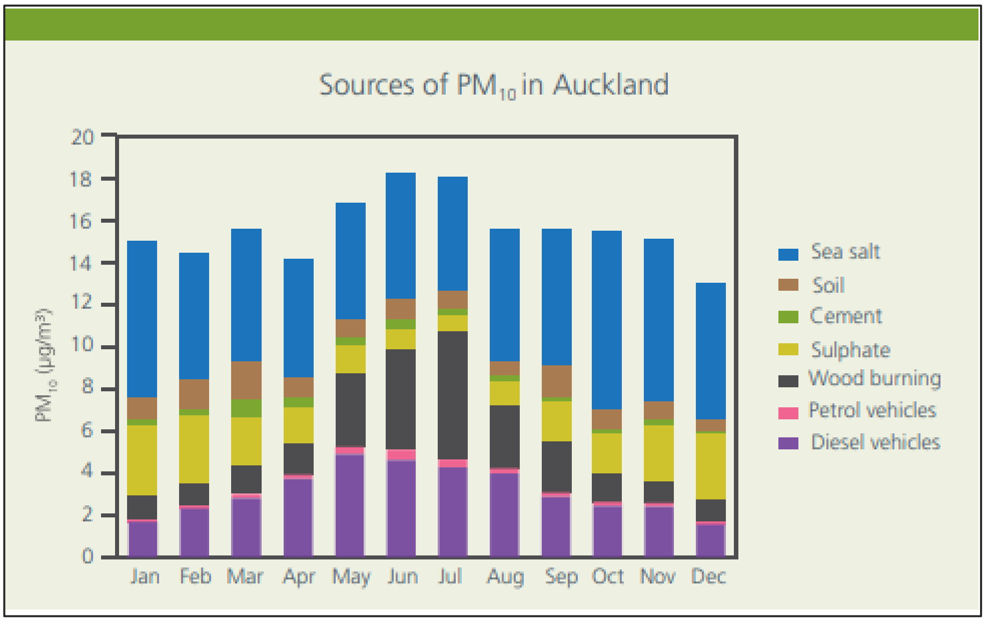


Figure 2.10: Source apportionment of PM_{10}, Takapuna, Auckland from samples taken every third day between 2006 and 2013(PCE, 2015) with permission.

A recent study by Davy et al. (2017) found that sea salt, motor vehicle exhaust emissions, residential wood burning, and crustal matter are the key contributing influences on ambient PM_{10} concentrations in Auckland (Figure 2.12). The marine aerosol component of urban air PM is part of the 'natural' background.

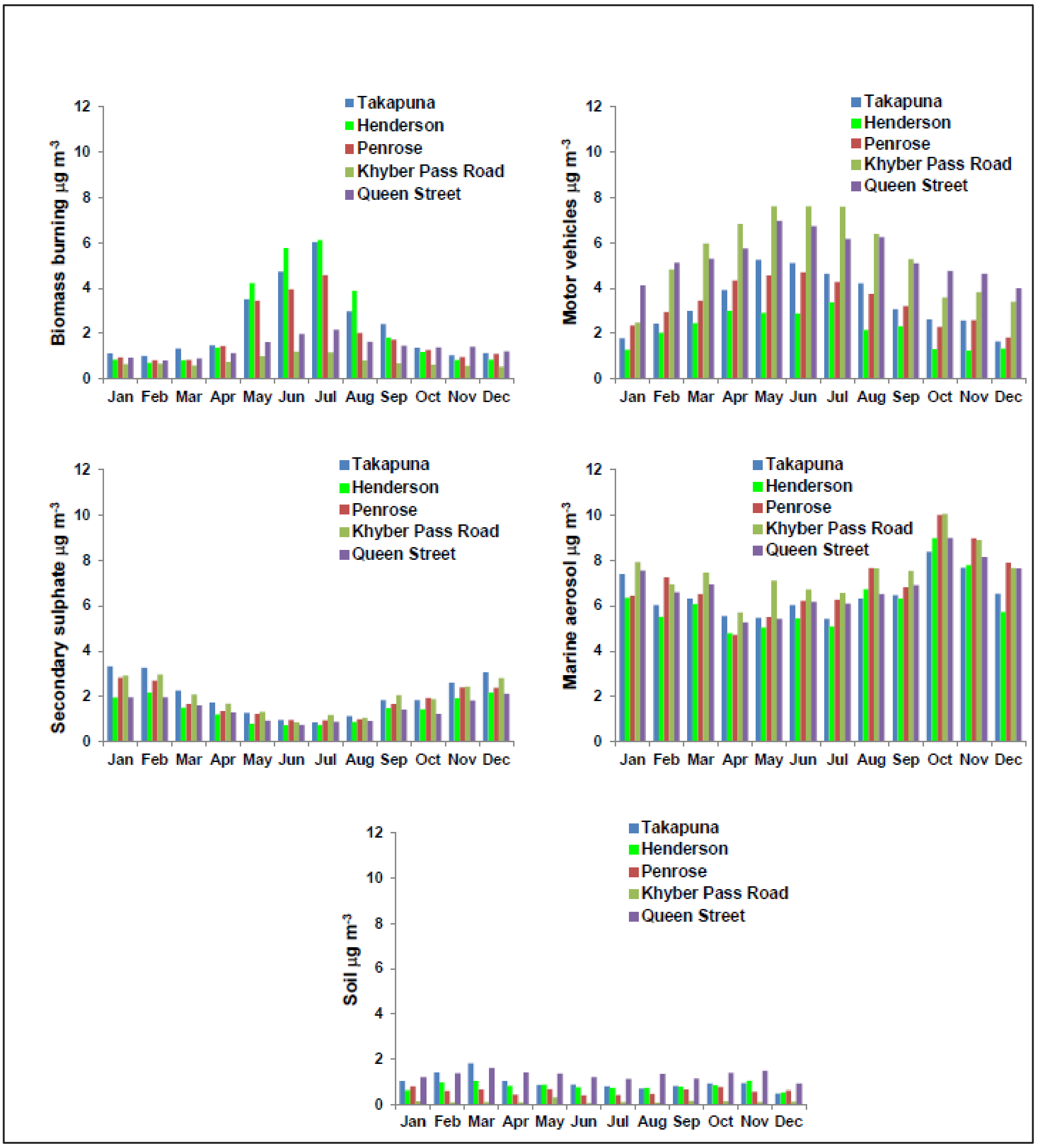


Figure 2.11: Source apportionment results on monthly average PM_{10} for Auckland monitoring sites.

According to Davy et al. (2017), marine aerosol concentrations showed a statistically significant, decreasing trend from 2005 to 2015 across all sites in Auckland. The decreased trend could be either a permanent trend or be partially related to a larger inter-decadal cycle of Southern Hemisphere circulation climate patterns. Marine aerosol and motor vehicle emissions were the main sources of PM_{10} in Auckland (Figure 2.14) while 72% of the wintertime increment was due to home heating. Analysis of crustal matter contributions in PM_{10} concentration at high density traffic sites in Auckland showed that road dust may be a major contributor. Other sources were windblown soil, earthwork, and construction dust. The PM_{10} crustal matter source contributions trend analysis showed that concentrations decreased over the monitoring period and was mostly site dependent. Temporal variations of crustal matter contributions showed lower concentrations during weekend suggesting the influence of human activities on source of emissions. Sea salt contributed about 60% to Auckland's PM_{10} concentration over a year (Davy et al., 2017).

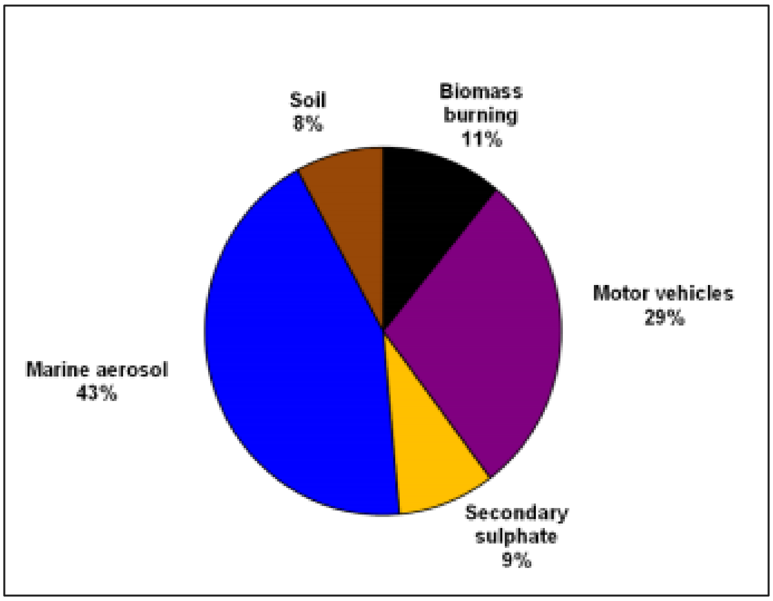


Figure 2.12: Source contribution for PM_{10} particulate between 2006-2015 (Davy et al., 2017)

The 2016 Auckland Air Emissions Inventory estimated the relative contributions from motor vehicles in PM_{10} concentration. The details are shown in Figure 2.13.

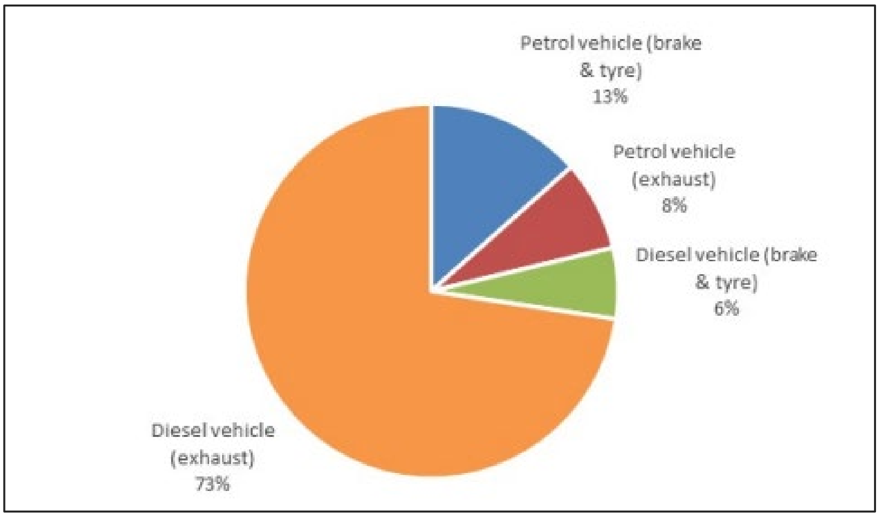


Figure 2.13: Estimated annual PM_{10} emissions for 2016 by emission type (Sridhar & Metcalfe, 2018).

During winter, domestic wood burners combined with poor meteorological dispersion conditions contribute to poor air quality, in urban areas. According to Auckland Council (2015) emissions from home heating fires, which are only lightly regulated, are the largest proportion of wintertime particulate emissions in most New Zealand regions including Auckland.

According to NIWA, who performed an analysis of PM_{10} and NO_2 level during COVID-19 lockdown in Auckland showed over the five weeks at Level 4 (26th March 2020 – 27th April 2020 inclusive), PM_{10} concentrations from traffic, heating, industrial and natural sources was reduced at all Auckland sites. This included a reduction of 14% at Queen St, 15% in Penrose, 28% in Takapuna, and 9% in Henderson (Figure 2.14). In the third week of lockdown, all sites except for Queen St, showed greater decreases in PM_{10} and NO_2 concentrations when compared to weeks one and two (Longley, 2020). Levels of dust, sea spray and smoke, were only marginally down across Auckland suggesting the emission sources are other than traffic.

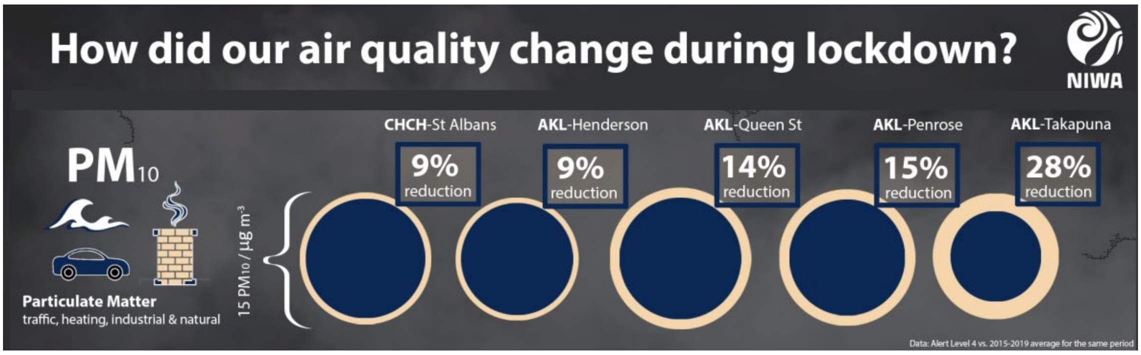


Figure 2.14: Covid-19 lockdown under level 4 restrictions and its effect on air quality changes in Auckland 2020, (Longley, 2020) with permission.

Another study showed 20.1% decrease in PM_{10} concentrations at the Auckland central, 16.2% at a sub-urban site of Henderson and 6.6% at the rural background site of Patumahoe during the first COVID-19 lockdown period (Patel et al., 2020).

For most Auckland sites, a peak in concentrations is observed during the morning (07:00 and 09:00) ‘rush hour’. PM_{10} concentrations in the afternoon are generally lower when compared to the morning due to increased mixing in the atmosphere. In the evening, a more stable atmosphere causes less dispersion hence an increase in concentrations. This natural process is worsened during winter with the increase in home heating emissions (Patel et al., 2020).

Regional air quality targets are set to control ambient air quality pollutants in Auckland. This includes those that are not covered within the National Environmental Standards for Ambient Air Quality (AQNES) including ambient pollutants or averaging intervals. Although a gradual decrease has been observed in long-term trends, short-term air quality issues in Auckland are not yet fully. Auckland Region’s NO_2, PM_{10} and $PM_{2.5}$ emissions from domestic fires and mobile sources need to reduce significantly if the area is to meet the target of zero breaches of the standards in residential areas, workplaces, playing areas (Auckland Council, 2010).

2.4.2 New Zealand Guidelines and Regulations

Development and application of guideline values is an iterative process due to the rapid growth of studies on air quality and the ongoing research findings on pollutants health effects. Air quality guidelines and their application are reviewed and updated by the Ministry for the Environment (MfE) at least every five years. The MfE cooperates with councils to improve air quality management. Most of the guideline values adopted in New Zealand have been taken from guidance provided by overseas organizations such as WHO (2006). Since 1994 local authorities have been operating under an air quality guideline of 120µg/m^3 for 24 hours (Ministry for the Environment, 1994a). The national ambient air quality guidelines were last updated in 2002. In May 2002, the strength of the medical evidence of the acute health effects of PM$_{10}$ persuaded the MfE to recommend a new air quality guideline for PM$_{10}$. The annual average value was thus amended to reflect the chronic health effects of PM$_{10}$. As the result the new PM$_{10}$ guideline values changed to 50 µg/m^3 (24-hour average) and 20µg/m^3 (annual average). In addition to the ambient air quality guidelines, New Zealand has the National Environmental Standards for air quality (Air Quality NES) which was introduced in 2004 to assure a minimum level of health protection for New Zealanders. This was due to increasing level of PM$_{10}$ in most parts of the country during winter. NES regulations are made under the Resource Management Act 1991 and came into effect on 8 October 2004. These air quality legislation, regulations, and guidelines are summarised in Table 2.3 (MfE & Stats NZ, 2018).

Table 2.3: Air quality legislation, regulations, and guidelines (MfE & Stats NZ, 2018).

	Guidelines	**Legislation and regulations**
International	World Health Organization air quality guidelines (global update 2005)	
New Zealand	Ambient Air Quality Guidelines 1994 (Last updated 2002)	Resource Management Act 1991 National Environmental Standards (NES) for Air Quality 2004 (last updated 2011) Regional air plans (required under the Resource Management Act) (regional councils)

NES is based on human health only and includes only five priority pollutants whereas ambient guidelines support both ecosystems and human health and cover wide variety of pollutants, including toxins. The ambient air quality standards permissible exceedance for annual average PM$_{10}$ is zero. Since there is no NES for annual PM$_{10}$ concentrations the same national ambient air quality guideline applies.

With the 2004 Regulations coming to force it was expected that all airsheds would meet the PM_{10} standard by 2013. However, by late 2009, the MfE estimated that the air quality standards in 15 airsheds would not comply in time. The 2013 deadline was unachievable for the Auckland airshed that accommodated at the time nearly 30% of New Zealand's population. Therefore, in 2009, the MfE reviewed the PM_{10} regulations addressing concerns regarding the perceived 'stringency' of the ambient standard, the absence of equity for industrial air pollution sources, and the unachievable original target timeline of 2013. On the 1st of June 2011 the revised regulations came into force. The 2011 amendments renamed these regulations to be the Resource Management (Air Quality NES) Regulations 2004 and changed the timeframes to be in line with those of the ambient PM_{10} standard. The Air Quality NES includes staggered transitional periods and exceedances periods. When the transitional periods ended (in September 2020) the number of allowed exceedances will decrease to one exceedance per 12 months (MfE, 2011a). The new allowances for meeting PM_{10} standard are provided in Table 2.4.

Table 2.4: Allowances for meeting PM_{10} standard (MfE, 2011a).

Average number of exceedances per year (before start date)	Transitional period	Number of exceedances allowed
1 or fewer	Always	1 or fewer in 12 months
2-9	1 September 2011 to 31 August 2016	Unlimited
	1 September 2016 onwards	1 exceedance per 12 months
10+	1 September 2011 to 31 August 2016	Unlimited
	1 September 2016 to 31 August 2020	3 exceedances per 12 months
	1 September 2020 onwards	1 exceedance per 12 months

Between 2006 and 2016, only 45 out of 93 monitoring sites had valid data for PM_{10} exposure. For 2006-2016, 38 of 45 monitoring sites exceeded the national 24-hour average (Haenfling, 2020).

2.5 Conclusion

Ambient air quality denotes the near ground level outdoor air quality, that is away from direct sources of pollution. High concentrations of pollutants lead to poor ambient air quality which affects human health and/or the environment. PM_{10} refers to particles with a diameter of less than 10 µg and is the major air pollutant monitored in New Zealand. These particles are derived primarily through suspension of dust and soil and other materials from roads, farming, construction, or mining activities, and combustion of coal. Other sources of PM_{10} include sea salts, dust, and secondary organic carbon results from the VOCs. Different PM_{10} measurement methods and instruments are used for measuring PM_{10}

concentration and size. In New Zealand, the recommended monitoring methods to establish compliance with the 2002 AAGQ were reviewed, and the US 40 CFR Part 50 is now recommended as PM_{10} monitoring method.

This Chapter also addressed air quality management in New Zealand. New Zealand is known for its clean and green credentials. Discharges to air such as products of combustion and particulate matter can be complex in nature and have the potential to cause adverse effects on ambient air quality and human health. In 2004, 14 national environmental standards relating to air quality using WHO guidelines were introduced to help reduce the negative effects of poor air quality. The primary purpose of the national ambient air quality standards is to set minimum requirements for outdoor air quality in order to provide a guaranteed level of protection for the health of all New Zealanders. In 2020, amendments to some provisions of the NES were proposed to better control the release of fine particles into our air. These national regulations place a requirement on regional councils to monitor air quality and to report ambient air quality exceedances under the Resource Management Act. The exceedances occur if annual average concentrations are greater than 20µg/m3) and number of monitored sites that exceeded the national environmental daily (24-hour) average standard for PM_{10} (exceedances occur when daily average concentrations are greater than 50µg/m3. From 1 September 2020 onwards one exceedance per 12-month period is allowed.

While the airsheds are established in locations with high likelihood of exceedance of standards, the quality assurance of monitoring sites is seeming to be neglected. The costs of maintaining already installed monitoring sites could be a reason for poor data quality but should be reconsidered as the effects of PM_{10} on human health is the main concern. This cost could be negligible as a cost of over $2 billion to the Auckland region in 2006 was reported as a result of exposure to ambient $PM_{10.}$

Chapter 3 ANALYSIS OF PM_{10} AND METEOROLOGICAL DATA

This Chapter provides an explanatory analysis of the PM_{10} concentrations and meteorological data used in this research. Descriptive statistical parameters and box plots of the data are constructed to illustrate the differences and similarities between the study sites. The Kruskal-Wallis (K-W) test is applied to show the statistical significance of changes over time for each site. The general characteristics of local PM_{10} concentration are investigated by comparing the daily average PM_{10} data within the sites. An assessment of trends in PM_{10} concentrations is carried out considering the daily variations in meteorological conditions and their consequent impact on PM_{10} concentrations. The "openair" package (Team 2011, Carslaw 2012, Carslaw and Ropkins 2012) for 'R' statistical software© was used in all statistical analyses presented in this Chapter.

Air quality data were supplied by Auckland City Council (AC). AC is the authority responsible for air quality monitoring of the Auckland airshed. A study by Aberkane et al. (2005) found that 90% of all particles measured as PM_{10} are made up of ($PM_{2.5}$) during winter. Therefore $PM_{2.5}$ was not included as a predictor to eliminate correlation in the dataset.

3.1 Study Area and Data

Air quality data usually includes pollutant concentrations and meteorological data of urban ambient air at a certain time. AC measure air quality in several stationary locations and with mobile stations. The AC network for continuous monitoring of PM_{10} is comprised of 13 permanent and two mobile sites. The PM_{10} monitoring network extends from Patumahoe in the south to the Whangaparaoa Peninsula in the north and from Glen Eden in the west to Botany Downs in the east. The locations of these PM_{10} monitoring stations are presented in Figure 3.1.

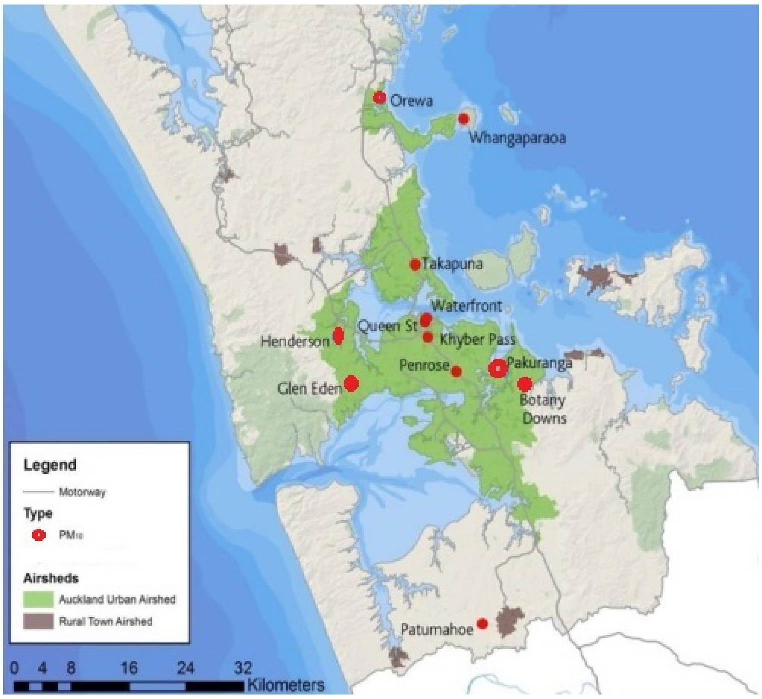


Figure 3.1: Auckland air quality monitoring network (AC, 2006).

All stations use Beta Attenuation Monitors (BAM) instruments for measuring the concentration of PM in the lower atmosphere. AC provided the air quality data set in 1-hour and 24-hour averages. The size of the dataset is site specific with notable differences between each site. Among the 13 stations, some provide less than four seasons of data, which is insufficient for training a temporal prediction model. In addition to the short period of some of the site's time series, missing data is also a problem. Even if the monitoring stations are available, system maintenance and incidental events can cause missing data.

This research employed the most current data available from AC, at the time of writing, and covers six full years of data from 2011-2016 inclusive. Because of the variation in data quality from the various sample sites some sites in the AC were necessarily excluded. In mid-2014, three long-term monitoring sites at Botany Downs, Orewa and Whangaparaoa were decommissioned. The loss of these long-term sites is regrettable, as it represents the closure of an extended dataset, which was invaluable for the observation and analysis of long-term trends in ambient air quality. For this research, these sites decommissioned in 2014 were not included because of a lack of more recent data 2014-2016. Auckland Waterfront was also excluded as the sampling station was mobile and the location of the station changed during this research's sampling period.

The remaining six stations used in this study are Takapuna, Henderson, Glen Eden, Penrose, Pakuranga and Patumahoe. In a region with heterogeneous land use, the spatial positioning of monitoring stations is important. The PM_{10} dataset used in this study has the geographical characteristics of urban residential, urban industrial, urban traffic/residential and rural residential. These classes are the ones determined and reported by AC (AC, 2006).

Glen Eden site is categorised by AC as existing in an urban background. The site is an air-conditioned shed at SE corner of a park and20m from the closest road. Most houses in the area were built in the 1980s and rely on electric heating. However, in the north of Glen Eden there are many older houses (built in the 1960s) and about 75% of these houses still have the original wood fire heating. The site is surrounded by hills to northeast which tend to influence wind flows from this direction.

Henderson is a suburb 13 kilometers west of Auckland's city center. The monitoring site is approximately 10m from the western side of Lincoln Road. Lincoln Road is one of Auckland's main arterial roads. Traffic volume is high, and the road suffers from congestion with an estimated 46,000 cars a day travelling along it in 2016. Along with this traffic has come significant development in the area. In a source appointment study marine aerosol was found to be the main source mass contribution to PM_{10} at this site. Biomass burning and motor vehicles were identified as the predominant sources of pollutants during peak PM_{10} events and during winter. Houses in area were first built in the 1960's and while building has continued; approximately 50% of homes still have chimneys and wood burning heating. In Henderson, there is a mixture of residential and commercial activities in terms of land use, with some industrial activities 500m to the northeast. Waitakere Hospital is located within 300m southeast of the site (Davy et al.,2017).

Pakuranga is a southeastern suburb of Auckland. The station is sited at the south west corner of Bell Reserve; approximately 7.5m southeast of Pakuranga Highway. Pakuranga Highway is a major arterial route. Vehicle and residential home heating are the predominant sources of pollution. Houses in the area are of mixed ages; the earliest from the 1960s.

Patumahoe village is located at the southern edge of the Auckland Region. The air-conditioned station is located approximately 2.5km west of the Pukekohe urban area. There are greenhouses and sheds 8m to the north and 20m to the west and southwest. Hedges surround the site on three sides: a 4m hedge 30m to the south; an 8m hedge 40m to the east; and an 8m hedge 50m to the north. The surrounding area is used for horticulture and agriculture. The site is categorized as a rural background site.

Penrose site is an industrial station in south-central Auckland with a different source of pollutant. The site is highly impacted by traffic as it is located only 50m northeast of the Auckland Southern Motorway. The motorway is approximately two meters lower than the ground level at the monitoring site and is the main route through Auckland to the rest of the North Island. There are main roads within a range of 500m to 1km north of the site. There are no main roads within 1km to the east northeast. Industrial premises and a glass manufacturer are located from northwest-south to the northeast.

Residential houses dated from the 1930s onward are situated to the north and southwest (Davy et al., 2017).

Takapuna is a central-northern suburb of Auckland with mixture of residential and commercial land use. The coastline is complex and topography in this area is low-lying undulating. Therefore, the surface wind flow has a complex pattern, mainly during low synoptic wind flows conditions when sea-land breezes dominate the surface winds. The site is subjected to winds from all directions. As shown in the wind rose diagram in Figure 3.28 constructed based on the 2011-2016 data, the west and northeast winds are the dominant wind directions. The residential houses vary in ages spanning back to the 1960s, and 75% of the houses use fossil fuel for heating during the colder winter months. A concrete batching plant producing ready-mix concrete is located 100m southeast of the site. The site is located 3km east of the coastline of the Hauraki Gulf. There is a commercial centre in nearly 3.5km southeast. The site is 50m west of State Highway 1, the main motorway from the north into Auckland City, connecting the Southern and Northern suburbs.

3.2 PM_{10} Data collection and processing

The 24-hour averages of PM_{10} provided by AC is based on the available hourly measurements within that day (midnight to midnight). A few standards for determining the minimum amount of data collected from a station to estimate the average value of air pollutant concentration is provided by WHO (Tiwary, 2010):

- 1-hour average value must have at least 75% of the monitoring data.
- The 24-hour average value must have a minimum of 50% of hourly data in a day.
- The seasonal and yearly average values must have minimum 50% of daily data in a year.

To ensure compliance with WHO regulations, 1-hour data was checked to against the WHO criteria to determine whether they could be used to estimate average PM_{10} concentrations. Daily average concentration was calculated for a site only if the hourly concentration data in one day was available for at least 50% of the day (i.e., 12 hours in 24 hours). Any days where there was less than 50% of the data were excluded. A particularly extreme example of this occurred on the 18th of January 2012, when only four Takapuna PM_{10} measurements from 1 a.m. onwards were available. Even if the WHO 50% threshold is met, calculating daily average from midnight to midnight can cause overestimation of daily average PM_{10} concentration when only the evening data are available. The daily average maybe underestimated if only early morning/afternoon measurements are available. Another extreme pattern occurred in May 2012, where PM_{10} measurements for Glen Eden where recorded only from 1 a.m. to 10 a.m. Days with such missing patterns were identified and excluded from the averaging procedure. Several negative PM_{10} measurements (ranging from -6.00 $\mu g/m^3$ to -0.1 $\mu g/m^3$) were encountered during data cleaning. Negative values are an indication that the actual PM_{10} concentration is below the detection limit of the BAM 1020 (Met One Instruments, 2016). If the monitoring device' background offset is adjusted correctly during the initial setup, then the hourly concentration below -4 µg are statistically unlikely (Met One Instruments, 2016). In this thesis these negative values were treated as missing values when processing the raw data. In the original AC data file however, these negative values were included in the calculation of daily average PM_{10}, and where this occurred it led to underestimation of the daily averages.

Apart from two days with remarkably high readings, BAM standard range is 0 - 1000 µg/m^3 (Met One Instruments, 2016), no obviously unusual reading were encountered. Notable events resulted in PM_{10} exceedances was identified at Patumahoe during February 2013, the influence of which can be seen in the time-series plots of PM_{10} concentrations presented in Figure 3.2. The PM_{10} level increased from 20.01 (µg/m^3) at 2:00 a.m. to 404.50 (µg/m^3) at 4:00 a.m. The concentration reached its highest of 3056 (µg/m^3) at 6:00 a.m. and then decreased to 44.35(µg/m^3) at 8:00 a.m. An online search by the author to find the official reports for atypical influences (such as local burning near monitors, bonfires, and fireworks) did not provide any information to identify a possible cause for these unusual readings. These later measurements were considered in the explanatory data analysis presented in this thesis but were treated as outliers in the modeling approaches and models. Two relatively continues high measurements (449 µg/m^3 and 329 µg/m$^{3)}$ were observed at the Takapuna site in March 2011. During 2011 the sports fields around the Takapuna monitoring station were substantially redeveloped and this may be the cause of these hourly outliers. Davy (2017) noted the effect of these field development activities in the source contribution data when reporting on a source apportionment study.

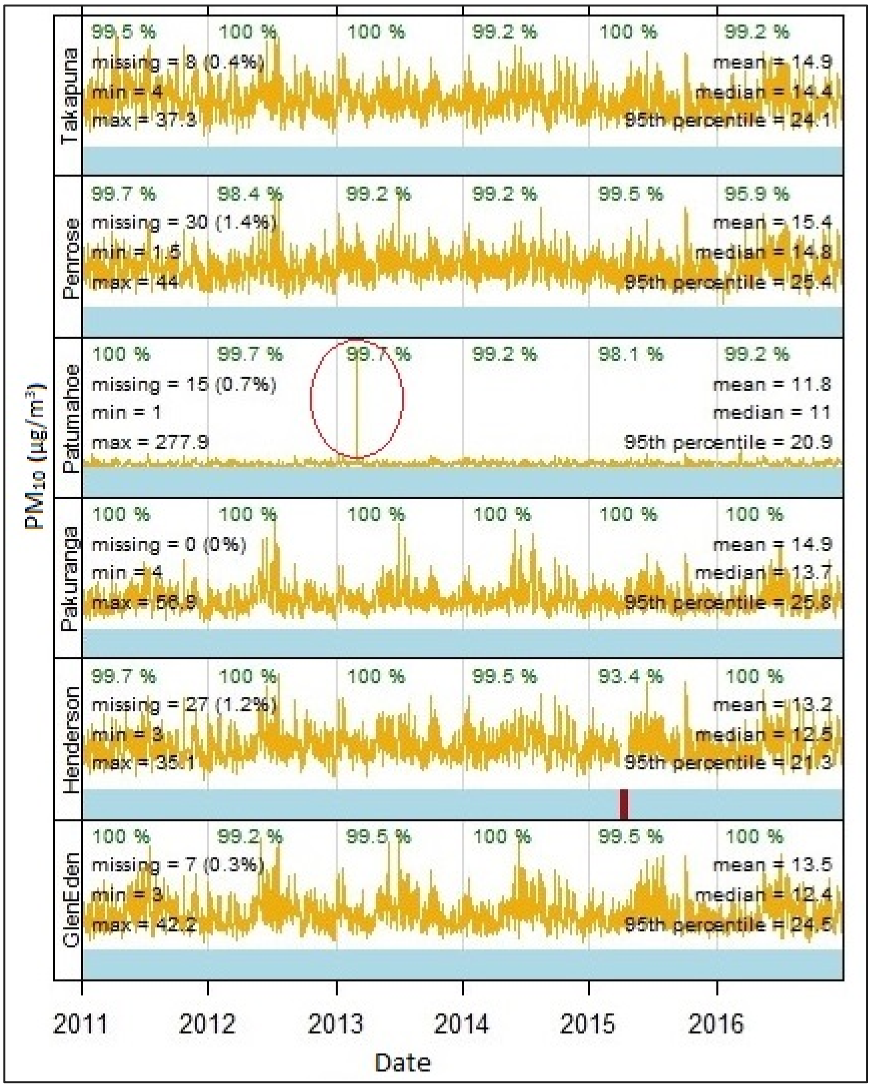


Figure 3.2: Final description of processed PM$_{10}$ datasets. Daily average PM$_{10}$ concentration in µg/m3 by year for each site. Notable exceedance at Patumahoe highlighted (red circle). Red bar in Henderson plot indicates a period of missing data.

3.3 Exploratory Data Analysis

PM$_{10}$ Concentrations

Site-specific summary statistics and temporal variation of PM$_{10}$ concentration over the period of 2011-2016 are shown in Figure 3.3. Each graph shows the median, 25th and 75th percentiles, concentration ranges within two standard deviations (indicated by whiskers), and extreme values (indicated by circles). Peak PM$_{10}$ events are classified as those that were higher than 60% (30 µg/m^3) of the National Environmental Standards (NES) 24-hour average.

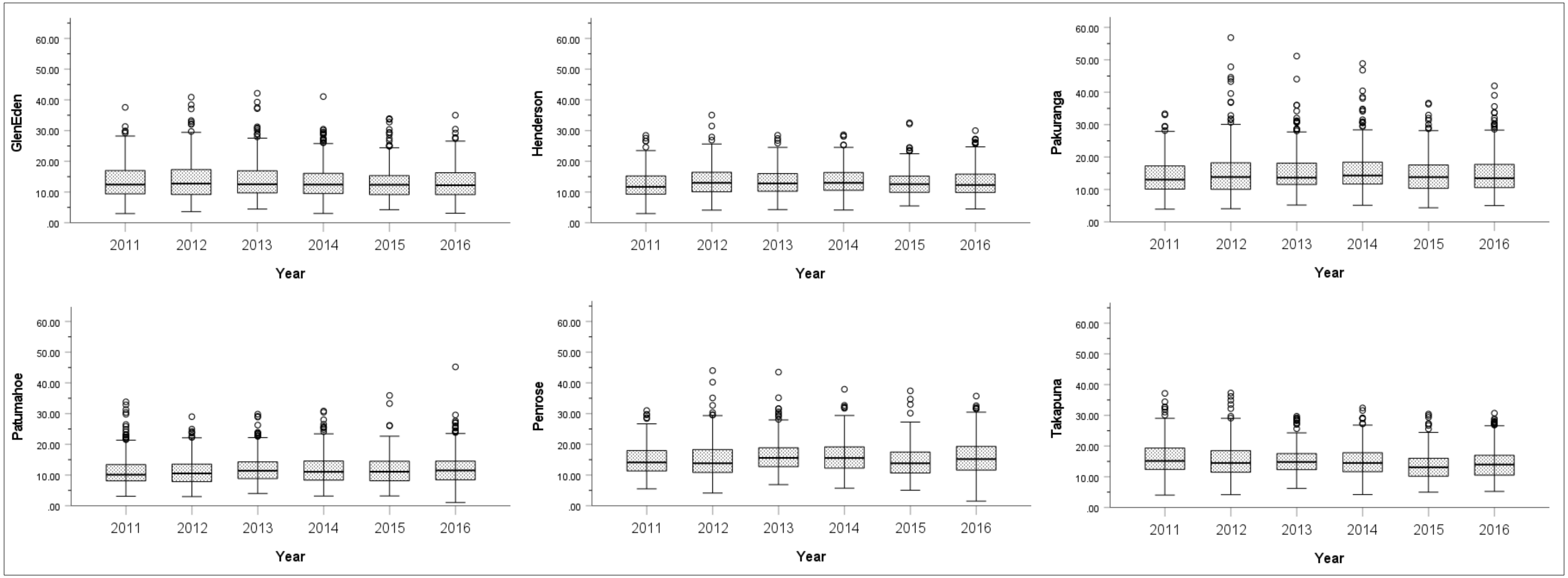


Figure 3.3: Annual distribution of 24-hour average PM_{10} (μg/m^3) concentration (2011-2016).

Within the study area daily 24-h average PM_{10} ranged from 1.05 µg/m^3 to 277.85 µg/m^3 (Patumahoe in the year 2013). The mean and standard deviation of PM_{10} concentrations were overall higher in Penrose. During the study period, in the year 2013, a highest PM_{10} daily average of 277.85 µg/m^3 was observed at the rural area of Patumahoe. The second highest daily average of 59.86 µg/m^3 was observed at Pakuranga during 2012. Descriptive statistics for PM_{10} concentration are summarized in Table 3.1.

Table 3.1: Descriptive statistics for daily average PM_{10} concentration in µg/m^3 within the study area, (2011-2016).

	Mean	**Median**	**Minimum**	**Maximum**	**NO. Peak PM_{10} Events**
Glen Eden	13.52	12.41	2.99	42.16	14
Henderson	13.21	12.55	2.98	35.06	1
Pakuranga	14.91	13.73	3.95	59.86	30
Patumahoe	11.82	10.97	1.05	277.85	5
Penrose	15.42	14.85	1.53	44.01	10
Takapuna	14.92	14.36	4.04	37.3	6

The annual average concentration for each site was found to be following the NES and WHO prescribed standard of 20 µg/m^3. The highest annual average of 16.43 (µg/m^3) was observed at Penrose for year 2013 (Table 3.2)

Table 3.2: Annual average of PM_{10} concentrations.

	Year					
	2011	**2012**	**2013**	**2014**	**2015**	**2016**
Glen Eden	13.57	13.84	13.94	13.60	13.2	13.21
Henderson	12.39	13.58	13.61	13.63	12.94	13.08
Pakuranga	13.92	15.13	15.39	15.72	14.54	14.71
Patumahoe	11.32	11.22	12.75	11.93	11.66	12.01
Penrose	14.91	14.93	16.43	16.07	14.32	15.86
Takapuna	16.13	15.27	15.25	14.96	13.54	14.4

The correlation matrix plot of the daily PM_{10} concentration between the stations is provided in Figure 3.4 (b) (page 71) showing the estimated Pearson product moment correlation coefficient. The obvious feature revealed by this comparison is the large difference in variability between the urban and the rural measurements. The variabilities are shown by the color, width, and direction of the ellipses in all subplots. A lower spatial variability is observed between Takapuna and Henderson sites. The corresponding results for the winter (Jun-August) daily PM_{10} averages in Figure 3.3(b) are also very similar.

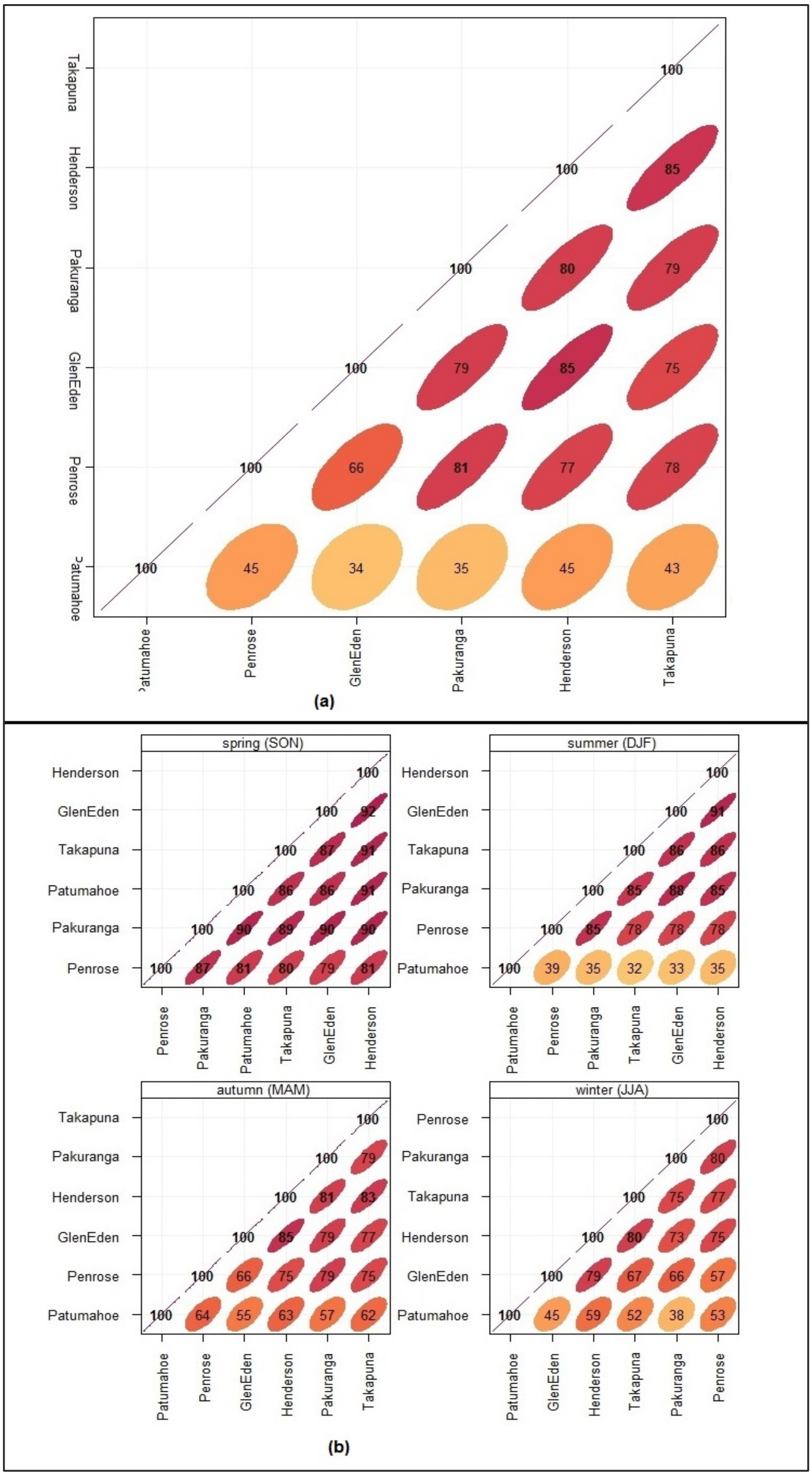


Figure 3.4: (a) daily and (b) seasonal, using daily average data**,** matrix plot of pairwise scatterplot of PM_{10} concentrations (2011-2016). The lower variability the darker the color and narrower the ellipse.

Diurnal Analysis

The highest pollution levels at the study sites were observed during the morning rush hour period (6:00-10:00 NZST) on weekdays (Figure 3.5 to Figure 3.7). The afternoon (15:00-22:00 NZST) peak levels of PM_{10} concentrations can be due to school pick-ups which now blends into rush hour commuter traffic. The solid and biomass fuels from heating is seasonal as well as diurnal.

The day-to-day pattern within a week indicated that mean PM_{10} concentration attained its lowest value on Sundays. The highest measurements are observed on Wednesdays at all sites except for Glen Eden. The highest concentration of PM_{10} at Glen Eden is observed on Saturdays and can be attributed to peak time activities in nearby Ceramco Park Function Center (Council, 2020). The weekly pattern of PM_{10} levels is consistent for Penrose, Takapuna, Henderson, and Pakuranga as the emissions at these sites are predominantly from motor vehicles on nearby motorways and major roads. Although Takapuna, Henderson and Penrose are grouped as 'Citywide' background under the influence of traffic and industrial activities (Talblot & Crimmins, 2020), their diurnal and weekly analysis of PM_{10} show less similarity between them compared to Henderson and Glen Eden (see Appendix A (1)) therefore Glen Eden and Henderson are grouped together for purpose of comparison in Figure 3.5 and Figure 3.6.

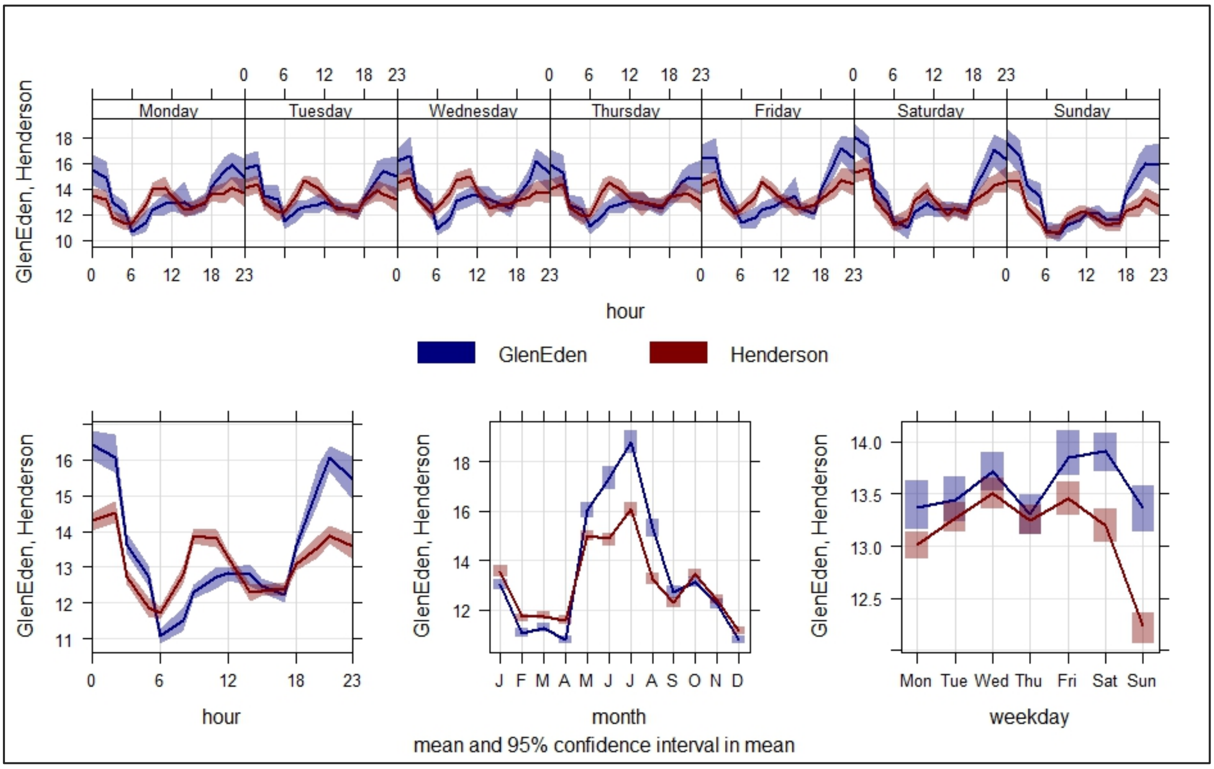


Figure 3.5: Temporal (2011-2016) variations in PM_{10} for Glen Eden and Henderson (the shades are the 95 % confidence intervals of the mean).

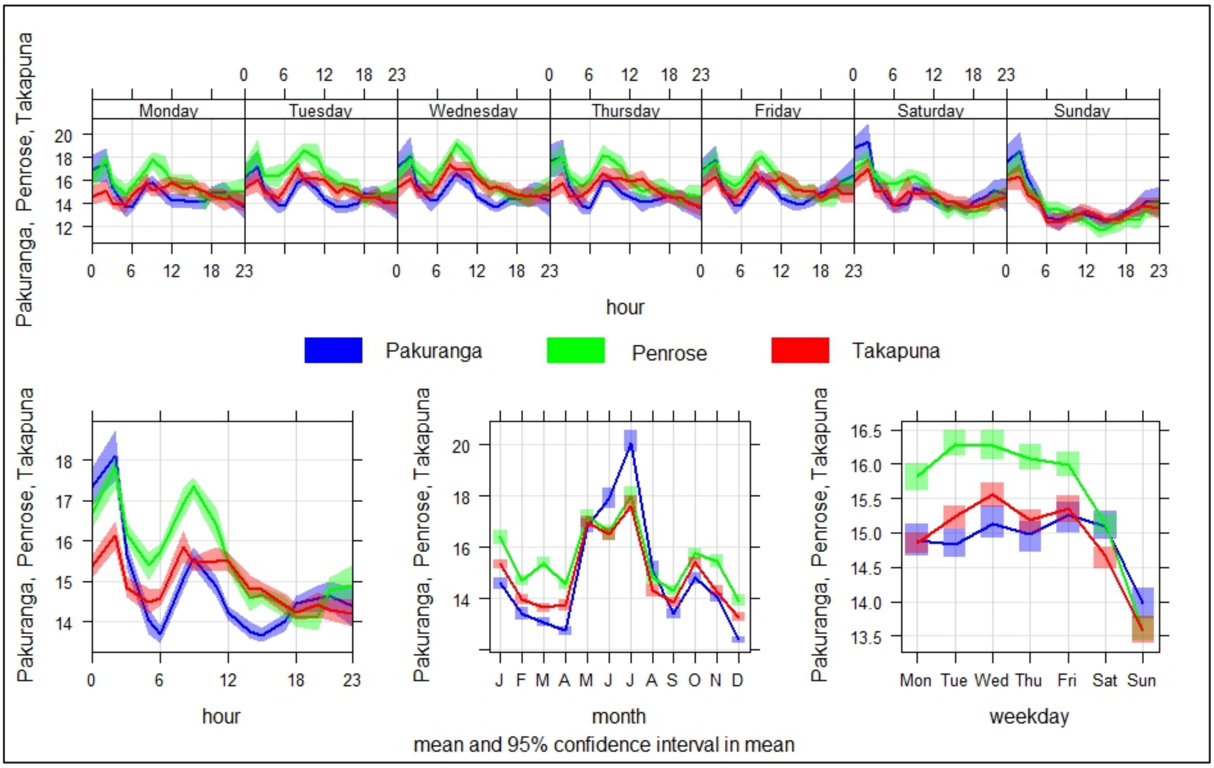


Figure 3.6: Temporal (2011-2016) variations in PM$_{10}$ in Pakuranga, Penrose, and Takapuna, (the shades are the 95 % confidence intervals of the mean).

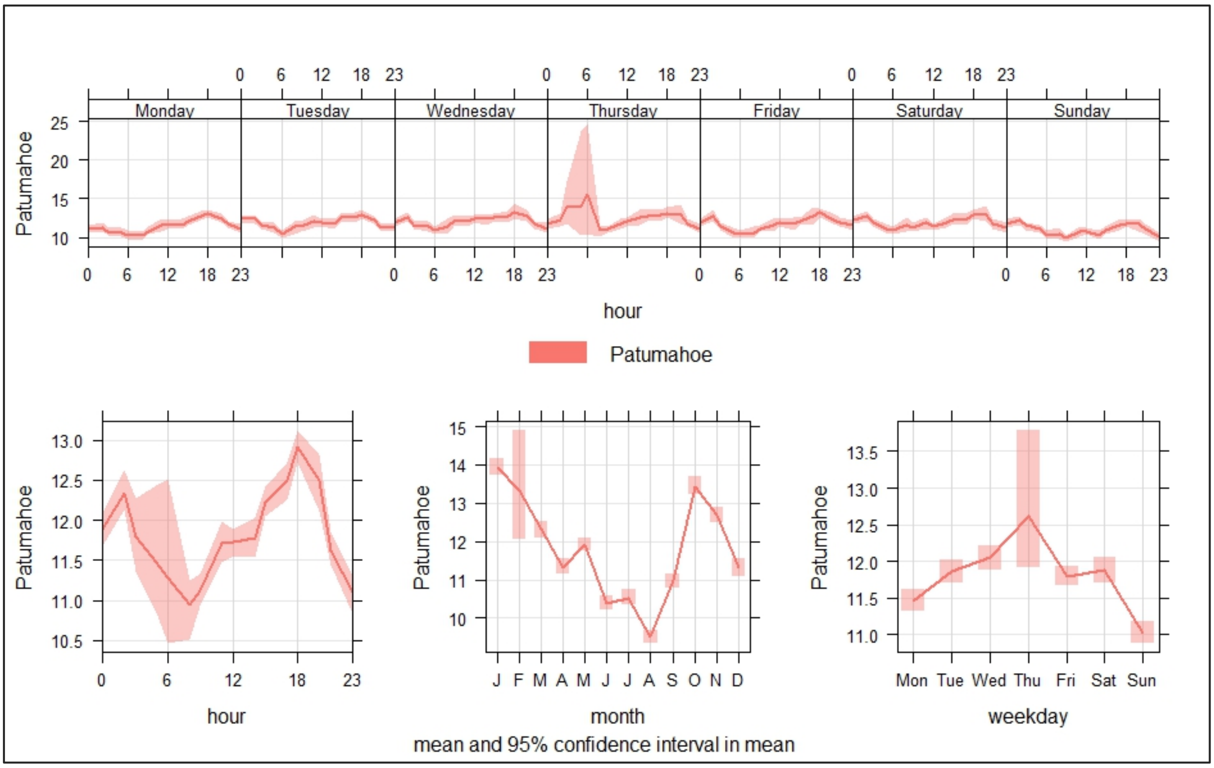


Figure 3.7: Temporal (2011-2016) variations in PM_{10} at Patumahoe rural site (the shades are the 95 % confidence intervals of the mean).

To test the significance of the observed "day effect" and to determine the seasonality given the days of the week, a nonparametric Kruskal–Wallis test (K-W) test for significant difference at a 99% confidence level was used. A low *p*-value indicates a 'significantly' different concentration.

Table 3.3 shows that there is no "day effect" in the daily PM_{10} concentrations for Glen Eden, Pakuranga and Patumahoe. This means that the 24-hour average concentrations of PM_{10} on Mondays are not significantly different than that on Tuesday or any other day of the week. The insignificance of diurnal changes in Patumahoe is notable as it is indicative of the absence of local emission sources at the background site. In contrast, daily variation of PM_{10} for Henderson, Penrose and Takapuna is significant and suggests that there are local emission sources. As mentioned previously, this local diurnal effect can be explained as the result of traffic emissions from nearby major roads and highways.

Table 3.3: Kruskal-Wallis test on day of the week on variation of PM_{10} concentrations (2011-2016). Significance is measured at the $p < 0.05$ threshold.

	Kruskal-Wallis Test	
	***p*-value**	**Remarks**
Glen Eden	0.88	Not significant
Henderson	0.014	Significant
Pakuranga	0.13	Not significant
Patumahoe	0.12	Not significant
Penrose	4.59e-11	Significant
Takapuna	2.38e-07	Significant

Monthly and Seasonal Analysis

Figure 3.8 depicts daily PM_{10} variation over the study area, during the study period, with the cold winter months (May-August) highlighted by a gray box. The dominant feature of this plot is the obvious seasonal nature of the PM_{10} concentrations, with episodic high levels observed during the cold seasons (May-September)/ winter months (June - August) in urban residential sites. The measurements less than or equal to target level of 50 μg/m^3 are plotted under the redline.

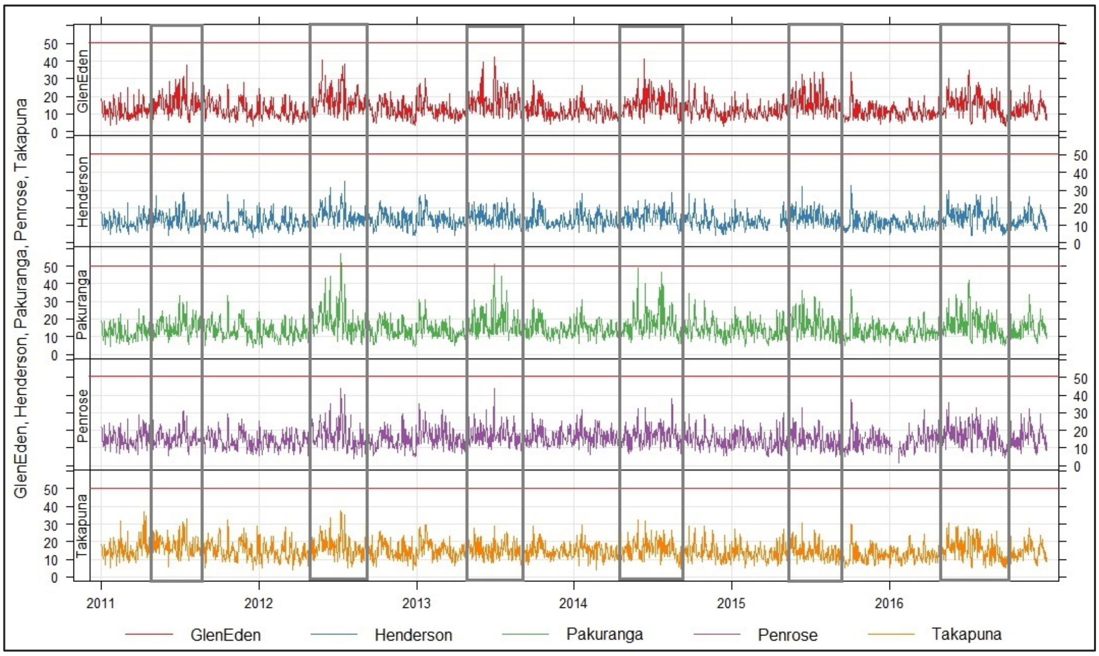


Figure 3.8: Time-series of PM_{10} (24-hour average) during 2011-2016 in urban area.

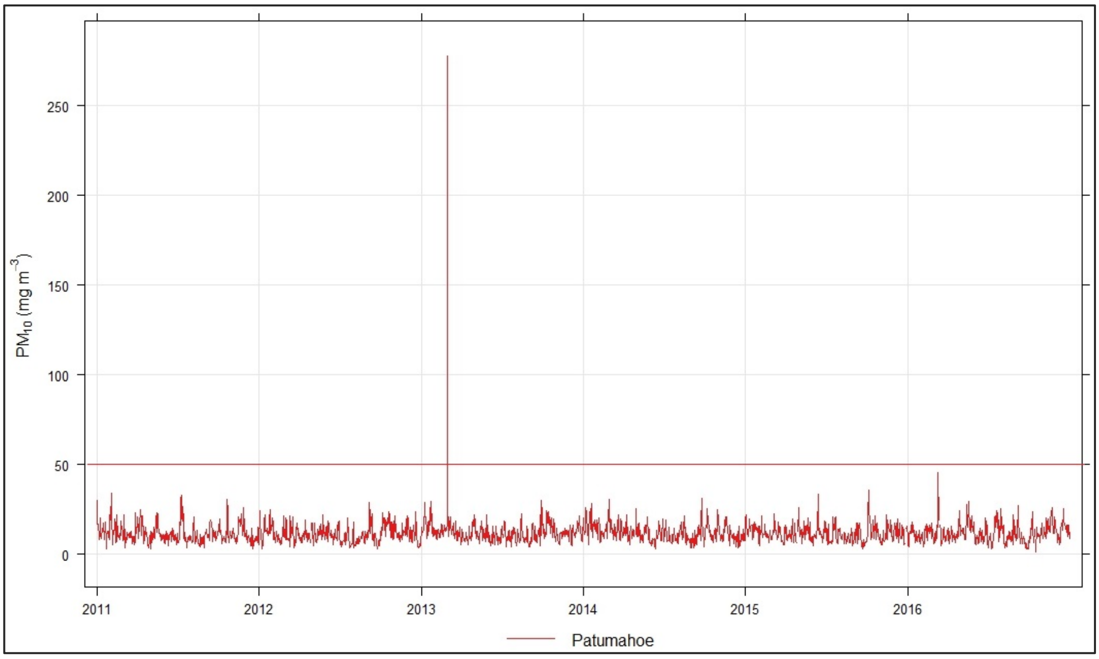


Figure 3.9: Time-series of PM_{10} (24-hour average) during 2011-2016 in rural area.

Average PM_{10} concentrations during winter are higher than spring and summer for all urban sites (Figure 3.10-11). The instances surpassing the prescribed limit of PM_{10} were the minority. However, distinct peaks in PM_{10} concentrations during winter months (June – August) are detected.

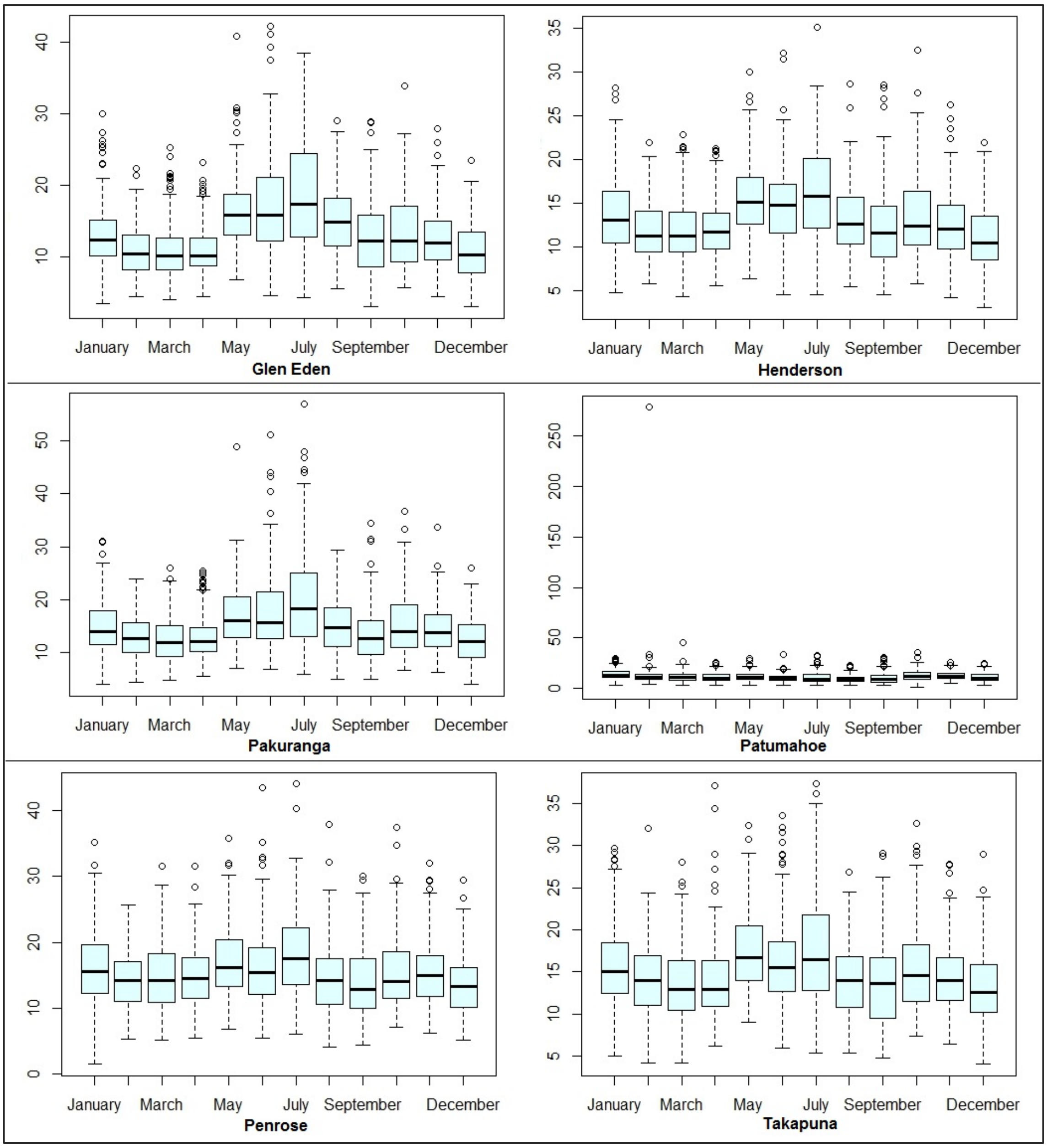


Figure 3.10: Monthly distribution of 24-hour average PM_{10} concentration (2011-2016).

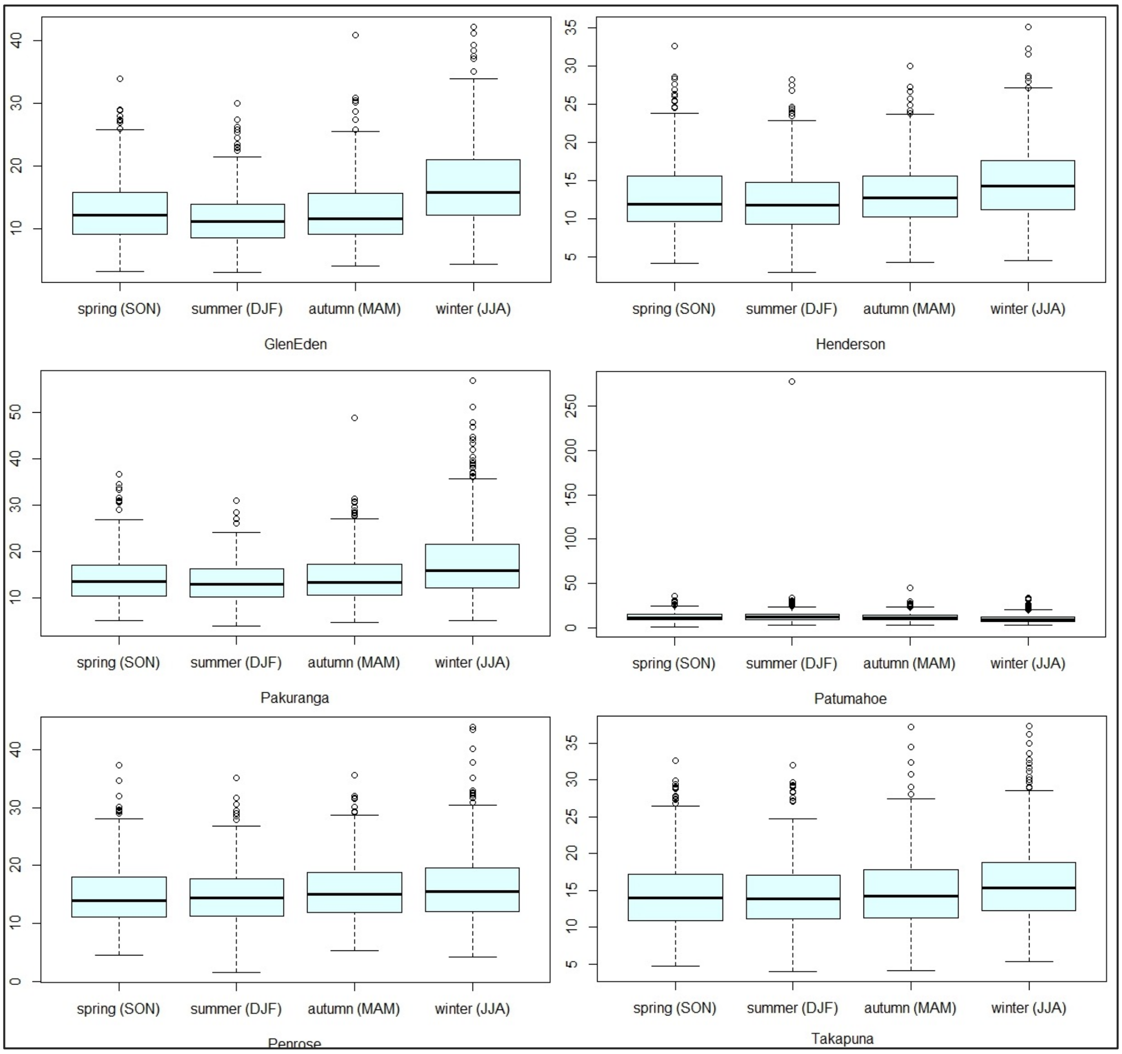


Figure 3.11: Seasonal distribution of 24-hour average PM_{10} concentration (2011-2016).

Better understanding of dispersion characteristics of PM_{10} can be achieved by analyzing the data with respect to seasons. The peak levels of PM_{10} concentrations seen in cold seasons, in contrast to the summer, in Pakuranga and Glen Eden (Figure 3.12) can be attributed to residential heating with solid and biomass fuels.

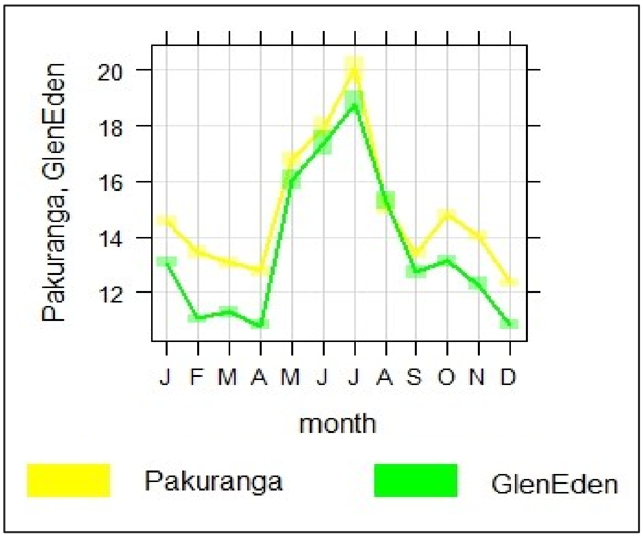


Figure 3.12: Monthly PM_{10} concentrations for Glen Eden and Pakuranga (2011-2016).

Peak PM_{10} concentrations are also evident at Penrose and Takapuna during spring. The summer-winter variations can be explained by the relative contributions to PM_{10} concentrations from different sources at different times of the year (Davy, 2017). The monthly and seasonal variations were analyzed using the K-W test for significance at a 99% confidence level. The result of the K-W test (Table 3.4) showed that the difference in PM_{10} concentration in urban areas during both winter and summer months was statistically significant at the 99% confidence level. The variation observed at the rural site of Patumahoe was not significant during colder seasons (autumn $p = 0.163$ and winter $p \approx 0.02$). This result can be explained by the lack of anthropogenic activities at Patumahoe. The variation between summer and winter concentrations was significantly different at the 99% confidence level for all sites (Table 3.4).

Table 3.4: Kruskal-Wallis test on monthly and seasonal variation of PM_{10} concentrations.

	Monthly *p*-value	**Seasonal *p*-value**			
		spring	**summer**	**autumn**	**winter**
Glen Eden	< 0.01	0.334	< 0.01	< 0.01	0.000
Henderson	< 0.01	0.005	< 0.01	< 0.01	< 0.01
Pakuranga	< 0.01	0.004	< 0.01	< 0.01	< 0.01
Patumahoe	< 0.01	< 0.01	< 0.01	0.163	0.017
Penrose	< 0.01	0.009	< 0.01	< 0.01	< 0.01
Takapuna	< 0.01	0.001	< 0.01	< 0.01	< 0.01

The yearly variations of daily PM_{10} concentration shown in Figure 3.3 were also analyzed using K-W test for significant difference at a 99% confidence level (Table 3.5). The results show that the variation of PM_{10} concentration is significant at all sites except for Glen Eden.

Table 3.5: Kruskal-Wallis test illustrating the extent of the yearly variations in PM_{10} concentrations.

	Kruskal-Wallis Test	
	***p*-value**	**Remarks**
Glen Eden	0.72	Not Significant
Henderson	0.000	Significant
Pakuranga	0.001	Significant
Patumahoe	0.012	Significant
Penrose	1.04e-08	Significant
Takapuna	2.913e-11	Significant

Meteorological Data

The lifetime of pollutant residence in the ambient atmosphere and the formation of secondary pollutants is typically controlled by wind speed, turbulence level, air temperature, and precipitation as well as the rate of source-emission. Wind velocity, wind direction, solar radiation, relative humidity and rainfall influence the concentration of TSP and PM_{10} concentrations in ambient air (Sumesh et al., 2017).

Meteorological monitoring is undertaken at most PM monitoring sites as local meteorology provides insight into pollutant sources, short-term events, chemical reactions, data trends and possible causes of exceedances. The meteorological data for most of the sites was obtained from AC in 1-hour resolution. However, onsite meteorological data were not available for the Patumahoe station. For Patumahoe daily averaged data was collected by author from one of the nearby weather stations owned by the National Institute of Water & Atmospheric Research (NIWA), New Zealand through their "*CliFlo*" online data repository (CliFlo, 2015).

In order to ensure only quality data was utilised in this research, the hourly meteorological data was cleaned by discarding obviously unreasonable data points. Mean wind direction was calculated by converting the wind vectors into its west-east (u) and south-north (v) components using the functions provided in the Openair R-package (Carslaw & Ropkins, 2012).

In Henderson station onsite rainfall was not collected at all. In Takapuna station 85 days of continuous rainfall data were missing. In Penrose station 146 days of missing rainfall were observed. In Penrose, 88 of these days were continuous missing days during 2015, 30 days during December 2013, and 21 continuous missing days during August 2013.

There is not a priori basis for excluding rainfall variables from a model because of the complex nature of atmospheric processes leading to PM_{10} formation. Therefore, to overcome the issue of missing rainfall data, satellite rainfall measurements were obtained from the National Oceanic and Atmospheric

Administration (NOAA) / National Environmental Satellite (NESDIS) website (NOAA, 2014). Data files contained one-hourly Hydro-Estimator (Hydro-Estimator, 2014) accumulations ranging from 0 to 256. A value of zero is a missing value and a value of two means no rainfall. The rainfall data was provided on a latitude/longitude grid with 8001 columns and 3111 rows. Data were taken from the nearest proximity to each of the AC station sites. These values were then converted to rainfall accumulation in millimeters (mm) using the equation specified by NOAA (STAR, 2014) :

$$R = (value - 2) * 3048$$

The satellite rainfall measurements were integrated with the AC meteorological dataset after conversion of rainfall values and adjustment of UTC time to NZ time.

Table 3.6 summarizes the availability of the complete dataset of meteorological measurements used in this research. A key to the abbreviated meteorological variable names used in the table is provided in the table's footnote.

Table 3.6: Summary statistics of available meteorological data after imputation of rainfall missing values (2011-2016).

		Glen Eden	Henderson	Pakuranga	Patumahoe	Penrose	Takapuna
Rain							
	Min	0	0	0	0	0	0
	Mean	0.03	0.06	0.02	3.43	0.03	0.02
	Max	1.86	2.15	0.5	79.2	1.29	2.08
RH							
	Min	52.71	50.6	49.07	40	50.12	49.2
	Mean	78.24	75.58	74.87	83.39	75.73	74.24
	Max	98.07	96.7	94.45	100	98.38	95.04
Temp							
	Min	0	2	4.625	2.3	4.5	5.62
	Mean	14.82	15.23	16.02	14.29	14.98	15.8
	Max	23.04	23.88	25.12	22.8	23.12	23.62
Solar							
	Min	0.45	0.1	0	5.44	0.2	0.24
	Mean	168.15	175.8	177.6	164.54	172.37	174.61
	Max	401.27	562.47	448.4	382.87	406.24	436.65
WS							
	Min	0.38	0.1	0	0	0.24	0.56
	Mean	2.07	1.7	1.1	2.29	2.42	2.52
	Max	9.07	6.4	4.33	8.1	6.82	7.42
WD							
	Min	0.53	0.30	1.69	0	0.06	0.03
	Mean	192.89	204.56	183.35	152.52	189.80	194.64
	Max	359.18	359.58	359.37	360.00	359.35	359.89

Rain: Rainfall (mm), RH: Relative Humidity (%), Temp: Temperature (°C), Solar: Solar Radiation (W/m^2), WS: Wind Speed (m/s), WD: Wind Direction (°), NA: Number of missing data points.

3.4 Influence of Meteorological Factors on PM_{10} Concentration

Partial regression plots were used to assess the usefulness of each meteorological variable once the effects of the other variables have been accounted for. Partial regression terms cannot be directly interpreted, and it is not easy to find other variables that the terms should be corrected for. Therefore, pairwise analysis and Principal Component Analysis (PCA) are considered as methods for parameter selection.

To assess temporal PM_{10} variations, daily time series data with information about local meteorological data and PM_{10} concentrations were created for the period of 2011–2016 using the guidelines described in section 3.2. Figure 3.13- 3.18 show matrix plots of the daily meteorological variables against the PM_{10} concentrations.

It is very clear that daily mean temperature has a negative exponential relationship with the daily average PM_{10} measurements. A study by Zhou et al. (2020) found that precipitation had a significant role on wet removal of PM_{10} in days with low relative humidity (~60%) and a lesser role in the near saturated relative humidity range (90–100%). Similarly, in this study rainfall appears to have some site-specific predictive ability as shown by the low pairwise Pearson's correlation coefficients (see Figures 3.13- 3.18[1]). Based on the matrix plots, solar radiation appears to have limited predictive ability except in the case of the Glen Eden site. Temperature is known to affect fuel usage and ambient chemical reactions and lower temperature result in higher PM_{10} concentrations (Czernecki et al., 2016; De Gooijer & Hyndman, 2006). In this dataset high PM_{10} concentrations also appear to correlate with low temperature days, that can be due to use of wood burners during cold seasons. The relationship between the wind speed and wind direction on PM_{10} concentrations are site-specific. The more complex effects of wind speed and wind direction on PM_{10} distribution are explored in more detail in section 3.3.3.1.

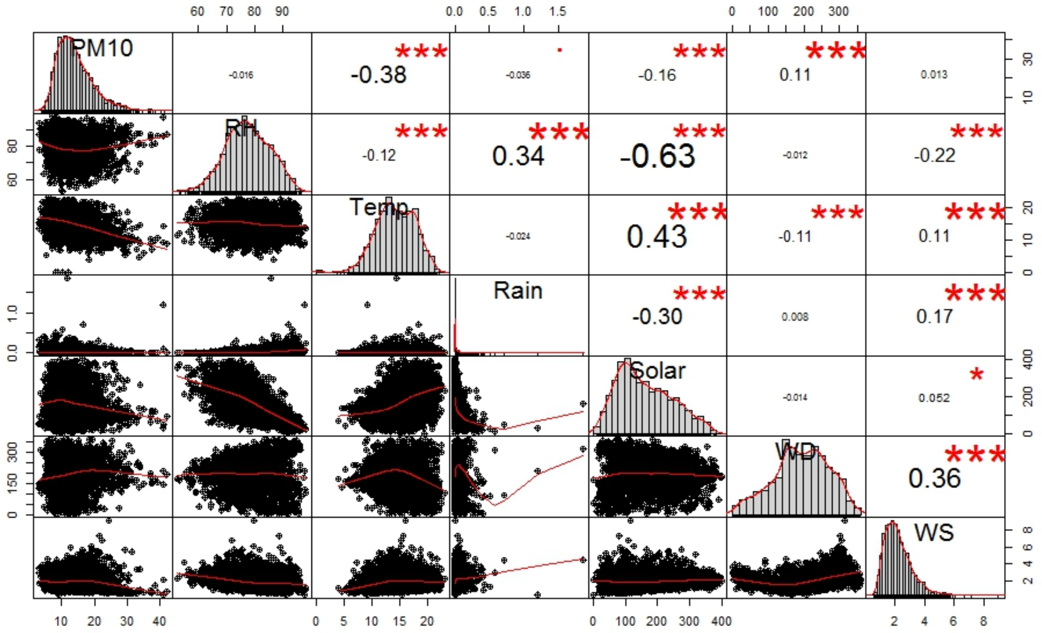


Figure 3.13: Matrix plot, for Glen Eden (2011-2016), showing the distribution of each variable is on the diagonal.

[1] Under the diagonal bivariate scatter plots are depicted with a fitted line. Above the diagonal gives the Pearson's pairwise correlation statistic and its significance level, where $p -$ values $(0, 0.001, 0.01, 0.05)$ is denoted by the symbols (***, **, *, ".") respectively.

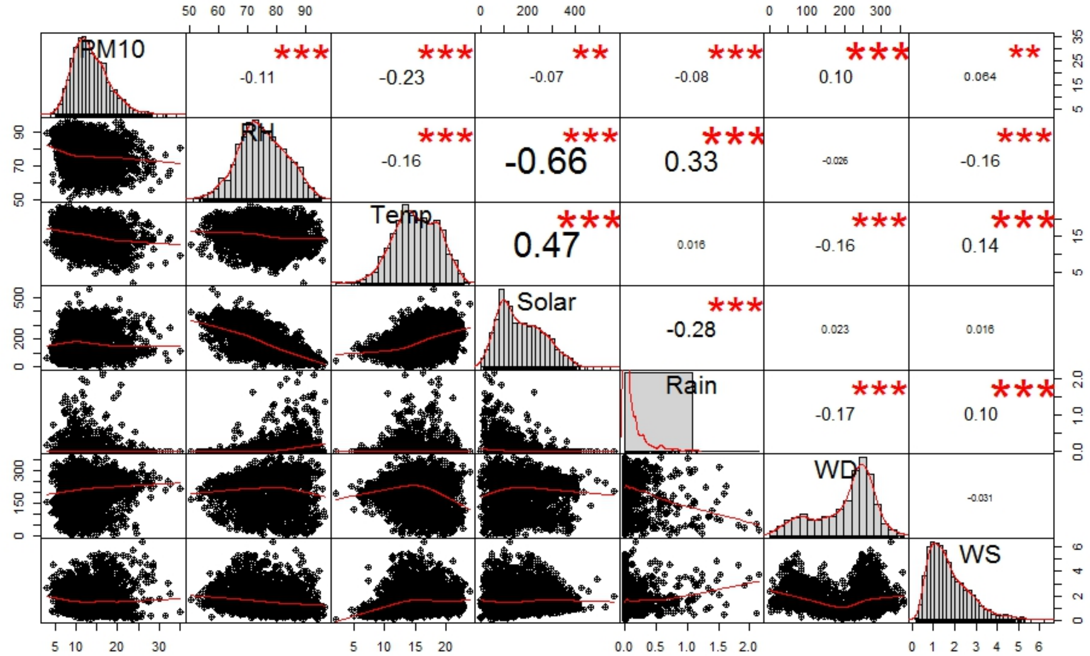


Figure 3.14: Matrix plot, for Henderson (2011-2016), showing the distribution of each variable is on the diagonal.

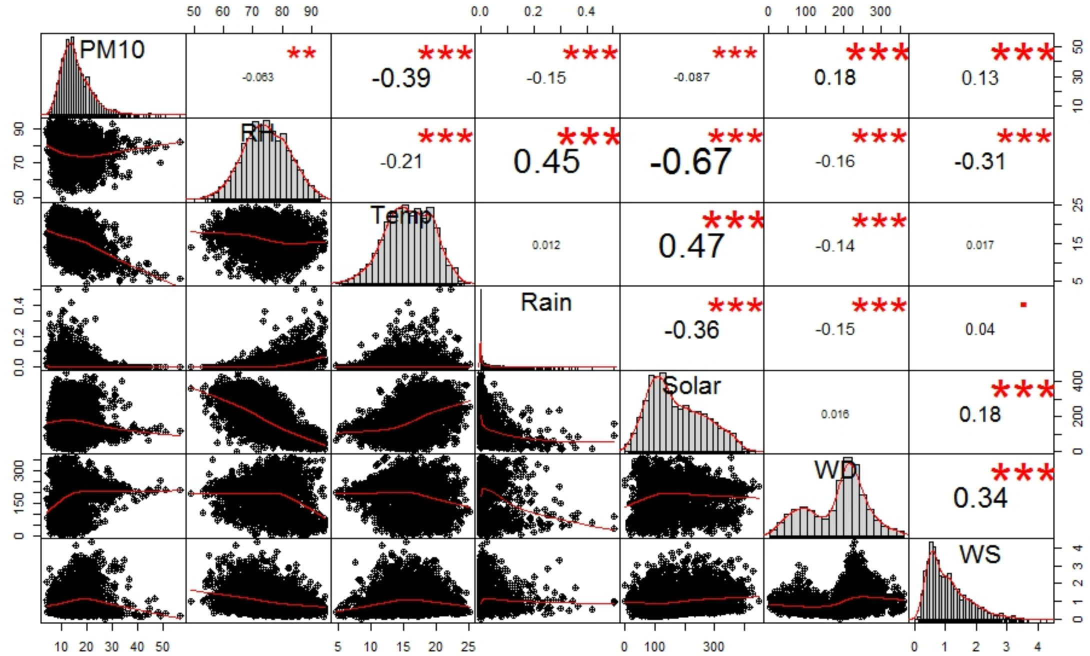


Figure 3.15: Matrix plot, for Pakuranga (2011-2016), showing the distribution of each variable is on the diagonal.

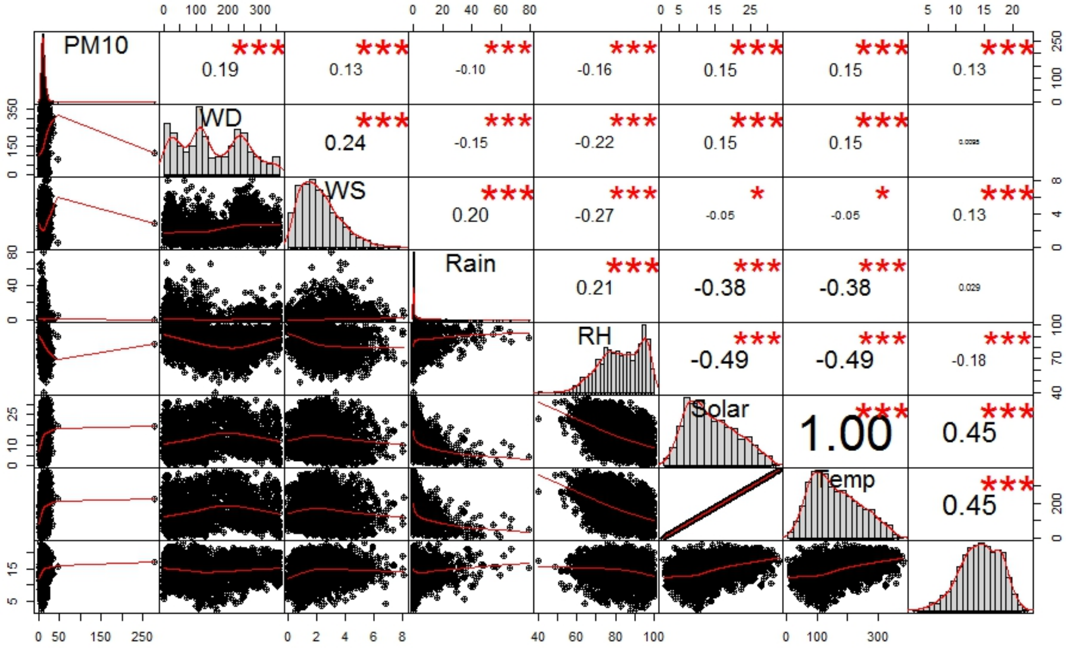


Figure 3.16: Matrix plot, for Patumahoe (2011-2016), showing the distribution of each variable is on the diagonal.

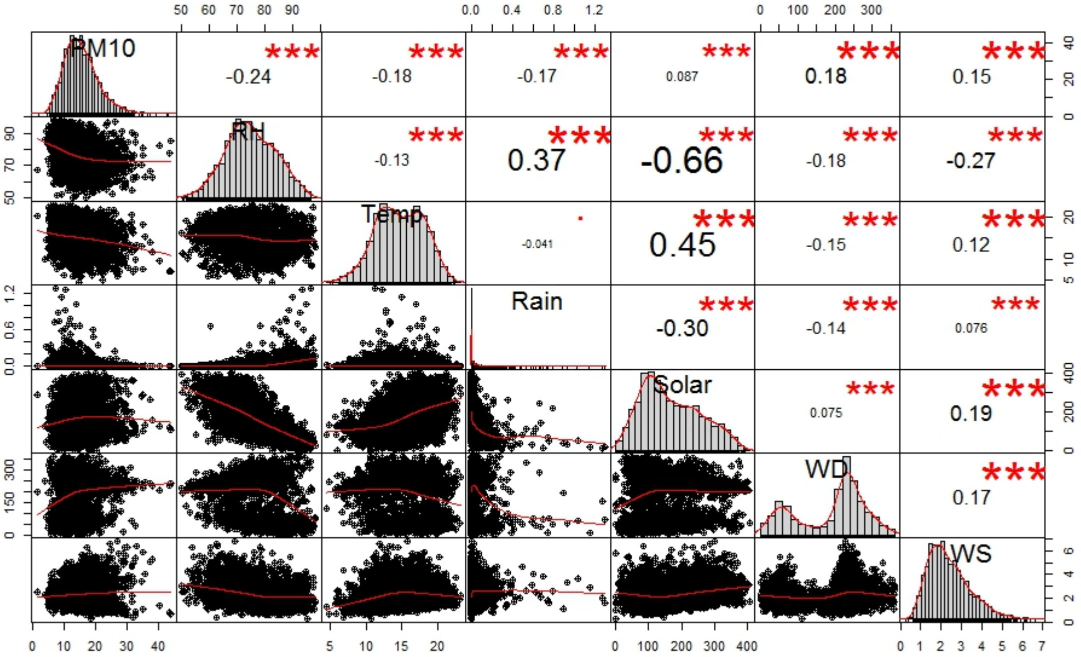


Figure 3.17: Matrix plot, for Penrose (2011-2016), showing the distribution of each variable is on the diagonal.

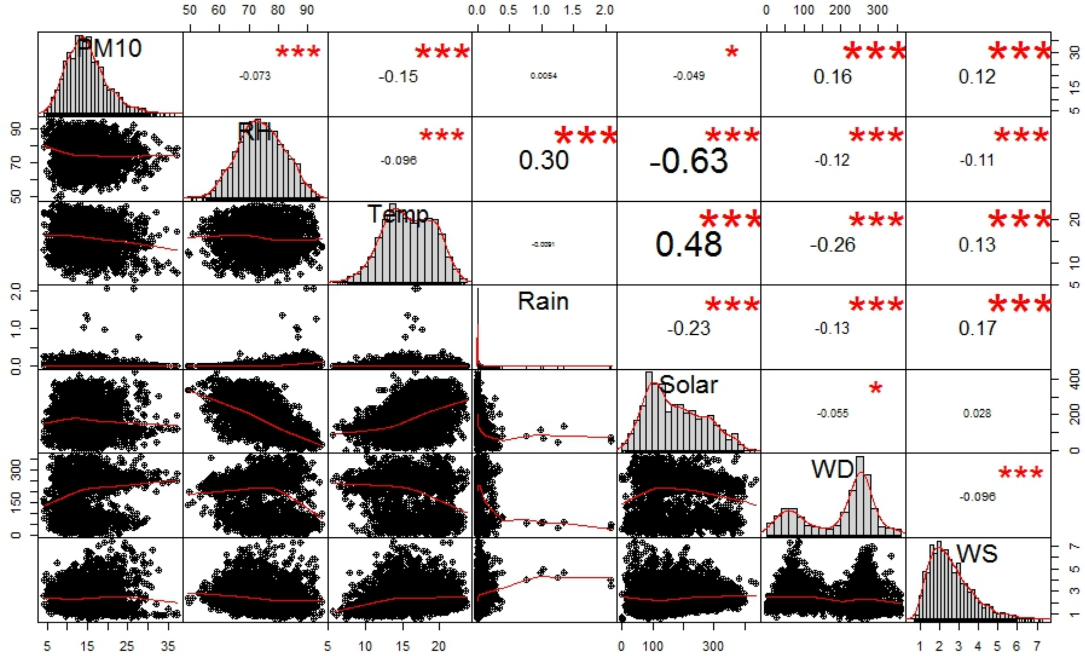


Figure 3.18: Matrix plot, for Takapuna (2011-2016), showing the distribution of each variable is on the diagonal.

3.4.1.1 Wind Speed and Wind Direction

Wind speed and direction can provide insight into pollutant transport within a region and are typically used to measure the relationships between emission sources and pollutant levels (M. A. Elangasinghe, 2014). Auckland mean wind speeds are relatively high due to Auckland's isthmus geography and its maritime environment. The westerly and southwesterly winds are prevailing, but northeasterly flows are also important (Hessell 1988). Bivariate polar plots were used to show the source contributions as a function of both wind speed and direction. Wind speeds and directions were vector averaged using the R's Openair package. For the readers reference, the vector averaging process adopted can be found in Carslaw's work (2012).

In Penrose, the southwesterly quarter wind direction is predominant (Figure 3.19). Indeed, the wind rose for both warm and cold months shows a predominant southwest wind component. In the colder season wind speed is lower and the southeast wind component has a greater spread. In the southeast, a sheltering effect is observed which may be caused by a nearby electricity substation.

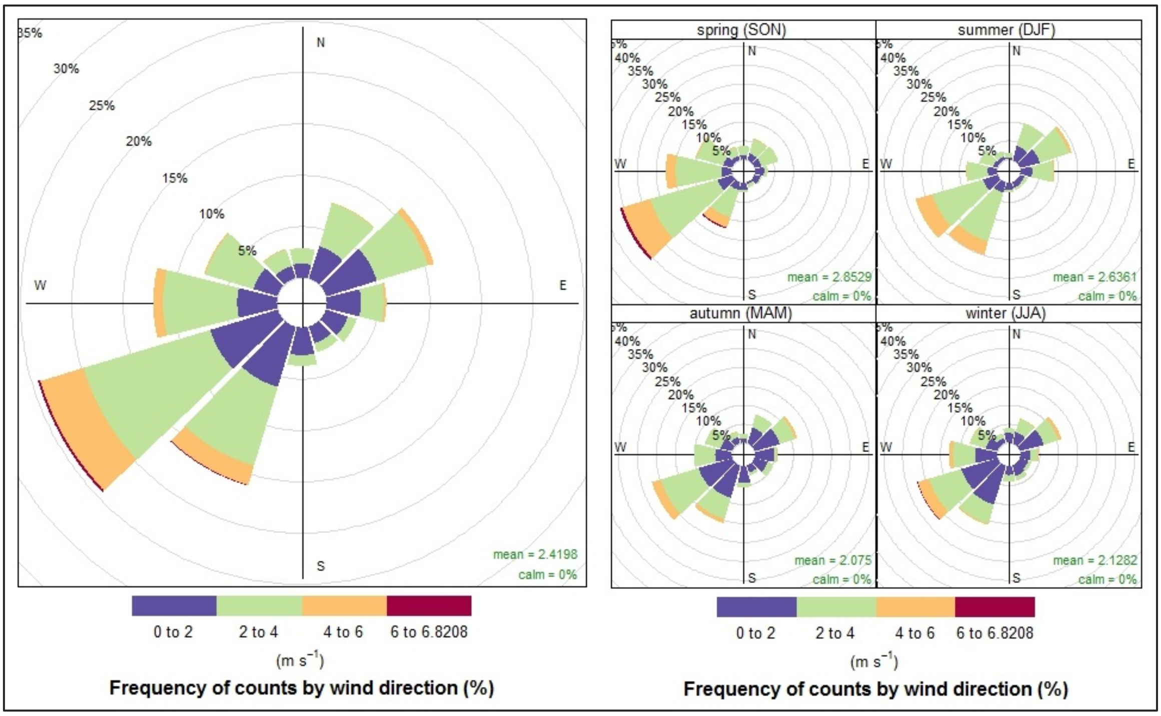


Figure 3.19: Penrose wind rose (left); Seasonal wind rose (right) for 2011-2016.

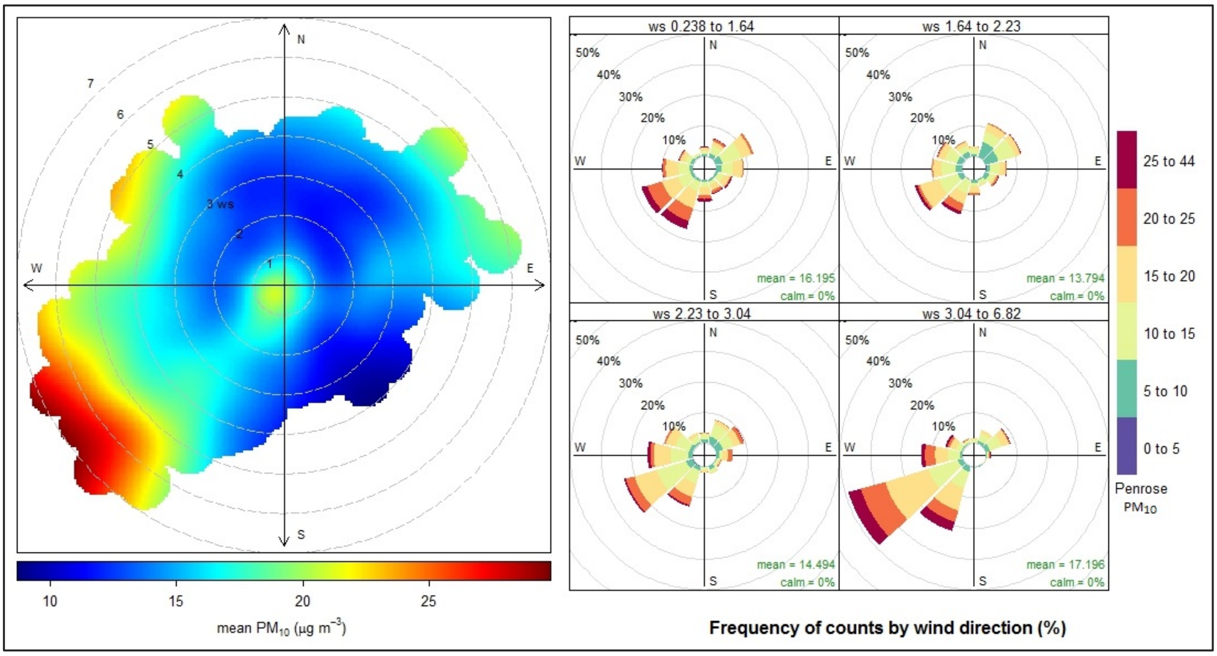


Figure 3.20: Penrose PM$_{10}$ concentration (a); rose (b) conditioned by wind speed 2011-2016.

In Henderson, a southwesterly quarter wind direction is predominant (Figure 3.21 (a)). The result of Figure 3.21(a) clearly shows highest PM$_{10}$ concentrations when the wind is from the westerly directions aligning with marine areoles driven in from Tasman Sea.

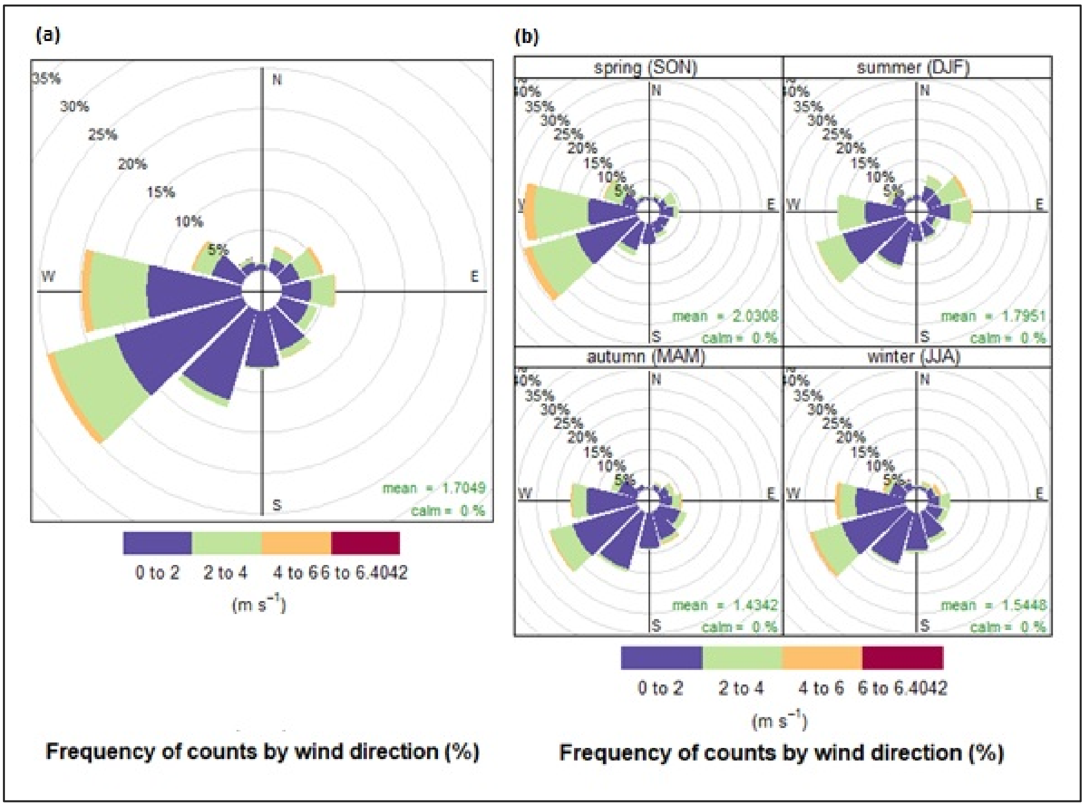


Figure 3.21: Henderson wind rose (a); Seasonal wind rose (b) for 2011-2016.

The prevalence of southwesterly winds is more significant during spring and winter. During the summer, east and northeast wind component are greater than those in other seasons. Figure 3.22(b) shows the variation of PM_{10} by wind speed, showing highest PM_{10} concentrations occur at the two extremes of wind speed. Peak PM_{10} are highest during cold calm winter days under inversion conditions or with a light southerly wind, particularly for anticyclones synoptic conditions.

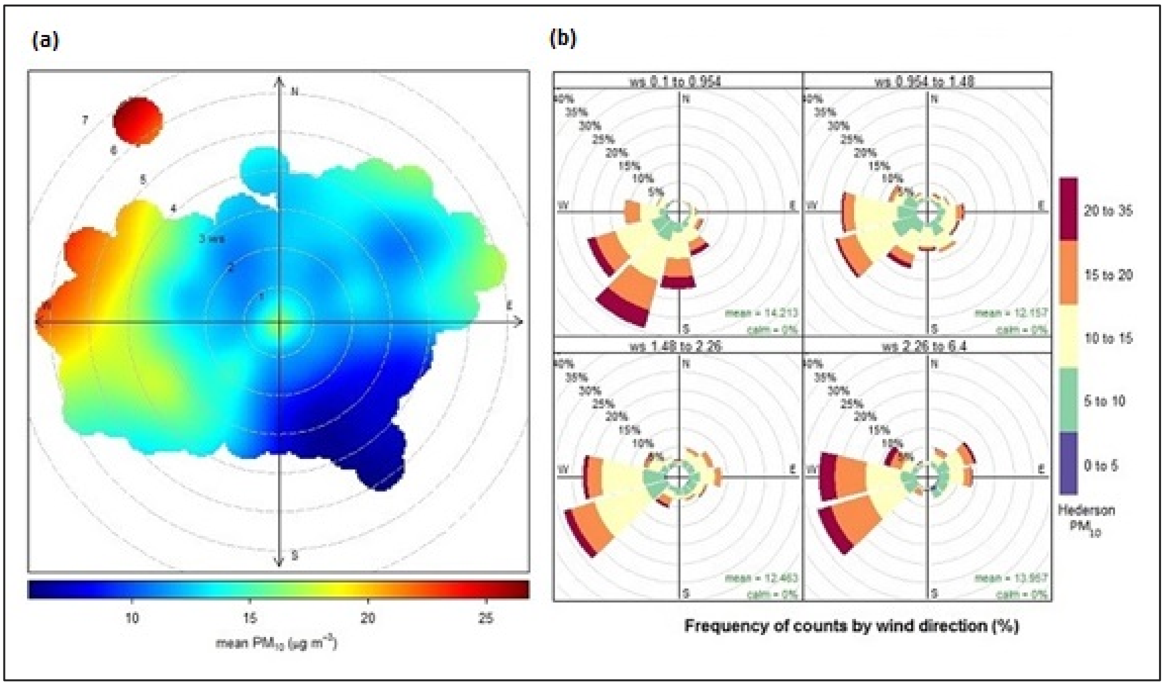


Figure 3.22: Henderson PM_{10} concentration (a); rose (b) conditioned by wind speed 2011-2016.

In Patumahoe the dominant wind is from easterly direction (Figure 3.23 (a)). During the summer months northeast wind component are predominant. Westerly wind component is weak during summer. The winter period has low wind speeds (4.9% calm). Wind directions are more spread from east in autumn.

The highest PM_{10} concentrations are observed during occurrences of high wind speed when the wind prevails in the westerly directions (Figure 3.24). The high concentration of PM_{10} observed in Patumahoe during warmer months can be attributed to dust and soil sources originating from agricultural and land use activities in the area. This conjecture is supported by Davy and Trompetter (2018) who showed that the contribution of soil in PM_{10} concentrations reaches to its lowest during cold seasons.

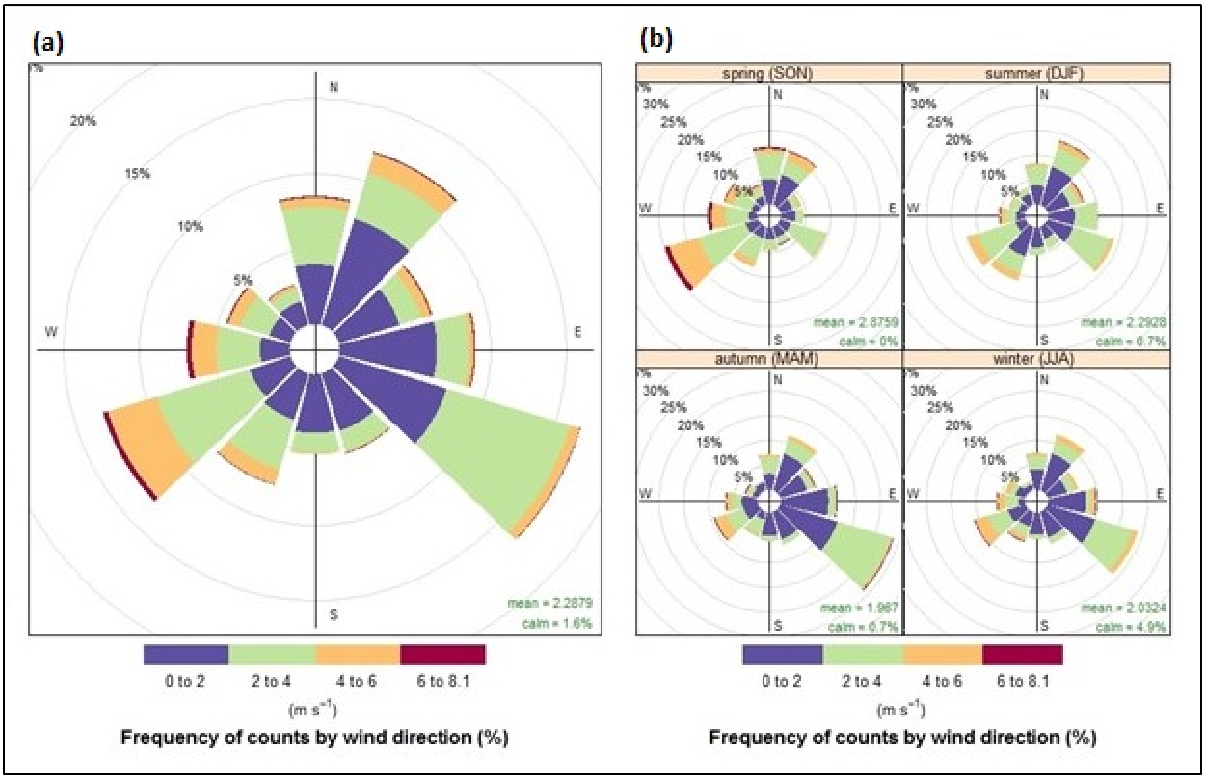


Figure 3.23: Patumahoe wind rose (a); Seasonal wind rose (b) for 2011-2016.

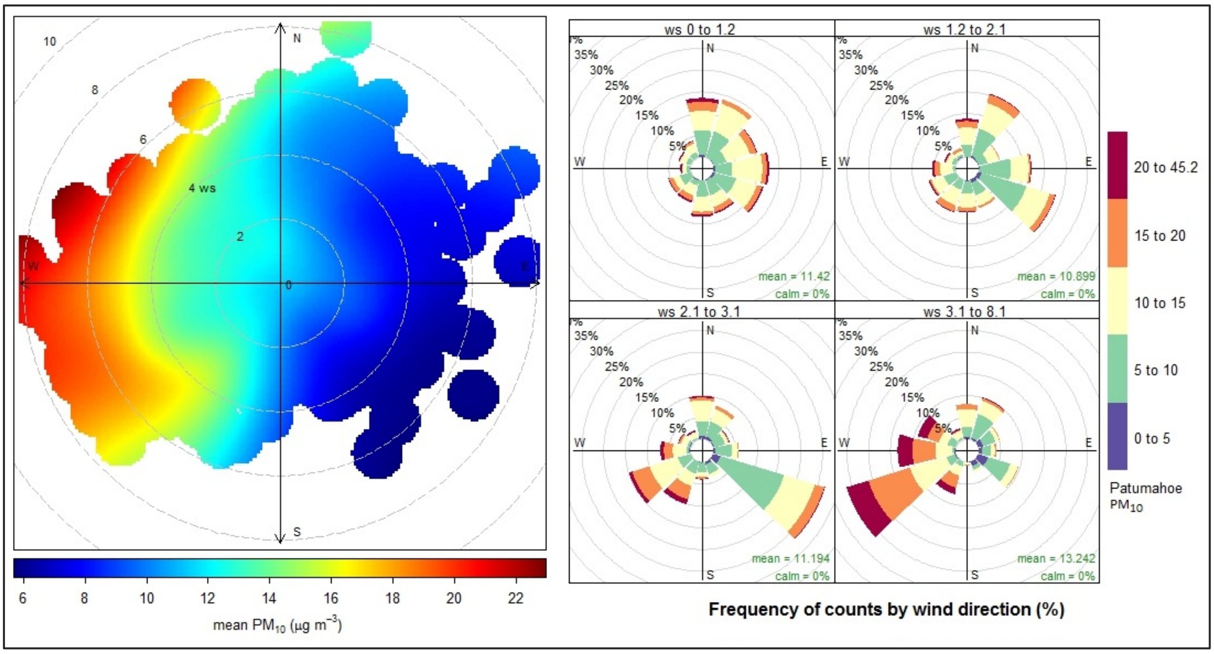


Figure 3.24: Patumahoe PM$_{10}$ concentration (a); rose (b) conditioned by wind speed 2011-2016.

In Glen Eden the predominant wind directions are from the southeasterly and southwesterly quarters (Figure 3.25 (a)). During spring, the northwest wind component is predominant while north and northeast wind component are weak. During winter wind speeds are low with a greater spread of southeast wind directions. The wind rose shows possible sheltering of the monitoring site from wind by housing and hills to the northeast (Figure 3.25 (b)).

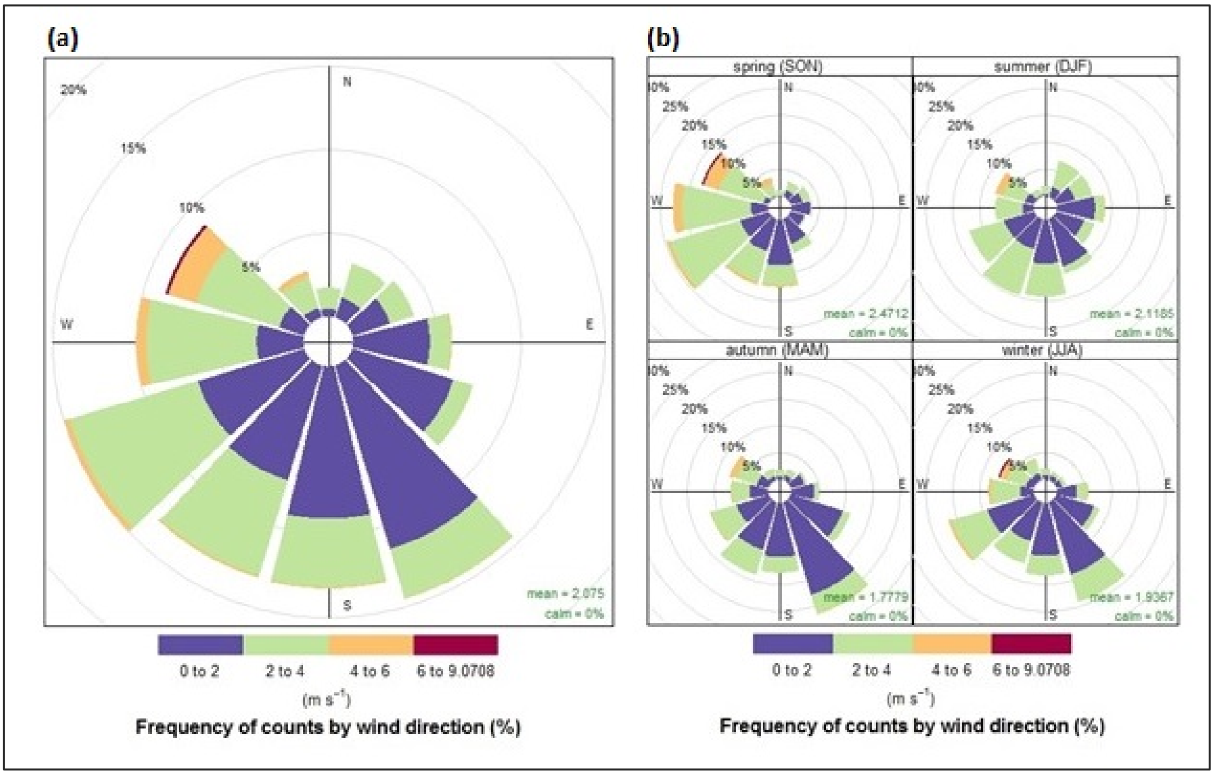


Figure 3.25: Glen Eden wind rose (a); Seasonal wind rose (b) for 2011-2016.

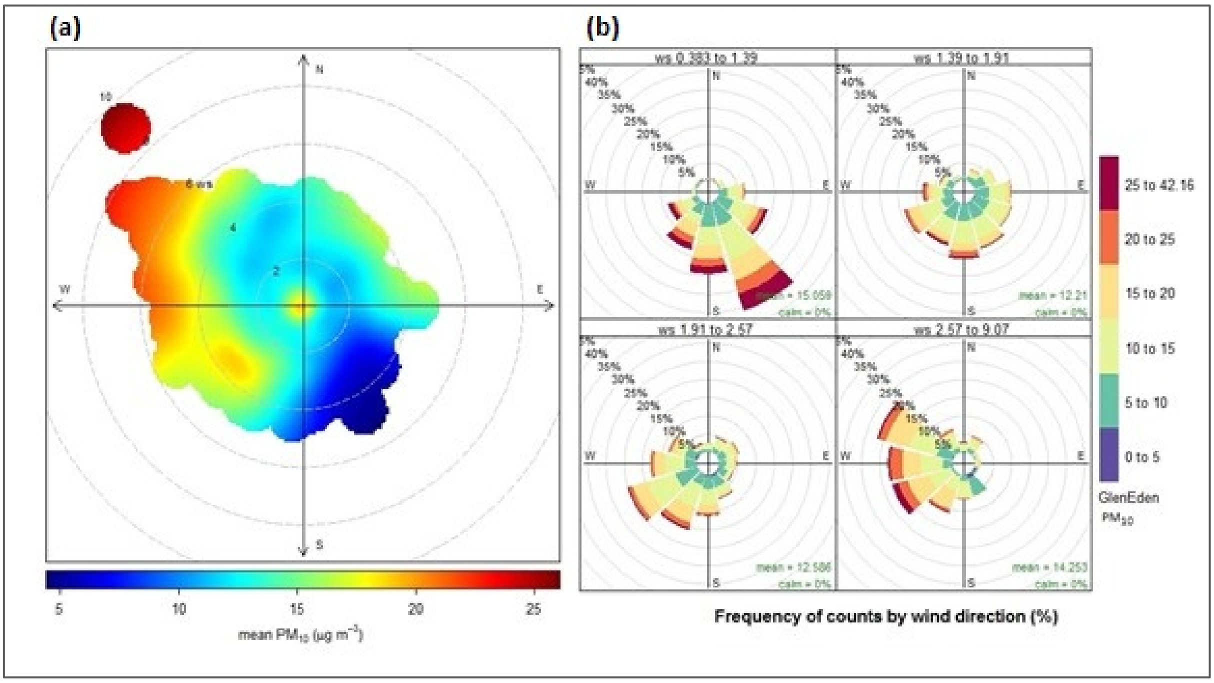


Figure 3.26: Glen Eden PM_{10} concentration (a); rose (b) conditioned by wind speed 2011-2016.

Figure 3.26 shows high PM_{10} concentrations to the northwest which aligns with high wind speeds. It is likely that the traffic to the west (Glendale Rd) as well as multi-functional conference center and local business activities along the road are contributors to the high PM_{10} concentrations. Peak concentrations are observed during periods of low to moderate wind speeds in Figure 3.26 (b).

In Pakuranga**,** the wind direction from th**e** southwesterly quarter is predominant (Figure 3.27). Predominant southwest wind component is observed in all seasons. The northwest sheltering effect is most likely due to the nearby residential buildings (approx. 2m) and in that direction.

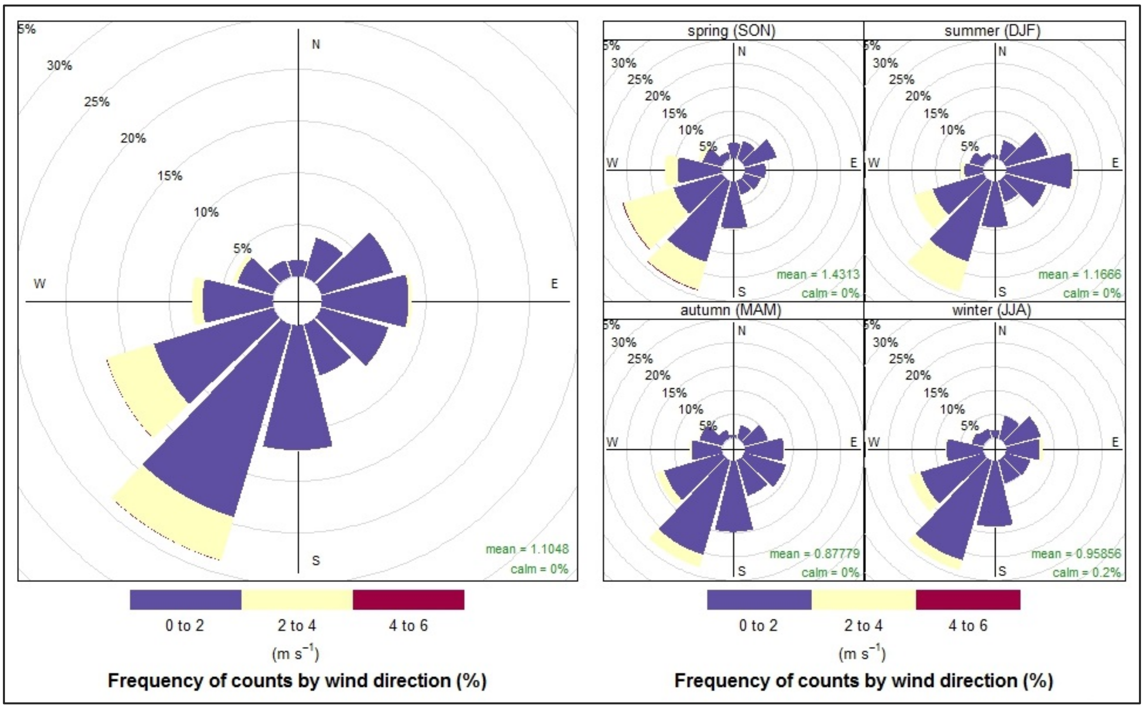


Figure 3.27: Pakuranga wind rose (left); Seasonal wind rose (right) for 2011-2016.

The highest PM_{10} concentrations are observed in the west and southwest directions under high wind speed conditions. Peak PM_{10} is observable in the southwest direction when calm to moderate winds is aligned with Pakuranga highway. Peak levels at the centerline of the polar graph with slow wind speed may be due to biomass burning by nearby residential houses in west side of the monitoring station.

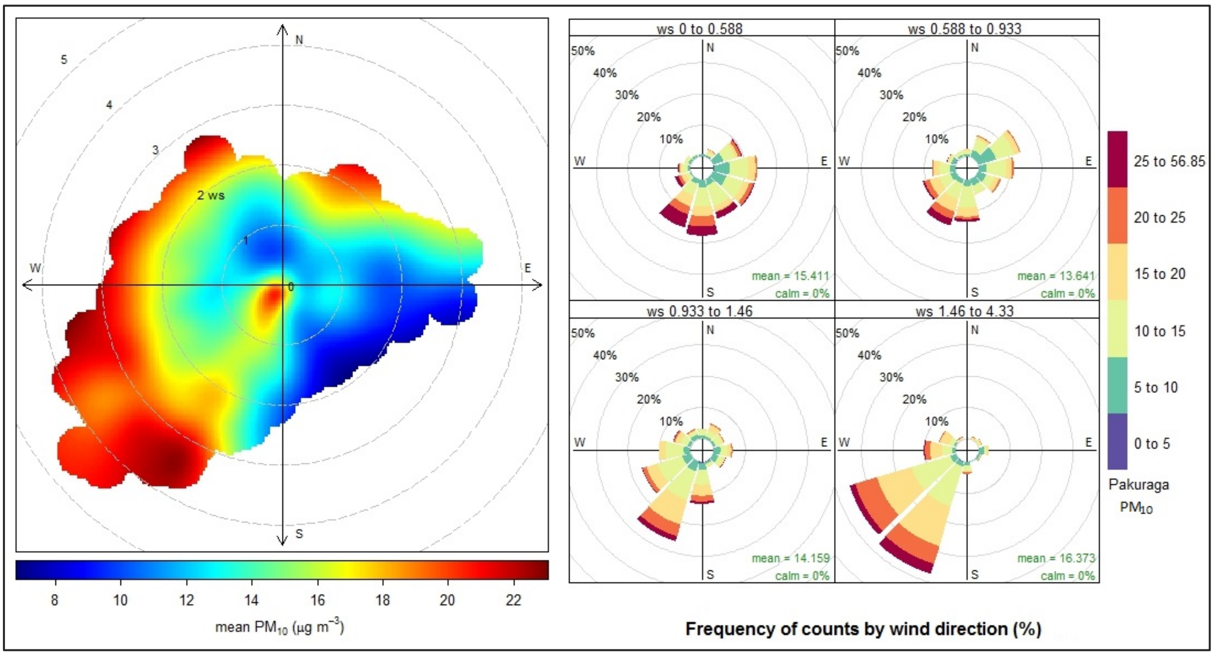


Figure 3.28: Pakuranga PM_{10} concentration (a); rose (b) conditioned by wind speed 2011-2016.

In Takapuna, the wind direction from the southwesterly quarter is predominant (Figure 3.29 (left)). In spring west to southwesterly winds are predominant, while during the summer the greatest components of winds are originating from the northeast and the southeast (Figure 3.29 (right)).

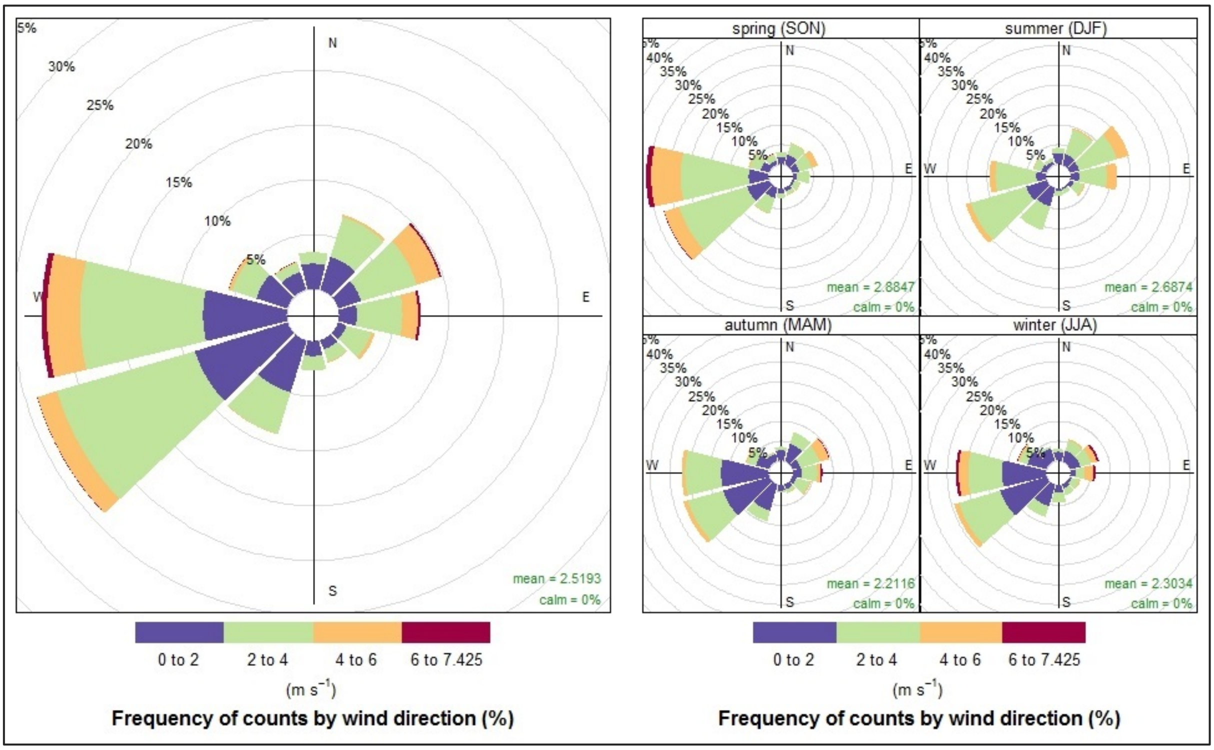


Figure 3.29: Takapuna wind rose (left); Seasonal wind rose (right) for 2011-2016.

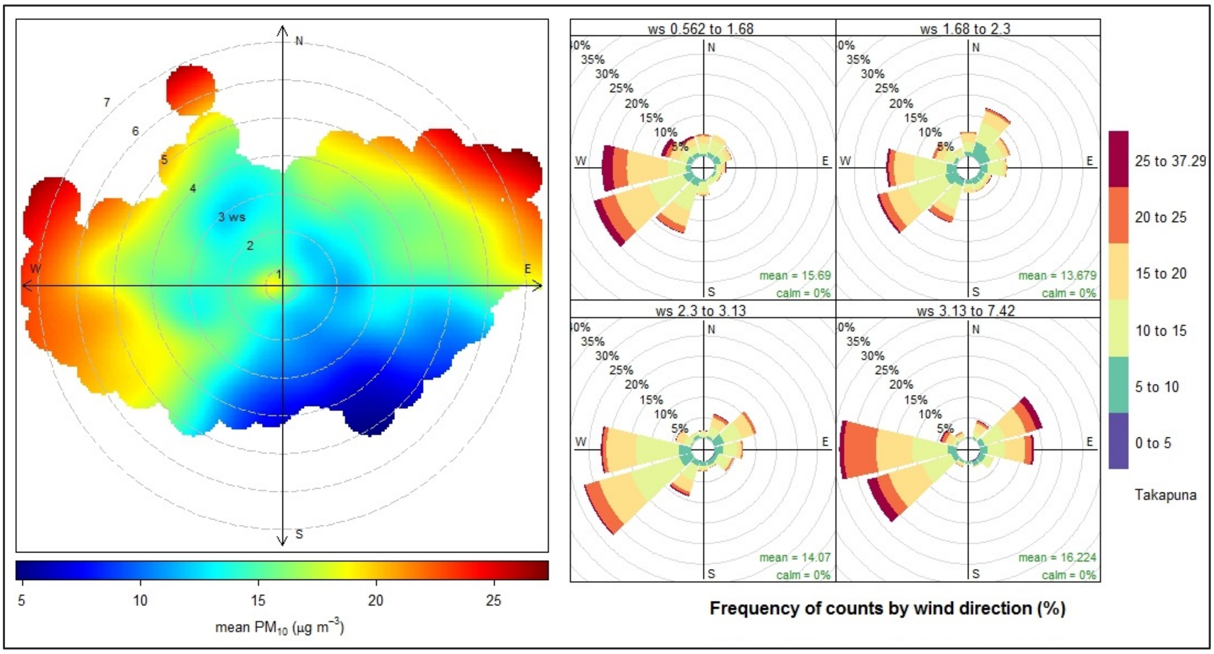


Figure 3.30: Takapuna PM_{10} concentration (a); rose (b) conditioned by wind speed 2011-2016.

The moderate to strong winds from the west/southwest direction (aligns with the nearby motorway) causes the highest concentration observed Figure 3.30 (a). The easterly concentration is more likely attributed to emissions from ships operating at Port of Auckland as suggested by Davy (2017). The same study by Davy (2017), suggested that the marine aerosol from the Tasman Sea and Pacific Ocean are the primary source of PM_{10} concentration in the west-southwest and east-northeast directions, and are associated with incidences of high wind speed.

3.5 Autocorrelation

The correlation between values of a process, with itself, at different time points is known as the autocorrelation of a random process (Stark & Woods, 2012). A measure of autocorrelation as the mean over time is calculated using the auto-correlation function (ACF). ACF is gives the correlation coefficient between observations divided by the specified time lags. The Partial Autocorrelation Function (PACF) is used to determine the order of the correlation structure.

The NES target has been specified in terms of a daily average from midnight to midnight. This means that evening measurements will run into the next morning. This overlap has the potential to induce correlation in PM_{10} concentrations between neighboring days. To investigate the time-dependent correlation of PM_{10} among neighboring days, site-specific estimation of ACF and the PACF are generated and analyzed (Figures 3.31-36). In these correlograms, ACF starts with lag_0 (the correlation of a value with itself) and PACF starts at lag_1 (correlation with the previous day).

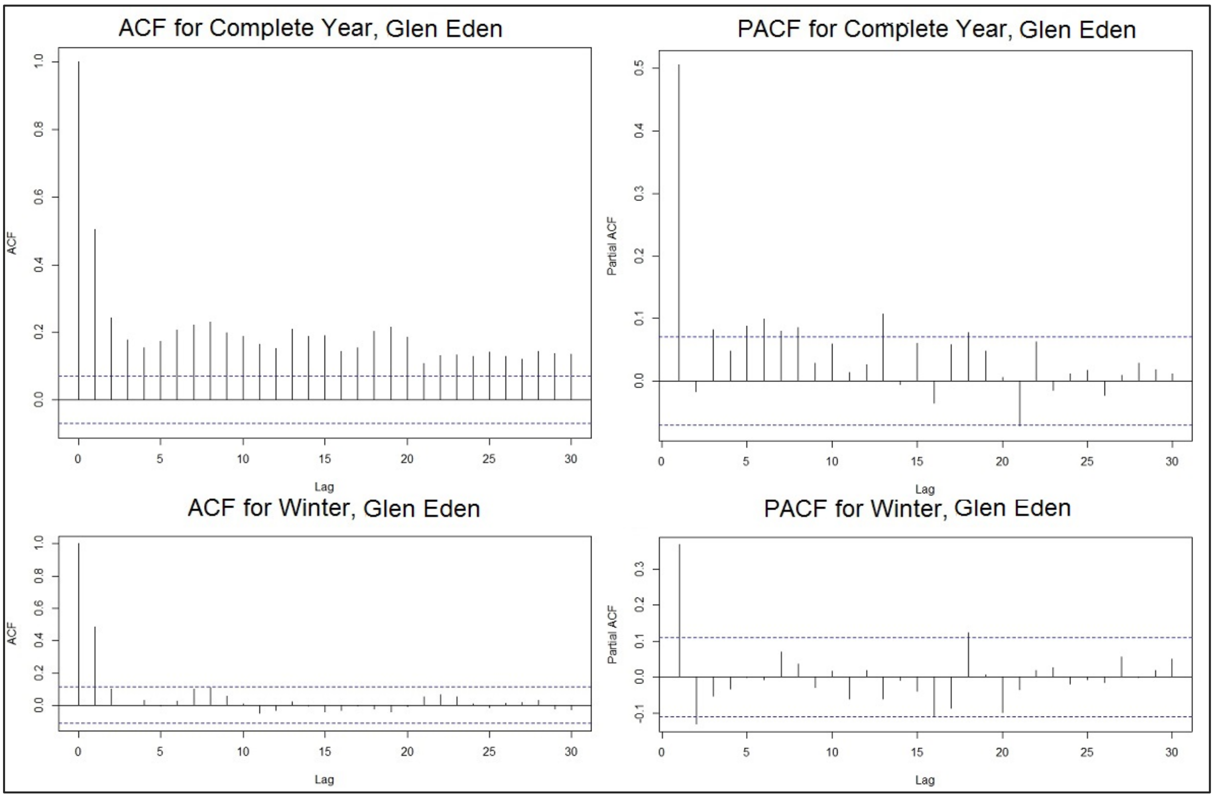


Figure 3.31: Auto-correlation Function of Glen Eden PM_{10}

The blue dashed lines represent lag wise 95% confidence interval signifying uncorrelated random variables within this limit. Nearby 5% of the estimated autocorrelations are likely to be outside of these limits. It is clear from the plots that the PM_{10} concentrations exhibit moderate positive autocorrelation at lag_1 (yesterday's PM_{10}) varying between 0.3-0.6 within the sites. The complete year partial autocorrelation is negative but not significant at lag_2 in all sites.

Lag_1 for Patumahoe (Figure 3.34) and Pakuranga (Figure 3.33) shows significant autocorrelation whereas for the remaining sites this correlation is considered to be borderline in terms of statistical significance. The differences observed between the autocorrelation in winter and that of the entire year may be due to the variability of meteorological conditions during in winter months. These differences are site specific with the lowest differences occurring at Patumahoe (Figure 3.34), Penrose (Figure 3.35) and Takapuna (Figure 3.36).

The lag_2 negative correlation in Takapuna, Pakuranga and Henderson during winter suggests that today's concentrations are negatively related to the concentration of two days ago.

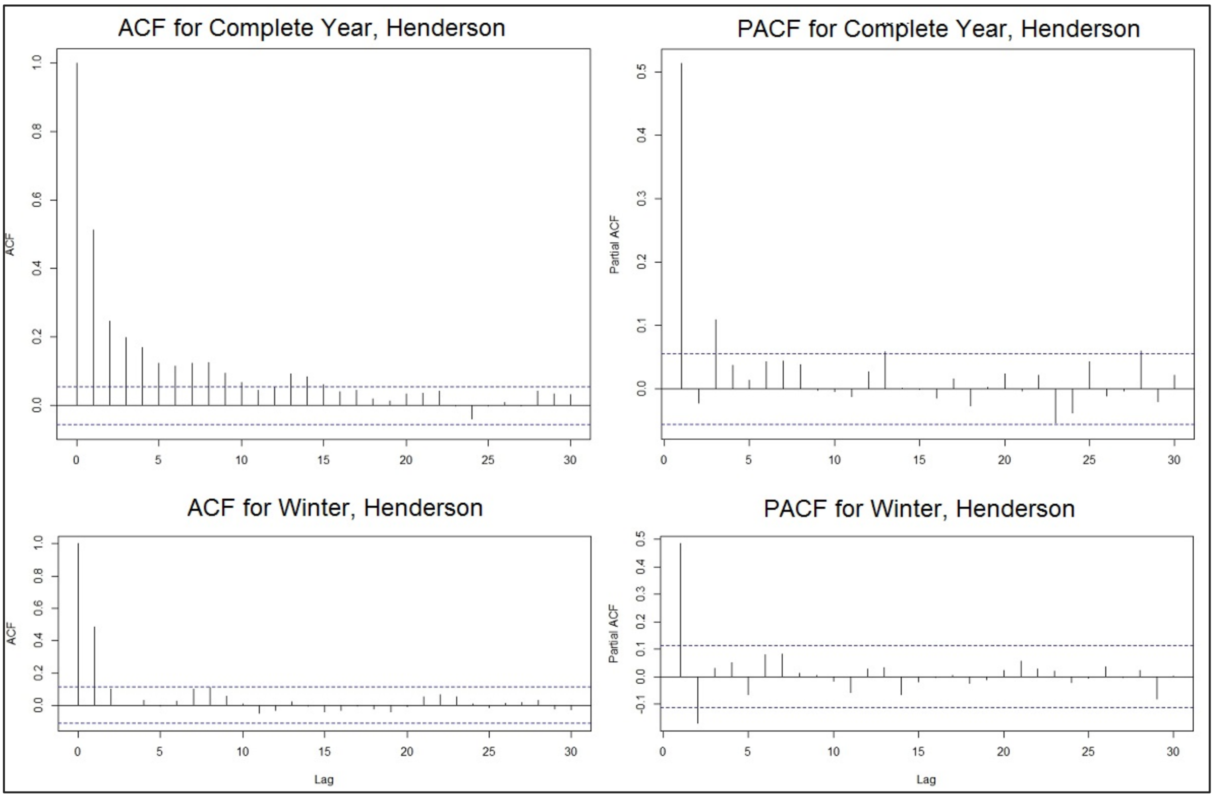


Figure 3.32: Auto-correlation Function of Henderson PM_{10}

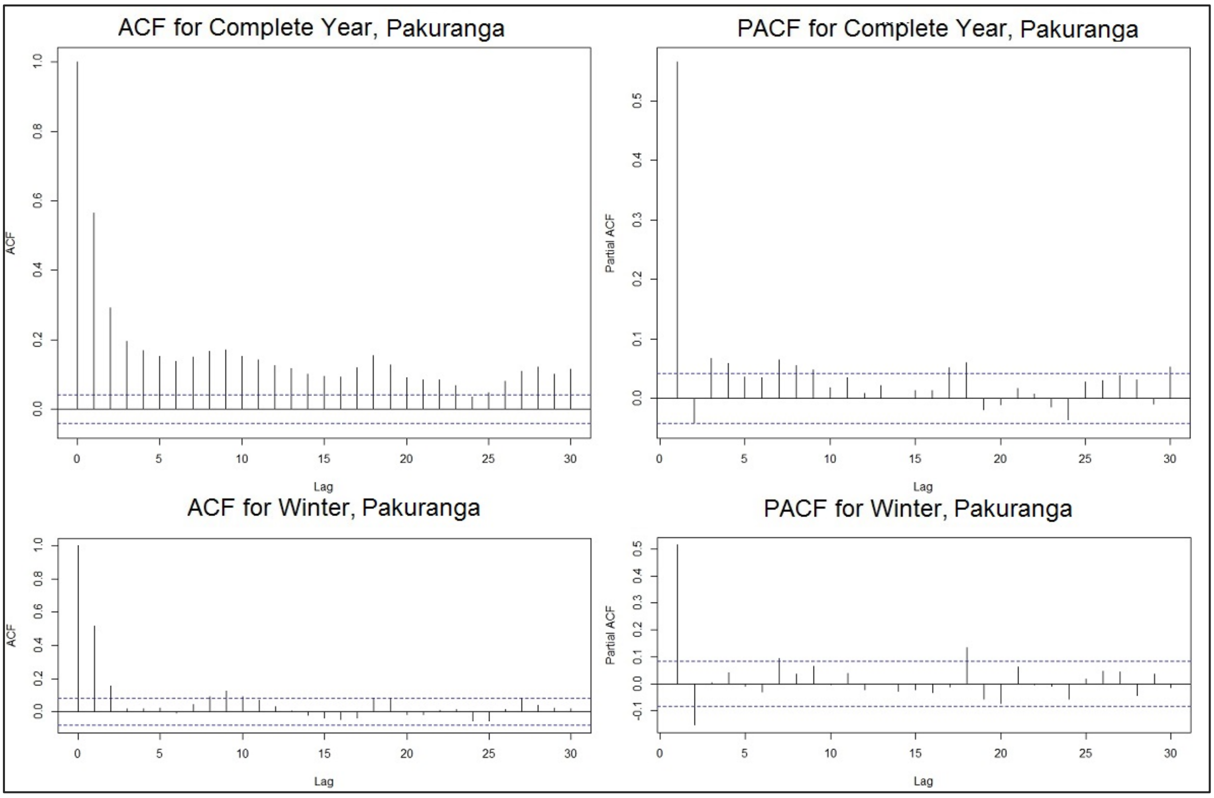


Figure 3.33: Auto-correlation Function of Pakuranga PM_{10}

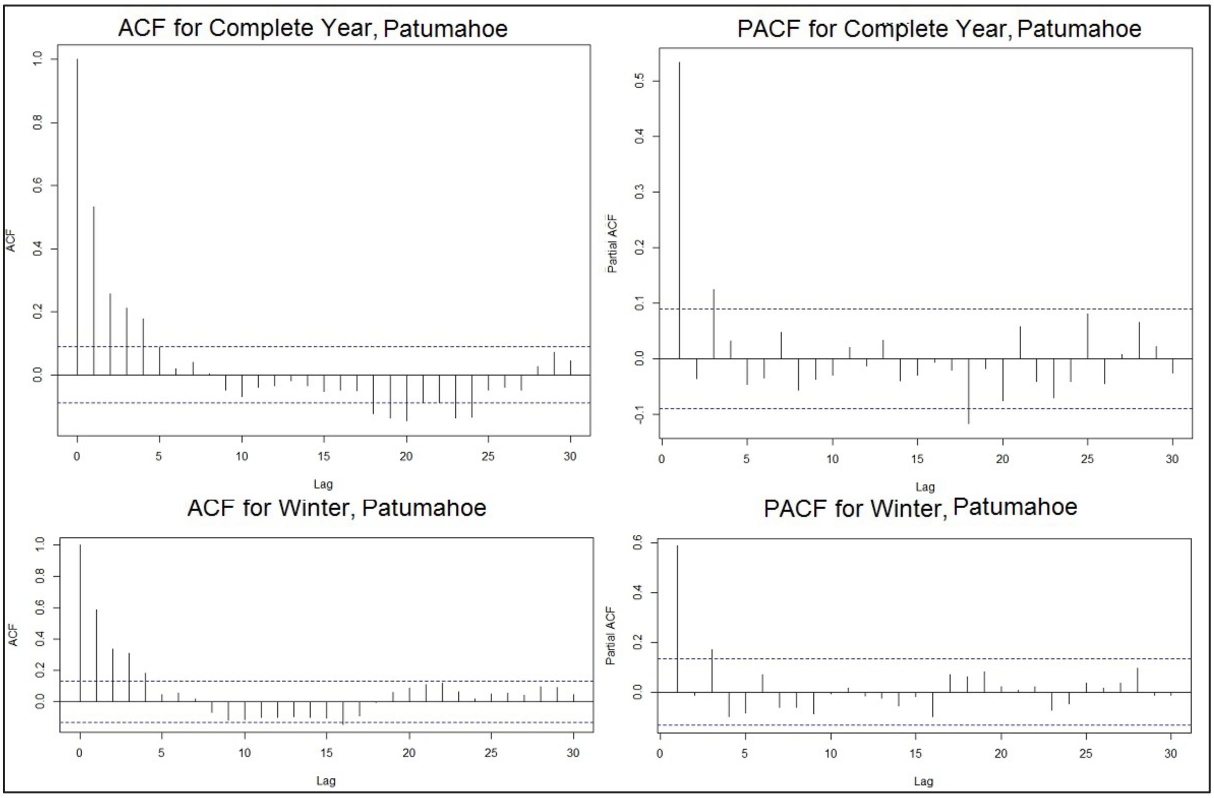


Figure 3.34: Auto-correlation Function of Patumahoe PM_{10}

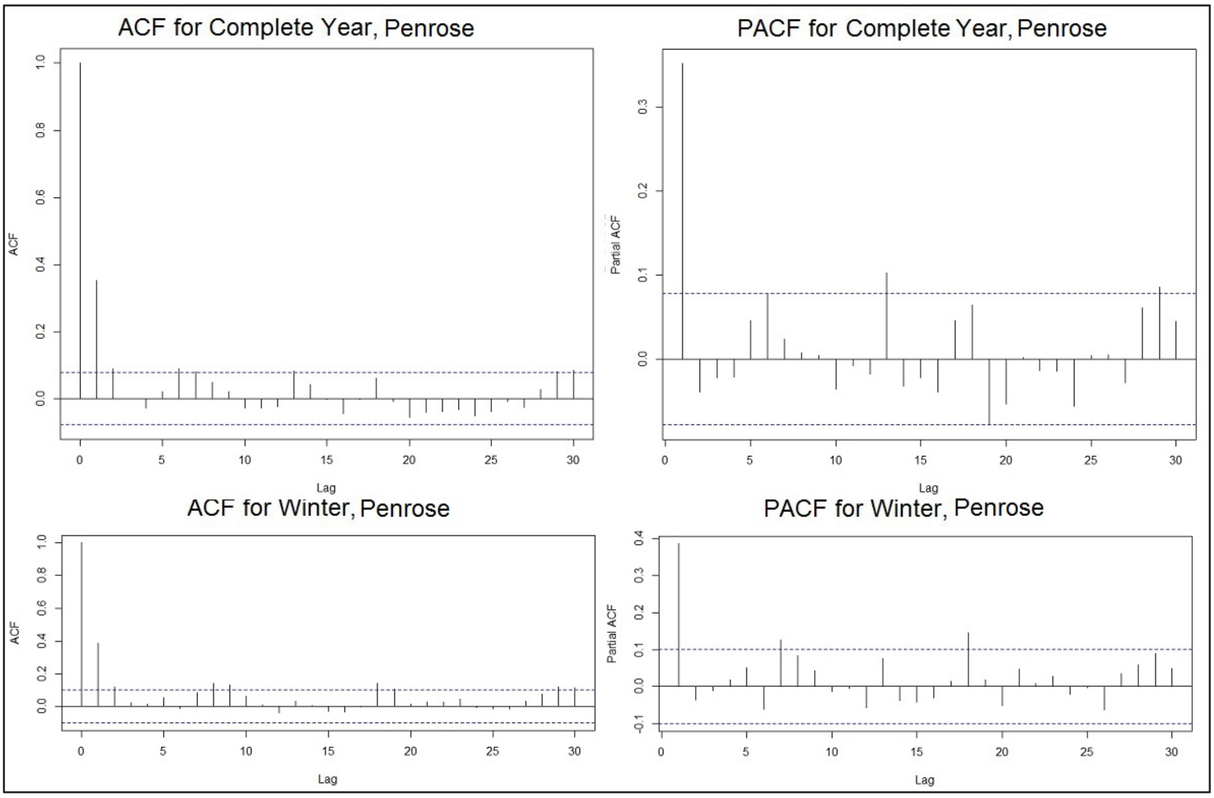


Figure 3.35: Auto-correlation Function of Penrose PM_{10}

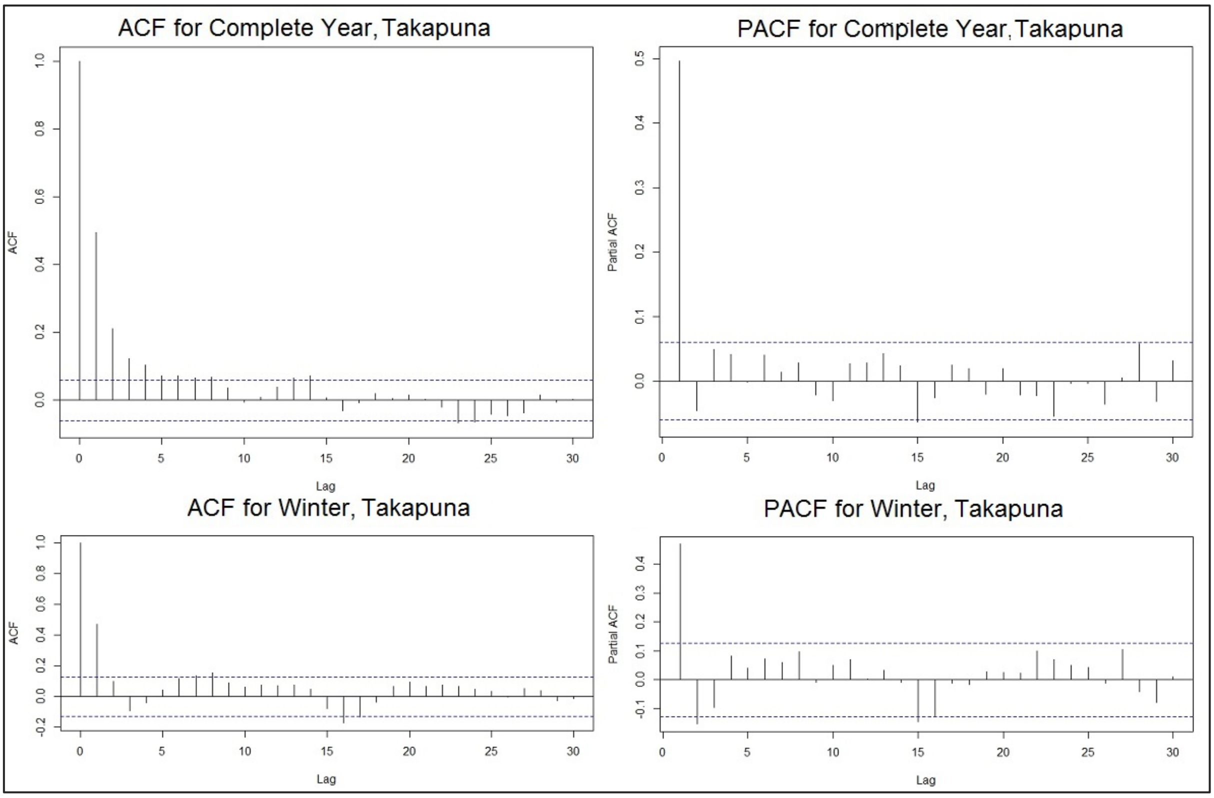


Figure 3.36: Auto-correlation Function of Takapuna PM_{10}

Because the PM_{10} at all this study's sites all exhibit autocorrelation, PM_{10} lag variables will be considered as potential predictors for the next day prediction (24 hours ahead) in future Chapters.

3.6 Long Term Trend Analysis

In this section, the Theil-Sen method (Sen, 1968; Theil, 1950) is used in the assessment of long-term trends in PM_{10} concentrations over the study period (2011-2016). The Theil-Sen estimator is known to be unaffected by outliers and tends to yield correct confidence intervals in non-normal data, which is the case for the PM_{10} data used in this study. The analysis of seasonal effects revealed significant seasonality therefore the Seasonal and Trend decomposition using Loess (STL) is applied prior to trend analysis. The Theil-Sen estimate of the slope is calculated as the median of all the slopes between all pairs of points. Ignoring mild autocorrelation (as reported in the previous section) would tend to give an optimistic impression of uncertainties. To overcome this issue, bootstrap simulations were carried out to account for autocorrelation. The estimator is nonparametric, which means that it does not draw from any probability distribution.

Trends in PM_{10} at the stations are presented in Figure 3.37 to Figure 3.41. The plots show the deseasonalised monthly PM_{10} concentrations. The solid red line shows the trend estimate. The 99.5% confidence intervals for the estimated trend using resampling methods are shown using dashed red lines.

The ∗∗∗ star symbols appearing next to a trend estimate, shown in green text at the top of the graph, shows that the trend is statistically significant at $p < 0.001$. Only Takapuna was found to have a statistically significant trend at the 0.001 significance level (Figure 3.42).

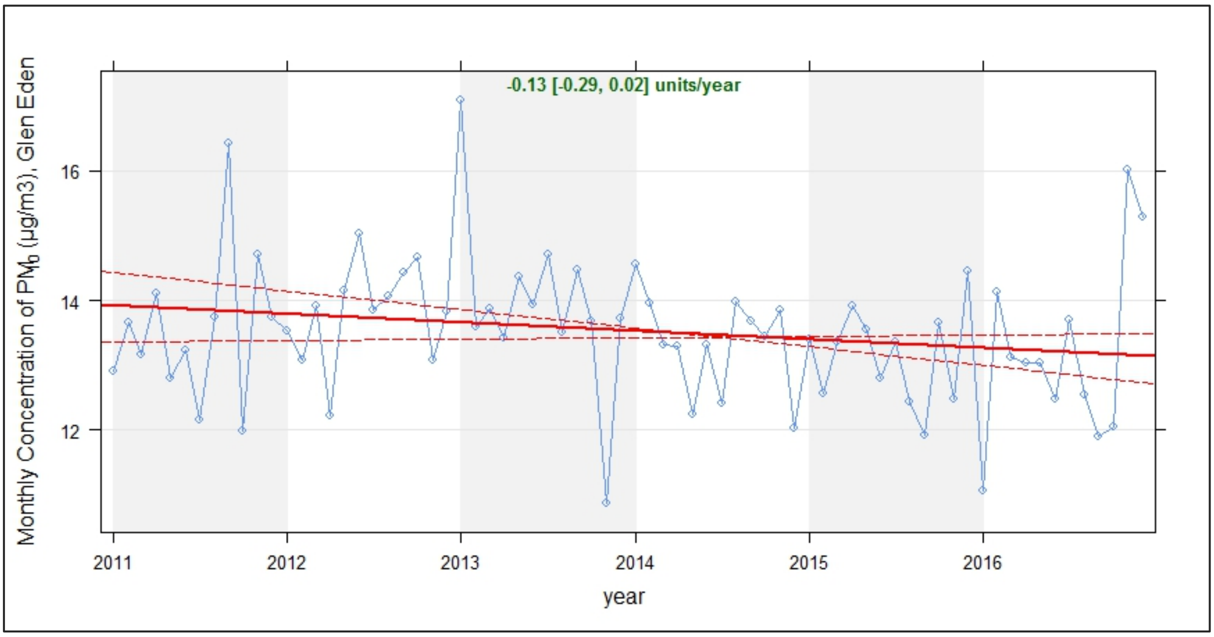


Figure 3.37: Deseasonalised monthly PM$_{10}$ concentration and trend line, Glen Eden.

Figure 3.37 shows the deseasonalised monthly concentrations of PM$_{10}$ in Glen Eden. The overall trend is negative (-0.13 µg/m^3) per year and the 95 % confidence intervals in the slope ranges between -0.29 to 0.02 µg/m^3/year.

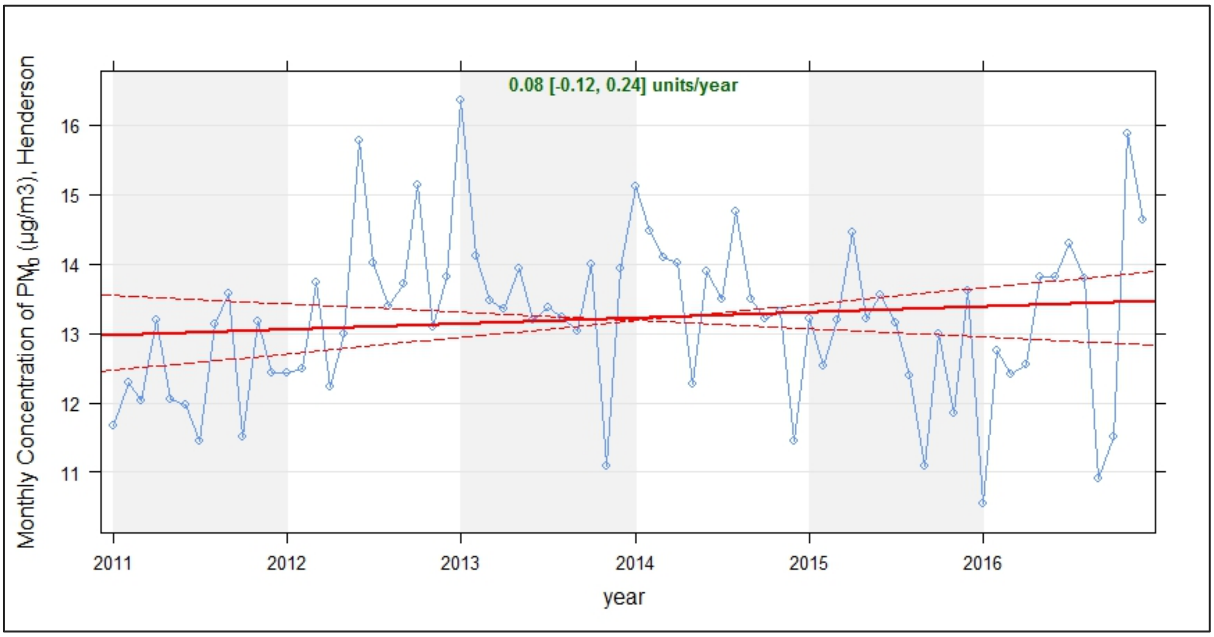


Figure 3.38: Deseasonalised monthly PM$_{10}$ concentration and trend line, Henderson.

A slight upward trend is observed in Henderson (Figure 3.38). The results found in this research agree with a source apportionment study at the Henderson site that also noted an increase in PM_{10} up until 2013. This increase was credited to an increase in biomass burning activities and vehicle emissions (Davy et al., 2017).

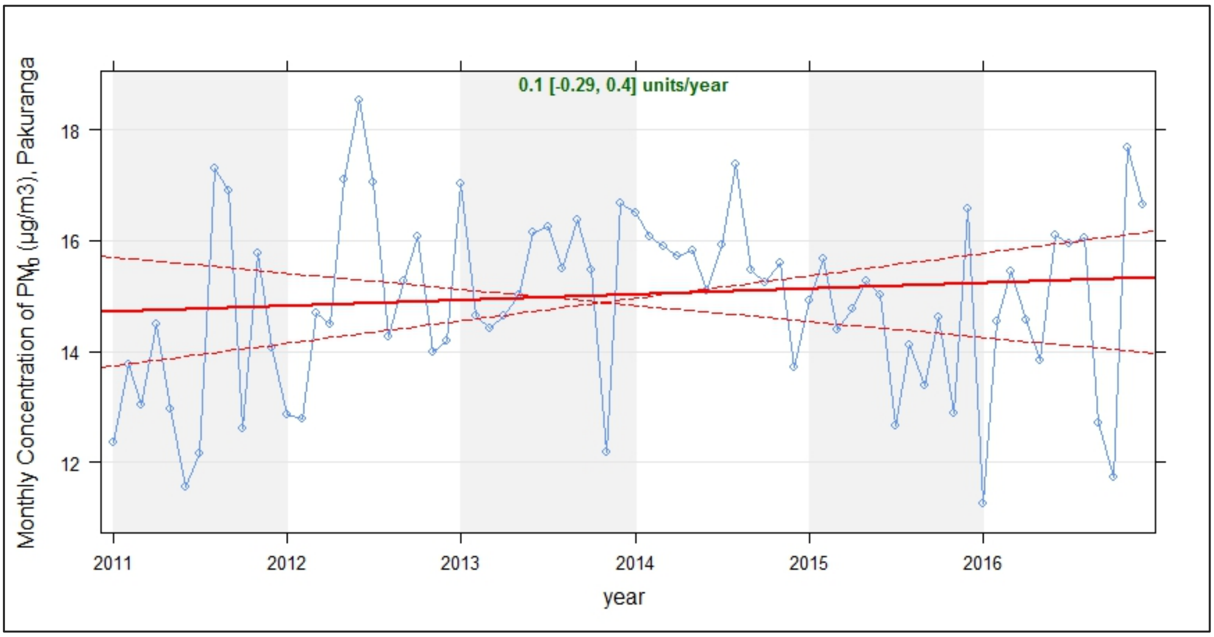


Figure 3.39 : Deseasonalised monthly PM_{10} concentration and trend line, Pakuranga.

The trend analysis for Pakuranga (Figure 3.39) shows no significant increase in PM_{10} concentrations during the study period.

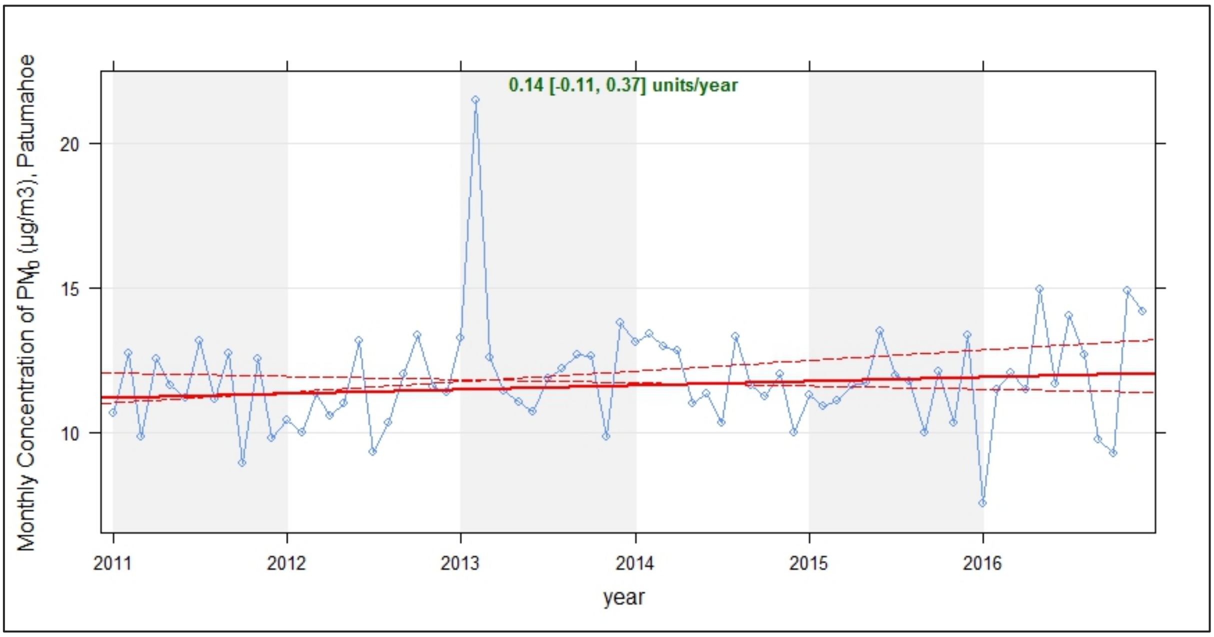


Figure 3.40: Deseasonalised monthly PM_{10} concentration and trend line, Patumahoe.

The rural site of Patumahoe, located in far south of the Auckland region, shows isolated peak PM_{10} incidences but overall, the trend is not notable.

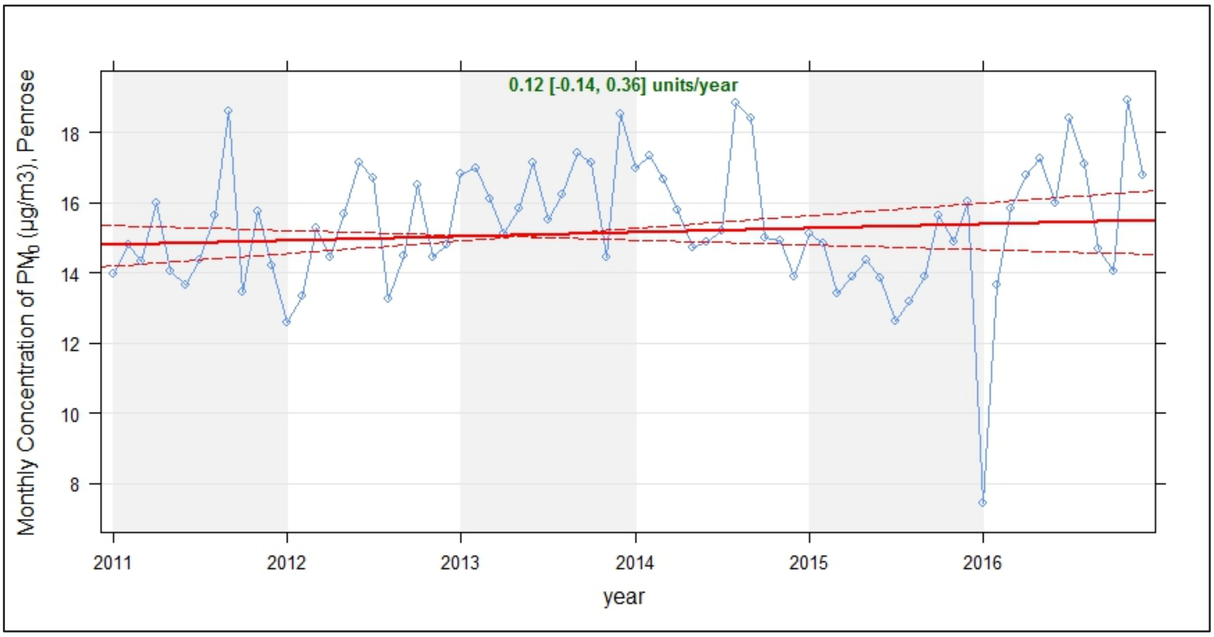


Figure 3.41: Deseasonalised monthly PM_{10} concentration and trend line, Penrose.

The trend at Penrose was not statistically significant (Figure 3.41). Decreases in contributions from motor vehicles, in secondary sulphate, and marine aerosol has been observed for 2007-2013 (Davey et al., 2017). However, the trend line from 2012 to late 2013 is almost unchanged but a slight increase can be seen from late 2013 onward (Figure 3.41).

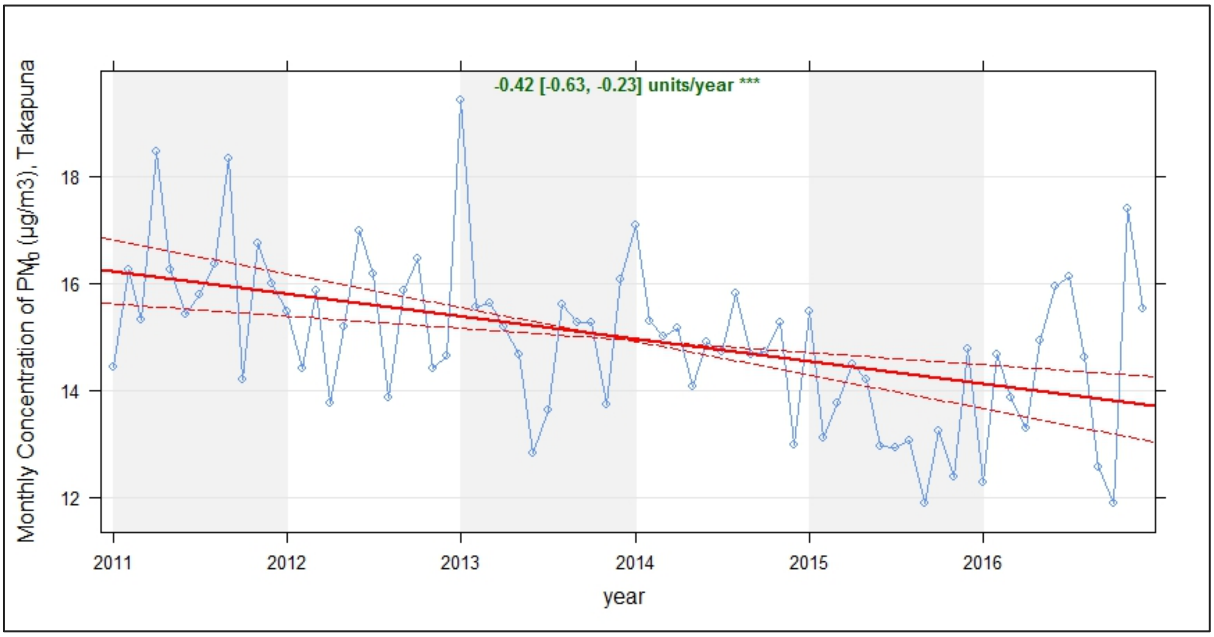


Figure 3.42: Trends in PM_{10} concentrations at the Takapuna site (statistically significant at the 99.9% confidence interval)

Figure 3.42 shows the deseasonalised monthly mean concentrations of PM_{10} in Takapuna. There is a significant (at the 0.001 level) downward trend in PM_{10} concentrations. The increasing trends seen at all sites in this work except for Takapuna disagrees with the findings of the report by (Talbot et al., 2017). This discrepancy could be due to their initial data processing and averaging processes which is likely to have resulted in an underestimation of their 24-hour daily average. Another possible contributor is the difference in the duration of observations used. However, the trend analysis result from this thesis's work agrees with Davy et al. (2017) who reported an increasing trend in PM_{10} concentrations which they attributed to an increase in biomass burning over the Auckland region. This research, while also extending the study period, confirms and gives weight to the findings of Davey et al. who explored data for a shorter period of PM_{10} concentrations at five of the six Auckland monitoring sites. It should be noted that although these trends were not statistically significant at the 0.001 threshold, a general upward trend in five stations is also visible in our study. The Takapuna PM_{10} monitoring site was the only site to exhibit a downward trend.

3.7 Conclusion

In Auckland six monitoring stations continuously collected both atmospherics (same location or nearby stations) and PM_{10} concentration during 2011-2016. PM_{10} concentration was collected on hourly basis and were averaged from midnight-midnight by AC for analytical purposes. As part of data cleansing and exploration, it was found that the 24-hour average data does not comply with WHO regulations for estimating the average value of air pollutant concentration. It was observed that there were 20 days of continuous missing data for Henderson during 2015. The number of peak PM_{10} events, those that were higher than 60% (30 µg/m^3) of the NES 24-hour average, ranged between 1 (Henderson) to 30 (Pakuranga) days amongst the monitoring sites. The 24-hour limit of 50 µg/m^3 was breached twice in Pakuranga (2012 and 2013) and once in Patumahoe (2013). The six monitoring stations used in this study have the geographical characteristics of urban residential, urban industrial, urban traffic/residential and rural/residential. A large difference in variability between these stations was observed using an estimated Pearson product moment correlation coefficient, showing a distinct correlation (84%) between Henderson and two other stations, namely Takapuna and Glen Eden. Patumahoe had the lowest correlation coefficient with other stations that can be explained by its rural/residential background. The lifetime of pollutant residence in the ambient atmosphere and the formation of secondary pollutants is typically controlled by atmospheric parameters.

There was a large amount of missing rainfall data for Henderson, Takapuna, and Penrose stations. There is not a priori basis for excluding rainfall variables from a model therefore satellite rainfall measurement for missing instances were obtained from the NOAA / NESDIS website.

High negative correlations between temperature and PM_{10} concentration were observed for all stations during colder months that can be attributed to use of wood burners. This was in alignment with previously reported findings of the effect of temperature on PM_{10} concentrations due to use of fuel usage and ambient chemical reactions (Czernecki et al., 2016; De Gooijer & Hyndman, 2006). In this dataset high PM_{10} concentrations also appeared to correlate with low temperature days, that can be due to use of wood burners during cold seasons

A study by Zhou et al. (2020) found that precipitation had a significant role on wet removal of PM_{10} in days with low relative humidity (~60%) and a lesser role in the near saturated relative humidity range (90–100%). In this study, Auckland rainfall appeared to have some site-specific predictive ability. This could be due to low variability of rainfall in these stations.

Based on the matrix plots created for this research, solar radiation appears to have limited predictive ability except in the case of the Glen Eden site.

Polar plots of wind direction and speed were created and analyzed with respect to PM_{10} concentration. In Penrose, the southwesterly quarter wind direction was found to be predominant causing high concentration of PM_{10} at southwest of this station. In Henderson, a southwesterly quarter wind direction is predominant showing highest PM_{10} concentrations when the wind is from the westerly directions aligning with marine areoles driven in from Tasman Sea and Pacific Ocean. The highest PM_{10} concentrations in Patumahoe are observed during occurrences of high wind speed when the wind prevails in the westerly directions. In Glen Eden, high PM_{10} concentrations to the northwest was aligns with high wind speeds. In Pakuranga, peak PM_{10} is observable in the southwest direction when calm to moderate winds is aligned with Pakuranga highway. In Takapuna, the moderate to strong winds from the west/southwest direction (aligns with the nearby motorway) causes the highest concentration observed. The easterly concentration is more likely attributed to emissions from ships operating at Port of Auckland as suggested by Davy (2017). The same study by Davy (2017), suggested that the marine aerosol from the Tasman Sea and Pacific Ocean are the primary source of PM_{10} concentration in the west-southwest and east-northeast directions, and are associated with incidences of high wind speed.

The NES target has been specified in terms of a daily average from midnight to midnight. This means that evening measurements will run into the next morning. To investigate the potential of this overlap in inducing correlation in PM_{10} concentrations between neighboring days, site-specific estimation of ACF and the PACF were generated and analyzed. As a result, PM_{10} at all this study's sites exhibited autocorrelation, suggestive of using PM_{10} lag variables for next day prediction.

Increasing trends seen at all sites in this work except for Takapuna disagrees with the findings of the report by (Talbot et al., 2017). This discrepancy could be due to their initial data processing and averaging processes which is likely to have resulted in an underestimation of their 24-hour daily average. Another possible contributor is the difference in the duration of observations used. However, the trend analysis result from this thesis's work agrees with Davy et al. (2017) who reported an increasing trend in PM_{10} concentrations which they attributed to an increase in biomass burning over the Auckland region. This research, while also extending the study period, confirms and gives weight to the findings of Davey et al. (2017) who explored data for a shorter period of PM_{10} concentrations at five of the six Auckland monitoring sites. It should be noted that although these trends were not statistically significant at the 0.001 threshold, a general upward trend in five stations is also visible in our study. The Takapuna PM_{10} monitoring site was the only site to exhibit a downward trend.

As Auckland is a growing city in terms of population and density as well as the road traffic therefore the number of people exposed to locally emitted particulate matter will increase. For 2011-2016, most sites within the Auckland airshed complied with the NES-AQ requirement of one or fewer exceedances of the 24-hour average PM_{10} standard per year with the exception Pakuranga site during 2015. Given the number of days with short-term peak PM_{10} concentrations, there are possibilities of breaching the exceedances limits to happen. Therefore, efforts to reduce the PM_{10} peaks should be taken by reducing the use of wood/coal burners as well as introducing Clean Car standards to reduce the traffic related emissions.

Chapter 4 TIME SERIES ANALYSIS OF PM_{10} CONCENTRATION

A general approach to time series analysis is provided in this Chapter to form the foundation for in the work presented in the remaining chapters. In this Chapter, the concept of dependence and stationarity is introduced prior to time series analysis and modeling approaches. Section 4.2 introduces details of the challenges encountered for time series where the seasonality is complex with multiple levels and reviews the relevant literature in the area. It should be noted that spatial features are not considered in the work presented in this Chapter, this work only investigates the effect of time on PM_{10} concentrations. The effect of time and space will be investigated in Chapter 6.

4.1 Introduction

Time series analysis aims to detect the nature of an event described by the structure of observations, and forecasting (StatSoft, 2013). To analyze characteristics of a time series (such as trend, seasonality and cycles) quantitative methods can be used (Wang & Chaovalitwongse, 2011). Linear Regression and Artificial Neural Networks (ANNs) are categorized as causal quantitative methods, where predictions are made using relevant influential factors. Moving Average (MA), Exponential Smoothing (Wang, 2012) , Box-Jenkins (Box & Jenkins, 1990), State Space (De Gooijer & Hyndman, 2006) and Spectral Analysis, (Wang, 2012) are categorised as Non-causal methods.

Since the time series of PM_{10} investigated in this research have complex seasonality (Section 5.6.5) it is necessary to explore recent, novel models that are designed to cope with such complexity. Thus, Section 4.7 of this Chapter describes in detail two new forecasting models, Harmonic Regression and TBATS, for seasonality complex time series. Performance of these two forecasting methods is compared and evaluated in 4.9.

4.2 ARIMA and SARIMA models in air-pollution forecasting

Autoregressive Integrated Moving Average (ARIMA) (Box and Jenkins, 1976) models are a type of linear model able to represent both stationary and non-stationary time series. Independent variables are *not* used in their construction only the variable of interest. These methods make use of the patterns in the time series itself to construct a model and are thus dependent on autocorrelation patterns in the model. Unlike most forecasting models, ARIMA models do not assume a certain pattern in the historical data of the time series to establish the forecasting model (Adhikari & Agrawal, 2013).

The Box-Jenkins methodology (Box & Jenkins, 1990) provides a number of procedures for identifying, fitting and verifying ARIMA models. Forecasts are then made based on the form of the fitted ARIMA model. This methodology is comprised of three key stages: identification, estimation and testing, and model application (Adhikari & Agrawal, 2013).

In ARIMA model application of finite differencing of the data points make a non-stationary time series stationary. Section 4.3 discusses the concepts of stationarity and non-stationarity of a time series and some common approaches to transforming a non-stationary process into a stationary one. The Seasonal ARIMA (SARIMA) model was proposed by (Box & Jenkins, 1990) to deal with series containing seasonal fluctuations. In the SARIMA model seasonal differencing is used to transform a non-stationary model to a stationary one.

Both the ARIMA and the SARIMA models have been used in air pollution studies to forecast different air pollutants. A SARIMA model was used to forecast CO and NO_2 concentrations in Malaysia (Ibrahim et al., 2009). Another study reported very good performance for short-term predictions of ozone and PM_{10} (72 hours ahead) in Blagoevgrad, Bulgaria using SARIMA models (Gocheva-Ilieva et al., 2014). In a follow-on study the same research team developed high performance SARIMA models for forecasting concentrations of PM_{10} and sulfur dioxide (SO_2) that included meteorological variables. These models were for the town of Kardzhali in Bulgaria and for 24, 48 and 72 hours in advance predictions. These researchers reported that SARIMA models built on transformed time series demonstrate better statistical performance with a coefficient of determination of up to 88% for SO_2 and 90% for PM_{10} (Doychin et al., 2015).

ARIMA models were found to outperform SARIMA models in predicting the Air Quality Index (AQI) between 2012-2015 for Kerala in India (Naveen & Anu, 2017).

A hybrid wavelet-ARMA/ARIMA model was used to forecast a PM_{10} time series for Taiyuan, China (Zhang et al., 2017). The hybrid model was found to effectively reduce forecasting error when compared with the ARMA/ARIMA method.

In an attempt to improve on existing PM forecasts, a deterministic decomposition model was built based on CO concentration measurements from time series analysis (Guarnaccia et al., 2014). The deterministic model captured the average slope of the CO concentration with low mean error but performed poorly in predicting the local variations and fluctuations of CO.

A more recent study on PM_{10} modeling and forecasting in Bulgaria, fitted SARIMA models to historical data in an attempt to provide longer term, 120 hours (five days) forecasts (Gocheva-Ilieva & Ivanov, 2019). The PM_{10} timeseries had no hourly trend however a 24-hour cycle was detected. The SARIMA model which used meteorological factors (wind speed, temperature, and pressure) fitted well to historical data (R^2= 90%, RMSE=0.114). The R^2 for the five-day forecast was reported to be 0.507

indicating that only 50% of the variance can be explained when the model is used forecasting. Another SARIMA model without meteorological factors showed an R^2=0.888 and R^2=0.195 for fitting to historical data and forecasting five days ahead respectively.

4.2.1 Definition of the Box-Jenkins methodology, ARIMA and SARIMA Models

In general, the Box-Jenkins methodology is a framework for the development of ARIMA models. It is this methodology which is followed in this research. The four-steps of the iterative procedure specified in the Box-Jenkins methodology are (Hu, 2008):

i. **Model Identification**: Use the ACF and the PACF of the stationary data series to identify a suitable Box Jenkins Model.
ii. **Model Estimation**: Use historical data to estimate the model's parameters.
iii. **Model Diagnosis**: Check the adequacy of the model using different diagnostics tests such as residuals' autocorrelation. Find an improved model if necessary and treat it as the new identified model.
iv. **Forecasting**: Use the final model to forecast future time series values.

In ARIMA (p, d, q) model, p, d and q denote the number of autoregressive terms, number of differences and number of moving average terms, respectively. The model is comprised of the following three components:

i. **Autoregressive (AR) Model**: A regression model that uses the dependencies between an observation and its lag(s).
ii. **Integrated (I)**: In non-stationary time series the non-stationary pattern is removed to ensure that other correlation structures in the series can be seen prior to model building. This can be achieved by calculating the differenced time series.
iii. **Moving Average (MA)**: An approach that takes the dependency between observations and the residual error terms into account.

The significance of the individual parameters in an ARIMA model are computed using the ACF and the PACF of the appropriately transformed/differenced series. According to Table 4.1 given by Janacek (as cited by Fauzi Raffee, Abdul Hamid et al. (2018)), a model has an appropriate $MA(q)$ process if the autocorrelations are zero after lag q. The model is an $AR(p)$ if the decay is exponential. In mixed model $ARMA(p, q)$, correlations will shrink after $lag\ (p - q)$.

Table 4.1: Behavior of the auto and partial correlation function.

	ACF	**PACF**
$AR(p)$	Exponential decay	Zero after lag p
$MA(q)$	Zero after lag q	Exponential decay
$ARMA\ (p,q)$	Exponential decay after lag (p-q)	Decay after $lag\ (p-q)$

With seasonal time series, both non-seasonal and seasonal factors are incorporated in a multiplicative model. This process is called SARIMA and is denoted as $ARIMA(p,d,q) \times (P,D,Q)s$ where p, q and d, are order of AR, MA and difference respectively. $P, Q\ and\ D,$ are AR order of the seasonal process, seasonal MA, and seasonal difference. The number of time steps in a single seasonal period is denoted by s.

Most existing time series models such as the SARIMA and other Error Trend Seasonality (ETS) models can typically handle simple seasonal patterns with a small integer-valued period or seasonality at single levels. Some methods have been developed for producing forecast models for time series with two seasonal patterns (Pedregal & Young, 2008) but these methods are unable to cope with more than two seasonal patterns and are not able to accommodate for the nonlinearity found in air pollution time series (Foxall et al., 2001; Marra et al., 2003).

The existing exponential smoothing models (Taylor, 2003; Taylor & Snyder, 2012) perform poorly when modeling time series with complex multiple levels of seasonality (De Livera et al., 2011) such as exhibited by Auckland's PM_{10}. The issues seen when adopting such methods for complex seasonal time series include :over parameterization, the failure to adapt both non-integer period, and dual calendar effects (De Livera et al., 2011). Forecasting problems involving high frequency time series data with complex multiple levels of seasonality investigated in literature are related to time series other than air pollution such as hourly electricity loads (Baek, 2008; Fan & Hyndman, 2015), gasoline supply, bank visitors and electricity demand (De Livera et al., 2011). According to Baek (2008), extremely wide-ranging intervals are not appropriate for hourly electricity load prediction. The TBATS method proposed by De Livera et al., (2011) is claimed to be able to identify and extract hidden seasonal components in timeseries and can address the above-mentioned nonlinearity problem by using Box-Cox transformation. TBATS models are discussed in section 4.9.2 and used for modeling Auckland PM_{10} concentration. To the author's knowledge TBATS has not been previously applied for modelling of PM_{10} time series with complex seasonalities.

4.3 Testing for (non)Stationarity

The assumption of strict stationarity is that the statistical properties of the space-time are constant over time or between locations. The assumption of stationarity is often too strict (Storch & Zwiers 2002) and difficult to be confirmed, so second-order stationarity is adopted in this research. Under the second-order stationary assumption the expected value (mean) of random function is constant over the area and its space-time covariance function depends only on the spatial and temporal separation of points (Jentsch & Subba Rao, 2015). Since a Gaussian process is entirely specified by its mean and variance, the strict stationarity and second order stationarity are similar (Bruno et al., 2009a).

In practice, many data sets do not meet the *second order* or *intrinsic* stationarity assumption. This problem is summed up by quote from Thompson (1994, p. VI-74) and still holds today (Thomson, 1994):

> "*Experience with real-world data, however, soon convinces one that both stationarity and Gaussianity are fairy tales invented for the amusement of undergraduates*".

Examples of non-stationary processes are random walks with or without a drift and deterministic trends. In a pure random walk, the value observed at time *t* is the value of last period plus an independent and identically distributed (mean of zero) and variance (σ^2) white noise stochastic component. A pure random walk can be a process of some order, with unit root or stochastic trend. According to Box and Jenkins (1990) differencing should be applied to homogeneous nonstationary sequences to make a process *difference* stationary. In this method, the difference of consecutive terms in the series is computed. Differencing is typically performed to remove a varying mean.

A time series that can be made strict stationary by differencing is considered to be difference stationary. In time series with a deterministic trend, detrending can be applied to remove the trend and drift. However, the variance will continue to go to infinity. Hence, applying differencing will remove the trend in the variance. Sometimes a non-stationary series may combine a stochastic and deterministic trend at the same time. In such cases, differencing and detrending together can avoid obtaining misleading results. The trend in the variance will be removed by differencing and detrending will remove the deterministic trend. A random walk with a deterministic trend can be transformed by detrending (Durlauf & Peter, 1988).

There are several methods that can be used to identify the stationarity /non-stationarity of data based on either the *unit root* hypothesis or on the *stationary null* hypothesis. *Unit root* indicates that the statistical properties of a given series are not constant with time this means the data contains a systematic pattern that is unpredictable. Early and pioneering work on testing for a unit root in time series was undertaken by Dickey and Fuller (1976). The Dickey-Fuller test is testing if $\theta = 0$ in this model of the data (Dickey & Fuller, 1976):

$$y_t = \propto + \beta t + \theta y_{t-1} + e_t \qquad \text{Eq. 4.1}$$

which is written as:

$$\Delta y_t = y_t - y_{t-1} = \propto + \beta t + \gamma y_{t-1} + e_t \qquad \text{Eq. 4.2}$$

where yt is the time series.

A linear regression of Δy_t against *t* andy_{t-1} is used to test if γ is nonzero. If γ = 0, then the process is random walk and non-stationary. If not and $-1 \; < 1 + \gamma \; < 1$, then the process is stationary (Holmes et al., 2019). The Augmented Dickey-Fuller (ADF) test is the basic autoregressive unit root test augmented to accommodate ARMA(p, *q*) models with unknown orders (Said & Dickey, 1984). The null hypothesis in ADF is that a time series $y_{t,}$ with ARMA structure, is $I(1)$ and the alternative hypothesis is that it is $I(0)$. Large p-values (p-values > 0.05) are indicative of non-stationarity. We fail to reject the *null hypothesis*, if the test statistic is larger than the critical values. The ADF test is also known as a difference stationarity test. Using the usual 5% threshold, differencing of the time series is required if the ADF *p*-value is greater than 0.05.

The Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test (Kwiatkowski et al., 1992) tests the null-hypothesis that the process is stationary around a mean or a deterministic trend (i.e. trend-stationary), the alternate hypothesis is that of a first difference stationary time series. Assuming a time series; $t = 1,2,\ldots,N$; can be written as summation of deterministic trend (β_t), a random walk(r_t), and a stationary error (ε_t), using the regression equation (Das & Bhattacharya, 2014):

$$x_t = r_t + \beta t + \varepsilon_t \qquad \text{Eq. 4.3}$$

The series is trend stationary if the intercept is a fixed element. In a level stationary, the series is stationary around fixed level, the null hypothesis would be $\beta = 0$ (Wang, 2006). If the test statistic is greater than the critical value, then the null hypothesis is rejected; the series is non-stationary. Table 4.2 shows the alpha values for critical values of 10%, 5% and 1%.

Table 4.2: Table of KPSS critical values (Kwiatkowski et al., 1992).

	Critical Values		
	0.1	**0.05**	**0.01**
Intercept Only	0.347	0.463	0.739
With Linear Trend	0.119	0.146	0.216

4.3.1 Identifying (non)Stationarity in Auckland's PM$_{10}$

The ADF test was applied to the daily PM$_{10}$ time series using the R 'urca' package (Pfaff et al., 2016). To include both trend and drift, the ADF test was performed with trend and the test statistics present in the time series. The test statistics were compared against the critical values presented in Table 6.3.

Table 4.3: Critical values for ADF test statistics

	Critical Values		
	0.1	**0.05**	**0.01**
tau3	-3.96	-3.41	-3.12
phi2	6.09	4.68	4.03
phi3	8.27	6.25	5.34

Test statistics greater than the critical values indicate that null hypothesis was not rejected. Failing to reject the $tau3$, means there is unit root $(\gamma = 0)$ and the process is a random walk. The null hypothesis for $phi2$ implies a unit root $(\gamma = 0)$, no trend $(a_t = 0)$ and no drift$(a_2 = 0)$. The $phi3$ hypothesis implies $\gamma = a_2 = 0$. $P - values$ of less than 0.05 means that the $phi3$ null hypothesis is rejected meaning at least one of these two terms are not zero.

Phi2 tests the hypothesis that there is a unit root, no time trend, and no drift term. The results for the *phi2* test on daily mean PM$_{10}$ time series for Auckland by station are presented in Table 4.4. For all sites, the *phi2* value (T-statistic) is greater than the corresponding 1% critical value of 4.03. Thus, the ADF test results indicate that the null hypothesis should be rejected in favor of the alternative hypothesis that the time series are stationary around a deterministic linear time trend.

Table 4.4: Results of ADF tests on daily PM_{10} time series.

Site	**Type=" trend"**	
	p-value***	**T-statistic**
Glen Eden	0.01	36.48
Henderson	0.01	47.63
Pakuranga	0.01	42.98
Patumahoe	0.01	59.57
Penrose	0.01	48.29
Takapuna	0.01	53.84

As discussed in above section *unit root* is one of the reasons that a process exhibits non-stationarity. Therefore, it is possible for a time series to be non-stationary without having a *unit root* and for the ADF to indicate that the time series is trend stationary.

The KPSS test was performed using R's VGAMextra package (Miranda & Yee, 2018). The outcomes of the KPSS tests on the same PM_{10} time series (see Table 4.5) indicate that considering the 'intercept' alone, the time series at all sites are level stationary. The results also show three of the six sites are stationary around a deterministic trend while the rest reject the null hypothesis of trend stationary in favor of the presence of *unit root*.

Table 4.5: Results of KPSS tests on daily PM_{10} time series.

	Level Stationary		**Trend Stationary**		
Site	***p-value***	**T-statistic**	***p-value***	**T-statistic**	***Conclusion***
Glen Eden	0. 24	0.169	0.24	0.081	Trend stationary
Henderson	0. 23	0.181	**0.02**	**0.184**	**Unit root**
Pakuranga	0. 24	0.173	**0.03**	**0.168**	**Unit root**
Patumahoe	0. 32	0.103	0.41	0.053	Trend stationary
Penrose	0. 10	0.229	**0.01**	**0.22**	**Unit root**
Takapuna	0. 10	1.53	0.30	0.07	Trend stationary

For the KPSS test, we failed to reject the null hypothesis of stationarity around a deterministic trend where *p-values* are greater than 0.05 and the test statistics are smaller than the critical value (0.146) in Glen Eden, Patumahoe and Takapuna. The null hypothesis is rejected in the Henderson, Pakuranga, and Penrose time series where the respective T-statistic is greater than the 5% critical value (Table 4.3).

After discussing with one of the package authors (Miranda & Yee, 2018), it was concluded that the contradictory results observed could be due to the presence of multiple seasonality in time series. A possible explanation would be that unlike KPSS, the ADF test did not observe the seasonality in the time series and thus the ADF results for all stations indicated stationarity when in fact that is unlikely.

4.3.2 Autocorrelation Functions

In autocorrelated time series, the time series itself is correlated with its lagged (1, 2, or more periods) time series. Computing the sample autocorrelation (covariance) function (sampleACF) of the observed data helps to assess the data dependency and to select a model that reflects this dependency. Assuming that the data are realized values of a stationary time series $\{X_t\}$, then the sampleACF provides an estimate of the ACF of $\{X_t\}$. If the sample ACF, denoted as ($|\rho\hat{}\ (h)|$), exhibits a slow decay as h increases this is an indicator of data containing a trend. Moreover, for data with a significant deterministic periodic component, the sampleACF function will show similar performance within the same periodicity. As a rule of thumb, the sample ACF and PACF can be used as ACF and PACF representative of a stationary process for lags no more than a third of the sample size.

Figure 4.1 to Figure 4.6 show the empirical autocorrelation plots for the PM_{10} time series, at each station, before any differencing is performed. The autocorrelation analysis, made by means of correlograms, suggest that the time series are not independent. It is notable from the ACF plot that autocorrelations are significant for many lags. However, the autocorrelations observed at lag_2 and above maybe due to propagation of the autocorrelation at lag_1. This was confirmed by the sharp cutoff observed in the PACF plot between lag_1 and lag_2. The slow decay observed in the ACF plots is sign of trend in time series and can be a useful indicator of non-stationary.

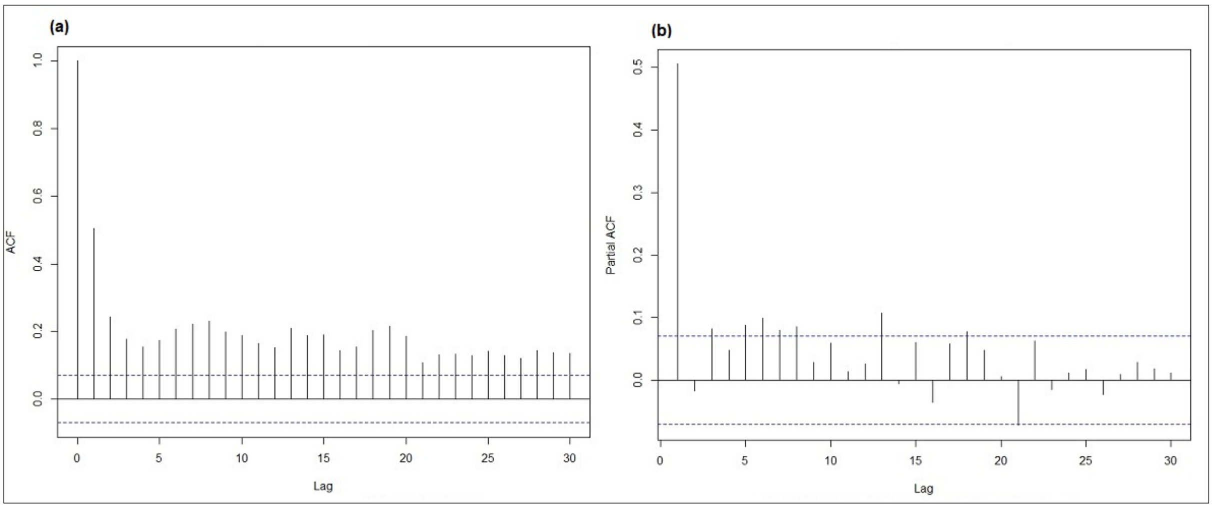


Figure 4.1: ACF plot (a) and PACF plot (b) of daily average PM_{10} in Glen Eden**.**

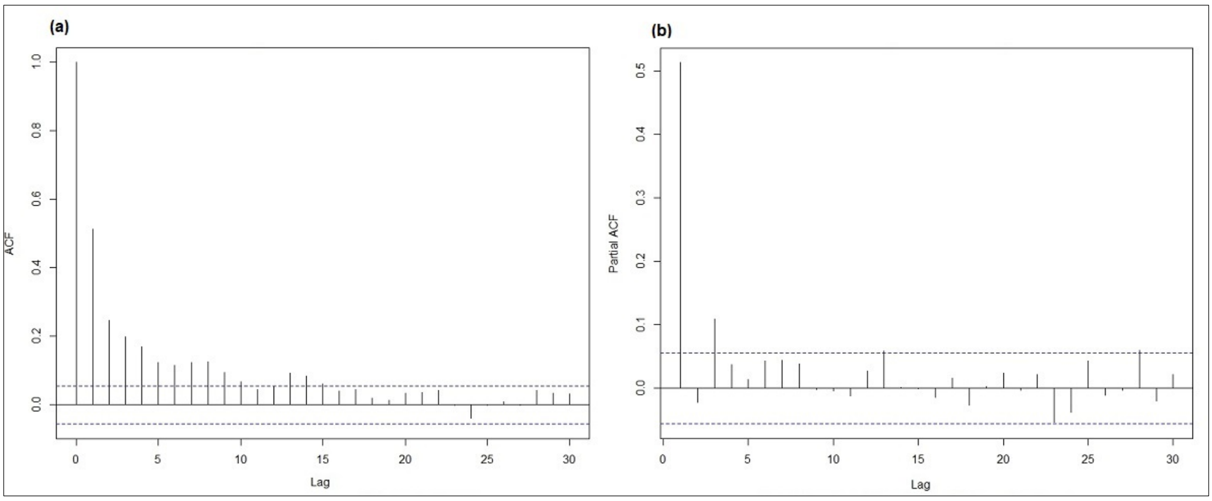


Figure 4.2: ACF plot (a) and PACF plot (b) of daily average PM_{10} in Henderson.

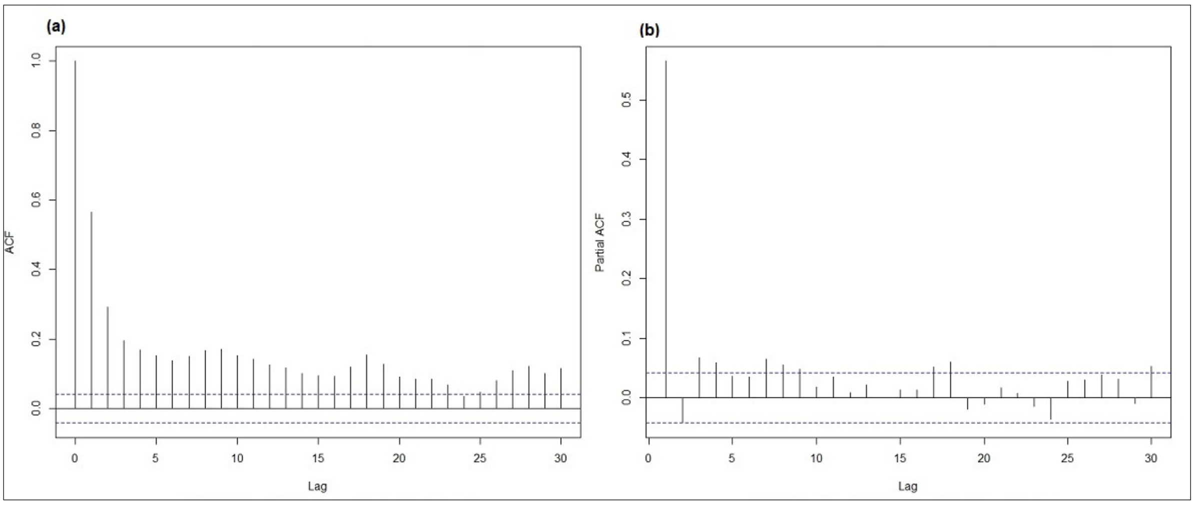


Figure 4.3: ACF plot (a) and PACF plot (b) of daily average PM_{10} in Pakuranga.

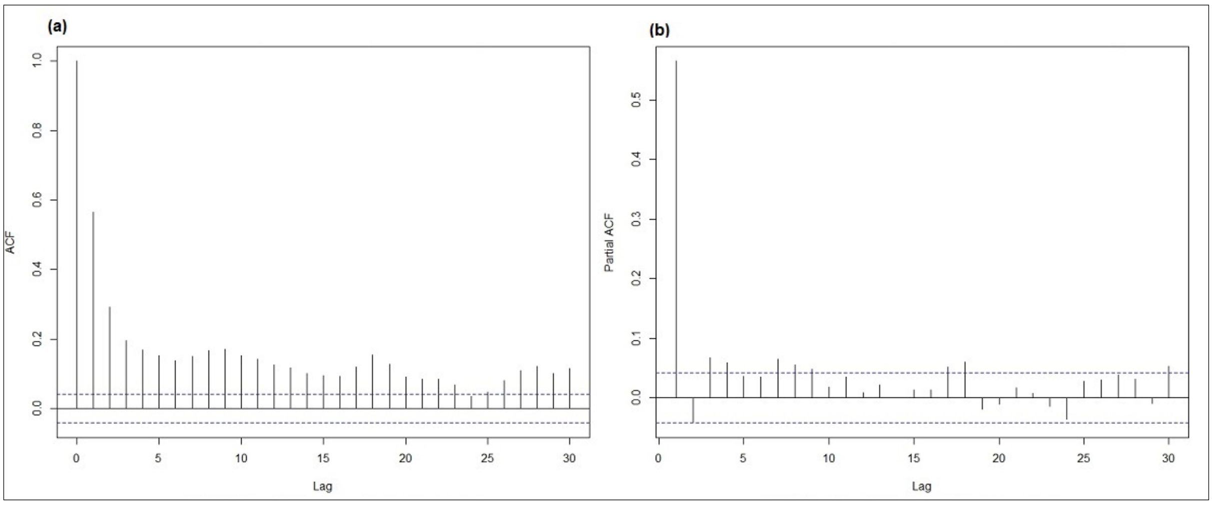


Figure 4.4: ACF plot (a) and PACF plot (b) of daily average PM_{10} in Patumahoe.

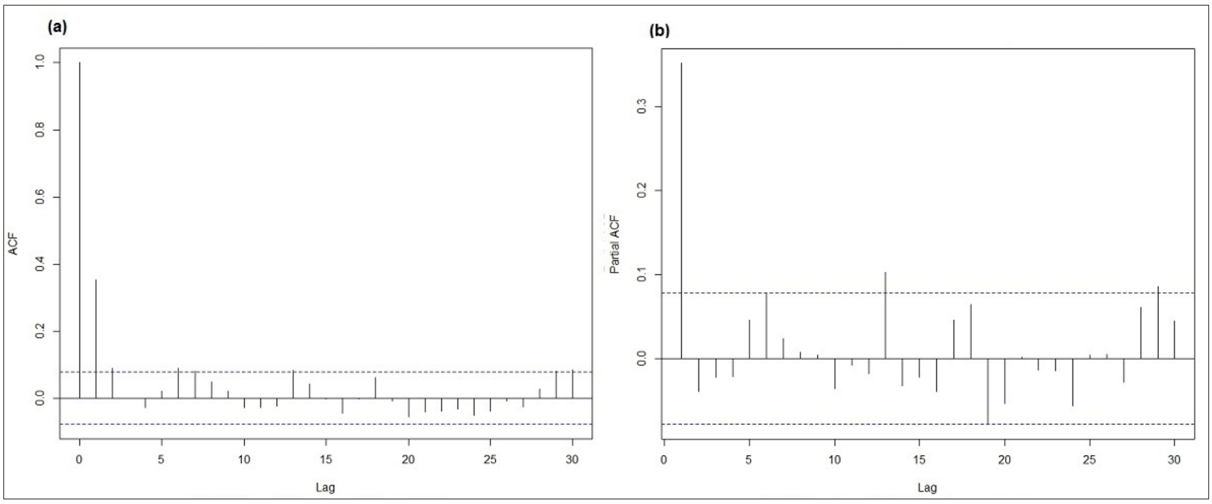


Figure 4.5: ACF plot (a) and PACF plot (b) of daily average PM_{10} in Penrose.

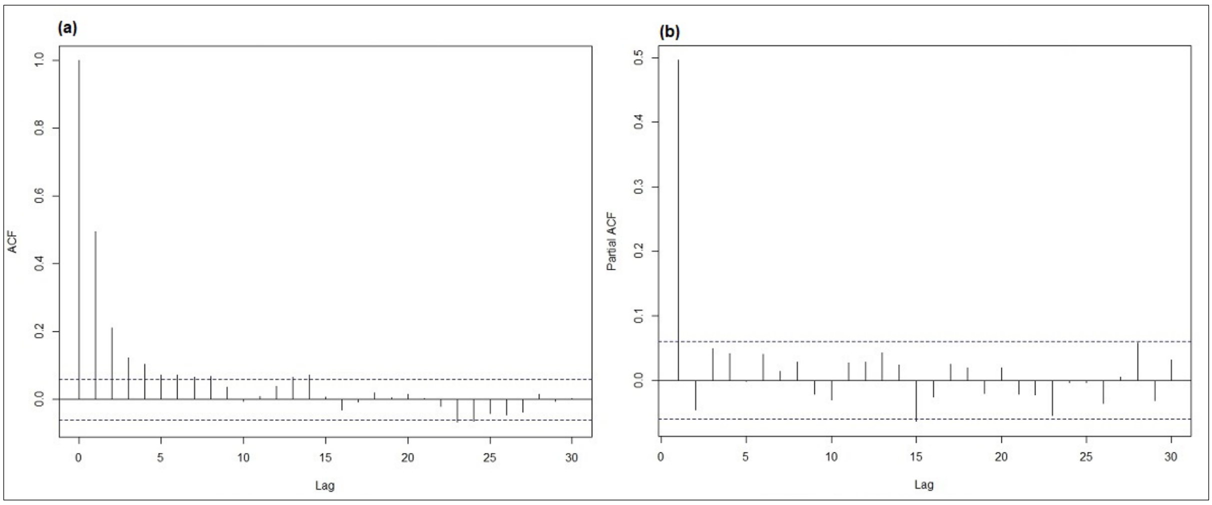


Figure 4.6: ACF plot (a) and PACF plot (b) of daily average PM_{10} Takapuna.

Approximately 95% of the sample autocorrelations lie between the $\pm 1.96/\sqrt{n}$ bounds, indicated by the dashed horizontal lines in the plots in Figure 4.2 to Figure 4.6, where n is the number of observations in the series. Because 95% of the sample autocorrelations are within these bounds the autocorrelation observed may be attributed to white noise. However, large, and repeating deviations from the bounds is indicative of the need for model to explain the autocorrelation dependency. As a rule of thumb, an $MA(q)$ model is recommended if the sampleACF of a stationary series falls between the plotted bounds $\pm 1.96/\sqrt{n}$ for lags $h > q$, while if the sample PACF of stationary series falls between the plotted bounds $\pm 1.96/\sqrt{n}$ for lags $h > p$, then an $AR(p)$ model is suggested (Brockwell & Davis, 2002). The sharp cutoff observed in the PACF plots between lag_1 and lag_2 and the gradual decrease in ACF values in the ACF plots is suggestive that the autocorrelation pattern for Auckland's PM_{10} concentrations may be explained better by adding AR terms rather than MA terms.

4.4 Time Series Models

The combination of one or more of the following typical components can be detected in time series data (A. M. Denham, 2012) :

1. **Trend**: a long-term increase or decrease in the mean that may appear in a linear or curvilinear form. The data is stationary if no trend is detected.
2. **Seasonal effects**: variations with weekly, monthly, or annual episodes. These intra-year fluctuations are almost stable year after year with regards to the time, direction, and scale of the effect. Seasonality of a time series can reveal normal variations that repeat every year (e.g., high exceedance of air pollution during winter months) or calendar related systematic effects.
3. **Cyclic changes**: variations occurring over a fixed period. Cyclic changes usually last longer than a year and are typically due to some physical influence rather than seasonal effects.
4. **Residual or error fluctuations or noise**: the unpredictable changes within a time series that do not follow a certain pattern.

To overcome the problem of non-stationarity caused by trend and seasonality, it may sometimes be necessary to apply a preliminary transformation to the data to produce a stationary series. This process as mentioned previously allows other correlation structures in the data to be identified prior to model building. One such method, differencing to remove seasonal patterns has already been discussed in Section 4.3. In this case the residual component is the data component that is removed from the data during the differencing transformation. Trend or regular pattern in residual indicates features which have not been attributed to the other components.

Trend and seasonality removal can be performed by estimation and subtraction of the trend/seasonal components from the data (Brockwell & Davis, 2002). There are two distinct purposes for performing such a decomposition of a time series. Firstly, to provide a summary description of time series significant features. The second purpose is prediction of the future values by forecasting the residuals and then inverting any data transformations, such as differencing, to get a forecast based on the original (untransformed) time series.

Additive models and multiplicative models are two classes of time series decomposition models. The decision of which type to use is a critical aspect of performing the analysis of a time series from the point of view of calculating the seasonal component. The trend, seasonal, and random components of a time series can be described using an additive model as follows (Hyndman, 2018):

$$Y(t) = T(t) + C(t) + S(t) + R(t) \qquad \text{Eq. 4.4}$$

Where

$T(t)$ is the long-term trend,

$C(t)$ is cyclic changes,

$S(t)$ is seasonal effects and

$R(t)$ is residual.

The assumption in additive model is that the difference between the trend and observed data is almost constant in similar periods of time (months, quarters) regardless of the trend tendency. This model is applicable when the seasonal component changes with the variations in the trend. The components are linked through multiplication as (Hyndman, 2018):

$$Y(t) = T(t) * C(t) * S(t) * R(t) \qquad \text{Eq. 4.5}$$

An alternative approach is to transform the data to stabilize the variation in the series over time, then use an additive decomposition (Hyndman, 2018).

A trend can be linear or curvilinear, cycles can have different durations and might appear at irregular intervals (Gaynor & Kirkpatrick, 1994). Therefore, modeling seasonality is easier than trend or cycle modeling as seasonality has an obvious frequent pattern (Dokumentov & Hyndman, 2015). A trend or cycle component may or may not be presented in a seasonal series. The effect of seasonality is assumed to be constant when performing a deterministic analysis of a time series. In the stochastic analysis method, moving seasonality is considered, thus avoiding the possible under/over correction made by a fixed seasonal pattern. Seasonally adjusted series can be useful if the variation in seasonality is not of primary interest. A seasonally adjusted time series is calculated by deducting or dividing the seasonal effect value of the given period from the initial time series. As a result, the obtained time series comprises the trend/cycle and random components. The next section of this Chapter outlines several different approaches to removing seasonality from Auckland PM_{10} time series.

4.5 Seasonally Adjusted Time Series

4.5.1 Decomposition of PM_{10} for the Auckland area using STL

The classical decomposition method (Persons, 1919) still forms the basis of many time series decomposition methods (Dagum, 2010). In classical decomposition, MA method is used for trend-cycle estimation. However, using this method means that a trend-cycle estimate for a specified number of first and last observations are not obtainable. In addition, the trend-cycle estimate has a tendency toward over-smoothing the data thus removing possible significant rapid changes in the data. In classical decomposition methods it is assumed that the seasonal component recurs in successive years which is not a sensible assumption for some longer series (Hyndman, 2018). Unusual observations can occur for a period which is relatively short in relation to the overall time series and affect the seasonal patterns during that time. For example, a building being constructed would result in not only significant PM_{10} readings during groundworks but also higher than normal air pollutant values being recorded from the increase in traffic volume due to construction vehicles (Chapter 3 describes such a situation at the Takapuna site during construction of new sports facility for a nearby school). Such an affect is a one-off event over a single period within the time series. Another PM_{10} example is the reduction in traffic related PM_{10} level during the COVID-19 lockdown, a situation that is different from the norm (Chapter 3). The classical method is not robust to these kinds of unusual values (Hyndman, 2018).

The STL method developed by Cleveland et al. (Cleveland et al., 1990) employs an iterative Locally Estimated Scatterplot Smoothing (LOESS) to obtain the trend estimate and then a subsequent LOESS process to remove a changing additive seasonal component. The LOESS procedure employs a nonparametric method to estimate local regression surfaces by fitting a simple model to localized subsets of the data and creating a function that explains the deterministic part of the data variation at each point. A wider window, an odd number not less than seven, results in a smoother LOESS curve. The rate of seasonal change and the trend-cycle smoothness can be controlled by seasonal and trend-cycle window parameters, respectively. STL only directly supports additive decomposition. In order to use STL for multiplicative decomposition first logs of the data must be taken prior to using LOESS, and later the components must be transformed back to their non-logarithmic form (Hyndman, 2018).

4.5.2 Seasonal Decomposition of PM_{10} for the Auckland area using STL

STL decomposition of Auckland PM_{10} data was applied using R's 'forecast' package. The type of seasonality was identified as additive as no exponential growth was evident in the time series. The seasonal window was set to be 'periodic' and the default of *next odd (ceiling((1.5*period) / (1-(1.5/s.window))))* was used. This window is used because it allows for a periodic seasonal window to be used to control the rate of the trend-cycle. This is based on the previous evidence presented in Section 4.3.1 which points to existence of seasonalities in the Auckland data.

Figure 4.7 shows the STL decompositions of the Auckland PM_{10} concentration time series into seasonal and local trend components. The expected variation in daily PM_{10} concentrations during the cold season is detected by the seasonal component which exhibits a peak concentration during mid-winter. The relative scales of the components are shown by the grey bars on right of each panel representing the same length. Vertical axis scales of each subplot should be used to interpret each plot. The variation in the gray bar sizes to different scales of the plots. The large grey bar in trend panels shows small variation in the trend component compared to variation in the data. Strong variability was found in winter seasonal components (see Figure 4.7, seasonal panels). However, the real seasonal variation is expected to be smoother in nature. Similar findings of rapid seasonal variation was reported through LOESS decomposition of PM_{10} data collected from a monitoring station in Christchurch (Scarrott et al., 2009). In the more complex modeling, which is presented in Chapter 5 this real seasonal component will be shown to be captured by the meteorological variables themselves.

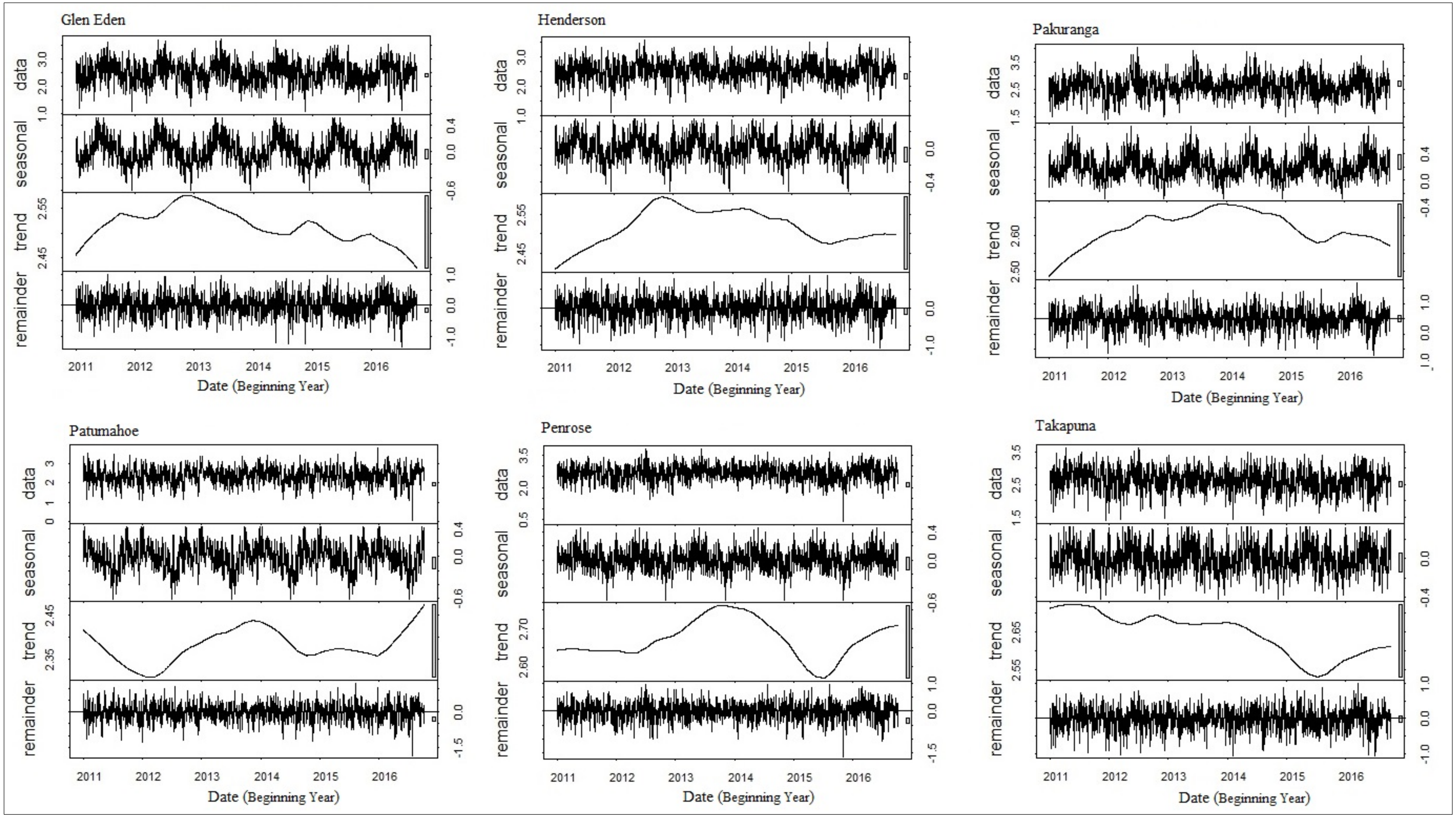


Figure 4.7: STL decomposition of daily average log PM_{10} series.

Distinct peaks of trend are present in the Glen Eden and Henderson 2012-13 time series with a substantial trend of decreasing PM_{10} concentration since 2014 at all the urban study sites. In contrast, an increasing trend was observed for Patumahoe rural site, between 2012 and 2015. Across the entire study area PM_{10} concentrations have notably increased in 2016. The clustering of the residuals (shown on the remainder) can be attributed to the autocorrelation in the PM_{10} concentration, therefore proving the significance of taking this lag structure into account when modeling the trend and testing for evidence of trend.

4.5.3 Seasonal Decomposition via Dummy Variables

Another method for removing the seasonal factor is the use of dummy variables (Hyndman, 2018). Dummy variables are categorical variables and are most commonly used in conjunction with regression analysis methods. Dummy variables are used to separate data into mutually exclusive classes for example weekday vs weekend day.

4.5.4 Seasonal Decomposition using Fourier Terms

Fourier terms can be used as an alternative to seasonal dummy variables, especially for long seasonal periods. Based on Fourier's theory a periodic function can be approximated using a series of sine and cosine terms of the right frequencies (Emmanuel Hernández Mayoral, 2017). Periodic seasonality with seasonal period of m, can be handled using pairs of Fourier terms (Hyndman, 2018):

$$x_{1,t} = \sin\left(\frac{2\pi t}{m}\right), x_{2,t} = \cos\left(\frac{2\pi t}{m}\right), x_{3,t} = \sin\left(\frac{4\pi t}{m}\right), x_{4,t} = \cos\left(\frac{4\pi t}{m}\right), x_{5,t} = \sin\left(\frac{6\pi t}{m}\right), x_{6,t} = \cos\left(\frac{6\pi t}{m}\right), x_{6,t}$$

Maximum ($K = m/2$) pairs of sine and cosine terms are the same as those achieved by seasonal dummy variables.

4.6 Time Series with Complex Seasonality

Time series with higher frequency, here daily observation of PM_{10} observations, often exhibit more complicated seasonal patterns. Figure 4.8 shows decomposition of the PM_{10} time series taking multiple seasonality in account. There are two seasonal patterns shown in site specific plots, one for the day of the week, labeled as 'Seasonal7' (the third red panel), and one for the year, labeled as 'Seasonal365' (the fourth green panel).

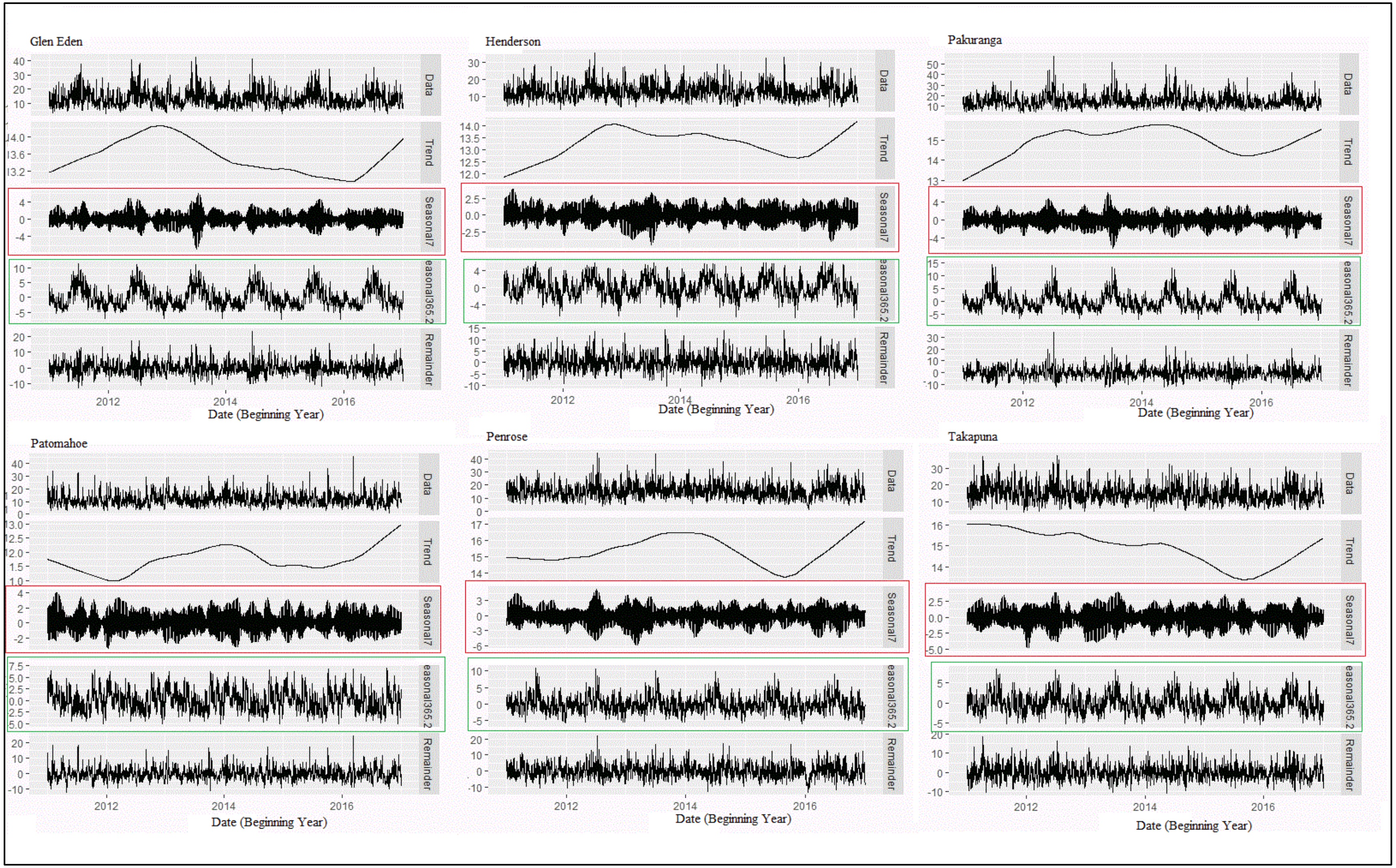


Figure 4.8: Multiple decomposition of the PM_{10} time series.

It is worth noting that the trend has relatively narrow ranges when compared to the other components. The weekly seasonality is also weak as result of the insignificant trend.

4.6.1 Multi Seasonal Adjustment Using Fourier Terms

To handle multiple seasonality time series, all the frequencies that might be relevant should be specified using dummy variables. These seasonalities may be represented through $S_{k}, k = 1, \ldots, m$. Strong seasonality at multiple levels is commonly present in air pollution concentration and road traffic volume. Diurnal fluctuation of meteorological conditions, seasonal fluctuation in solar radiation, and human activities cause clear yearly, seasonal, weekly, and daily periodicities in the time series. The initial explanatory analysis in Chapter 3 showed daily PM_{10} concentration exhibiting both a weekly and a monthly (winter months) elevation pattern. These patterns are mostly associated with anthropogenic resources. In time series with multiple seasonalities, Fourier terms are added for each seasonal period (see Section 4.7).

4.7 Trend Modeling without Meteorological Variation

The consistency in air pollution rates rising over previous years is a concern for environmental and air quality management. Trend results which provide a reliable uncertainty estimation and are able to be clearly communicated are vital for proper management of environmental planning. In this section, the aim is to look for evidence of trends in the daily PM_{10} concentrations and to explore the possible sources of uncertainties. The idea behind this exploration is to analyze possible trends in PM_{10} concentrations, prior to capture the meteorological impacts on the PM_{10} concentrations (Chapter 5) which is a rather complex process.

There are several statistical modeling techniques with parametric and nonparametric tests for trend detection and analysis. The use of these tests supports the interpretation of the results among existing uncertainties predictions (Zuma-Netshiukhwi et al., 2013).

Trend estimation is a complex approach as it is greatly influenced by the characteristics such as persistence and data irregularities in form of ties (Kisi & Ay, 2013). Parametric approaches such as Ordinary Least Squares (OLS) regression requires assumptions of normality in the distribution of residuals, no heteroscedasticity or autocorrelation, and a linear relationship to be valid. The assumption of no autocorrelation is frequently violated in time series analysis using OLS regression, which affects statistical inference through underestimating the standard errors, and hence the confidence intervals. The non-normal distribution of PM_{10} data could therefore lower the accuracy of the parametric test results.

There are two data-related assumptions that underlie nonparametric trend tests. Firstly, data point occurrences are assumed to be independent and identically distributed over time. Ties and degrees of persistence in time-series data can falsify this assumption which in turn influences the test statistic variance. Like the effect of anti-autocorrelation, increases in the data ties reduces the variance of the test statistic. The commonly used model to address the influence of short-term persistence on trend is the lag$_1$ autoregressive $AR(1)$ process (Yue et al., 2002). Secondly, the collected data points should hold the characteristics of the population. Insufficient sampling coverage, results in trends that are very likely biased and cannot represent the true value of the population of interest (Lang et al., 2019). A general process for determining trend assessment provided by NZ-Stats (2019) is based on at least three years of sub-annual data points, or at least six years of consecutive annual data points (with maximum missing data tolerance of 25%). According to New Zealand's Environmental Reporting Series, the seasonal Theil-Sen test should be carried out if there is at least three years of sub-annual data points. The Mann-Kendall (MK) test is only carried out if there are at least six years of consecutive data points available. The third test assumption is that the data are assumed to be collected under the "real conditions" of the sampling times. Imprints of anthropogenic influences and/or climate variability on air pollution levels are violation of this assumption and adds to the complexity of trend analysis in time series.

In New Zealand, a seasonal MK test for monotonic trends was used to detect any underlying trends in PM$_{10}$ time-series data sets for 2016 in the Waikato region (Wilton & Caldwell, 2018). The authors discussed that reasonably long historical data is required to confirm or exclude the existence of trend in their study area thus their results are limited.

New Zealand's Environmental Reporting Series is produced by the MfE and Statistics New Zealand. In their 2018 report, one of two methods of parametric or non-parametric statistical approaches was used to determine the direction of the PM trends. In their study, the trend analysis on monthly PM$_{10}$ concentrations in Auckland's airshed with varying time ranges was performed using the Theil-Sen function (MfE & Stats NZ, 2018). It was reported that the sites with the longer periods of historical data showed statistically significant declines in concentrations. The authors argued that the decreasing trends at most sites was suggestive of a reduction in home heating contributions to ambient PM$_{10}$ concentrations. However, the authors highlighted that their finding disagreed with rising trend from biomass burning over the Auckland region (Davy et al., 2017). The latest available report on Auckland's PM$_{10}$ trend used a de-seasonalised Theil-Sen analysis **without** taking the autocorrelation into account (MfE & Stats NZ, 2018). The results of this report therefore may be limited as the presence of autocorrelation in Auckland's PM$_{10}$ may effect inference of Theil-Sen estimates. Consensus is required on how to make such adjustments (Stats, 2019) – this consensus has yet to be reached at the time of writing this thesis.

4.7.1 Mann-Kendall (MK) Test

Rank-based tests are often preferred over parametric approaches in air pollution studies, and the exploration of method-related uncertainties are purposefully biased towards Mann Kendall (MK) (Mann, 1945; Kendall, 1955) and SMR tests (Onyutha, 2016). The non-parametric MK test is commonly used for monotonic trend detection in series of environmental data, climate data or hydrological data (Onyutha, 2016). The null hypothesis (H_0) is that the data are identically distributed and come from a population with independent realizations hence there is no trend or serial correlation structure among the observations. The alternative hypothesis, H_A, is that the data follows a monotonic trend (Pohlert, 2018). The Mann-Kendall test statistic is calculated according to (Mann, 1945; Kendall, 1955):

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^{n} \operatorname{sgn}(X_j - Y_k) \qquad \text{Eq. 4.6}$$

with

$$\operatorname{sgn}(x) = \begin{cases} 1 & \text{if} \quad x > 1 \\ 0 & \text{if} \quad x = 0 \\ -1 & \text{if} \quad x < 1 \end{cases}$$

The mean of S is $E[S] = 0$ and

the variance σ^2 is

$$\sigma^2 = \left\{ n(n-1)(2n+5) - \sum_{j=1}^{p} t_j\,(t_j - 1)(2t_j + 5) \right\} / 18 \qquad \text{Eq. 4.7}$$

where

p is the number of the tied groups in the data set and

t_j is the number of data points in the j^{th} tied group.

The statistic S is approximately normal distributed provided that the following Z-transformation is employed:

$$Z = \begin{cases} \dfrac{S-1}{\sigma} & \text{if} \quad S > 1 \\ 0 & \text{if} \quad S = 0 \\ \dfrac{S+1}{\sigma} & \text{if} \quad S > 1 \end{cases}$$

The statistic S is associated with Kendall's τ:

$$\tau = \frac{S}{D} \qquad \text{Eq. 4.8}$$

$$\text{where:} D = \left[\frac{1}{2}n(n-1) - \frac{1}{2}\sum_{j=1}^{p} t_j(t_j - 1)\right]^{1/2} \left[\frac{1}{2}n(n-1)\right]^{1/2} \qquad \text{Eq. 4.9}$$

The trend analysis was applied for the log transformed PM$_{10}$ of all sites. For the significance level (α = 0.05), the threshold value is 1.96. A positive (negative) value of Z signifies an upward (downward) trend. An absolute Z value higher than 1.96 indicates a significant changing trend (Ye et al., 2015). According to (Pohlert, 2018) S and the $p - value$ can also be used to detect the direction and significance of a trend. The non-parametric Cox and Stuart trend test examines the first third of the series with the last third used to detect a monotonic trend. Results of both Mann-Kendall and Cox and Stuart trend test are presented in Table 4.6:

Table 4.6: Mann-Kendall and Cox and Stuart trend test on log of daily PM$_{10}$.

Non-parametric Tests					
	Mann-Kendall trend test			**Cox and Stuart trend test**	
	s	***z***	***p-value***	***z***	***p-value***
Glen Eden	-3.339800e+04	-1.04	0.30	3.2	**0.001**
Henderson	1.053300e+04	0.33	0.74	1.16	0.25
Pakuranga	4.685400e+04	1.46	0.14	0.70	**0.48**
Patumahoe	5.742700e+04	1.78	0.07	0.78	**0.43**
Penrose	1.706700e+04	0.53	0.60	0.02	0.98
Takapuna	-2.142820e+05	**-6.66**	**2.765e-11**	**6.37**	**1.903e-10**

Table 4.6 shows that the absolute *Z*-value calculated in Mann-Kendall test for the first five stations did not reach 1.96, indicating no significant changing trend in PM$_{10}$. The negative *Z*-value of Takapuna (-6.66) with absolute value above the 1.96 indicates a significant decreasing trend. The *S* and p-value also indicate a significant decreasing trend ($S = -2.14,\ p < 0.001$) in Takapuna.

Like results of Mann-Kendall test, the Cox and Stuart test showed p-values smaller than *Z*-value for the first four sites accepting the null hypothesis of no monotonic trend in these sites. With the *Z*-value of 6.37 we reject the null hypothesis on a level of $p < 0.0001$ concluding the existence of a monotonic trend in Takapuna. However, the results showed monotonic trend in Glen Eden ($p < 0.05$) which is in contradiction with the Mann-Kendal test. The Cox and Stuart test is slightly weaker than the Mann–

Kendall test (Rutkowska, 2014) as *"Positive serial correlation among the observations would increase the chance of significant answer, even in the absence of a trend*" (Stuart,1955), p.95).

Therefore, the conclusion of a monotonic trend in Glen Eden may not be properly supported. To further investigate the existence the trend, a modified MK test is used to deal with autocorrelation in Auckland's PM_{10} time series.

4.7.1.1 Modified Mann – Kendall Test with Adjusted Auto Correlation

In the case of true correlation between the PM_{10} time series, the Modified Mann-Kendall Test (Hamed & Rao, 1998) for serially correlated data should be used in order to account for the existing serial correlation in the daily PM_{10} concentration values. The variance correction approach considers only the significant lags of autocorrelation coefficients. In such approach, the trend is removed, and the ranks of significant auto-correlation coefficients are used to calculate the effective sample size which is then used to adjust the inflated (or deflated) variance of the test statistic. According to Hamed, Rao and Chen (2003) only the first three auto-correlation coefficients are used in their proposed function. The autocorrelation and partial auto-correlation plots of the time series of daily PM_{10} concentration time series presented in Chapter 4 show a weak autocorrelation for two days and therefore the Modified Mann-Kendall test with 95% confidence interval for the slope of the trend is calculated. The result of the Modified M-K test is presented in Table 4.7.

Table 4.7: Modified Mann-Kendall Test. *p-value* is the original Mann-Kendall *p*-value. *p*-value^b is the *p-value* after variance correction. Decision on significance is based on *p*-value^b.

	tau	***p*-value**	***p*-value^b**	**trend**	**significant?**
Glen Eden	-0.016	0.30	0.55	Negative	No
Henderson	0.004	0.78	0.87	Positive	No
Pakuranga	0.021	0.16	0.41	Positive	No
Patumahoe	0.025	0.07	0.34	Positive	No
Penrose	0.007	0.63	0.76	Positive	No
Takapuna	-0.097	2.1883e-11	4.309396e-05	Negative	Yes

The reported *p*-value after variance correction (*value^b*) shows a significant downward trend for Takapuna. Glen Eden also shows a negative but not significant trend. The trend for the remaining sites is positive but not significant.

4.7.2 Spearman's Rho (SMR) Test

SMR was applied as a comparison to the Mann-Kendall tests. The SMR test assumes that time series data follows an identical distribution and are independent. The null hypothesis signifies no trend over time; the alternate hypothesis shows existence of an increasing or decreasing trend.

Table 4.8: SMR tests results for Auckland's PM_{10} time series.

	p-value	***rho***
Glen Eden	0.31	-0.02
Henderson	0.72	0.01
Pakuranga	0.15	0.03
Patumahoe	0.26	0.04
Penrose	0.58	0.01
Takapuna	5.525e-11	-0.14

The results of SMR test presented in Table 4.8 agrees with MK tests showing a negative trend at the Glen Eden and Takapuna sites. The trend however is again only statistically significant for Takapuna site only.

The time series of PM_{10} investigated in this research was shown in section 4.6 to have complex seasonality so a decision was made to explore recent, novel models that are designed to cope with this complexity. Thus, Section 4.8 of this Chapter describes in detail two relatively new, at the time of writing, forecasting models: Harmonic Regression and TBATS, for complex seasonality time series. Performance of these two forecasting methods and a combination model are compared and evaluated in discussed in Section 4.9.

4.8 Time Series with Complex Seasonality

Harmonic Regression Models: Harmonic regression is referred to as a regression model with Fourier terms where the consecutive Fourier terms signify the harmonics of the first two Fourier terms. For long seasonal periods, such as daily data that has an annual seasonality (365 days), seasonal differencing involves comparing today's event with last year's events on the same day even if the seasonal pattern is not smooth. For such time series, a harmonic regression approach can model the seasonal pattern using Fourier terms and handle the short-term time series dynamics with an ARMA error. The harmonic regression approach has advantages over ARIMA models as it can handle multiple seasonality of any length. The number of Fourier sine and cosine pairs (k) can control the seasonal pattern smoothness so that the smaller values of k the smoother the seasonal pattern. The only disadvantage would be the assumption of fixed seasonality over time. However, this drawback is minimal in impact as seasonality is usually and remarkably constant in practice, except for long time series (Hyndman, 2018).

TBATS model is a relatively new state space modeling framework developed by De Livera (2011). TBATS uses a Box-Cox transformation to tackle the non-linearity in data. To capture the autocorrelation in the residuals TBATS use the ARMA model. In TBATS notation, (w, p, q, ϕ, {m_1, k_1}, {m_2, k_2} … {m_M, k_M}), w and φ indicate the Box-Cox and damping parameters, respectively. The error is modeled as an $ARMA\ (p, q)$ process, the seasonal periods are listed as $m_{1}, \ldots, m_M$ and $k_{1}, \ldots, k_M$ is the number of Fourier terms for each seasonality.

The accuracy of forecast methods is commonly measured by the Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE) metrics. In the following experiments these are the measures that will be reported.

4.9 Experiments and Results

In order to identify the best method for forecasting Auckland's PM_{10} concentration the Dynamic Harmonic Regression and TBATS model were used. In these experiments the training set contains the daily data for the period from 2011 to 2015 for the purpose of model development and training. The test set (the data from 2016 is used) is used for model validation. A clear weekly seasonality as well as annual seasonality (average annual year length of 365.25 days) was evident in ACF and PACF plots reported in Section 4.3. Therefore, the existing frequencies in the PM_{10} series are $m1 = 7$, $m2 = 365.25$.

4.9.1 Dynamic Harmonic Regression with Multiple Seasonal Periods

Fourier terms are used for each seasonal period as discussed in Section 4.6.1. For the Auckland PM_{10} time series, the seasonal periods are identified as 7 and 365.5 therefore the Fourier terms are derived as:

$$\sin\left(\frac{2\pi kt}{7}\right), \cos\left(\frac{2\pi kt}{7}\right), \sin\left(\frac{2\pi kt}{365.5}\right), \cos\left(\frac{2\pi kt}{365.5}\right), \ldots$$

The model with an ARMA error structure is fitted. The total number of Fourier terms for each seasonal period have been chosen to minimize the AIC. The R code for tuning the number of Fourier terms based on lowest AIC values is provided in Appendix B (1). Two variables were used generate Fourier series with a period of 7 days and 365.5 days. ARIMA model was fitted using the two variables as external regressors and the corresponding AIC value was calculated. Once all the AIC values were obtained, the best parameters corresponding to the minimum AIC value was chosen. The value of lambda was set to zero so that the forecasts and prediction intervals stay positive. The effect of both day-of-week and national holidays were included by bonding the two regressors as a new parameter. Once the final model was fitted to the time series, the fitted model was then trained on the training set. A combination of terms from the two variables was used as an external regressor for the prediction using the training data as the input as suggested by Denham (2012b).

The best fit was returned with an ARIMA (1, 0, 1) error, including three autoregressive terms and three moving terms, with three Fourier terms corresponding to a period of 7 days and three Fourier terms corresponding to a period of 365.5 days. The coefficients and the standard errors (s.e.) for the various terms for Glen Eden are shown in Table 4.9 . The results for the other sites are provided in Appendix B (2).

Table 4.9: Estimated coefficients and the standard errors, Glen Eden.

	coefficients	**standard errors**
intercept	13.60	0.17
ar1	0.21	0.05
ma1	0.28	0.05
s1_7	0.13	0.20
c1_7	0.20	0.20
s2_7	0.08	0.14
c2_7	-0.12	0.14
s3_7	0.14	0.09
c3_7	0.05	0.09
s1_365.25	-0.64	0.24
c1_365.25	-3.02	0.24
s2_365.25	0.01	0.24
c2_365.25	1.66	0.24
s3_365.25	0.85	0.24
c3_365.25	-0.38	0.24

To test the dependency of the residuals, the Ljung-Box test was performed on residuals. A *p* value greater than 0.05 indicates that the residuals are independent. Results of Ljung-Box test on residuals from the Dynamically Harmonic Regression with ARIMA(1,0,1) errors is provided in Table 4.10. The *p*-values returned by the test are all greater than 0.05 (Table 4.10) for all sites and are indication of the residuals being independent.

Table 4.10: Results of Ljung-Box test on residuals from Dynamic Harmonic Regression with ARIMA (1,0,1) error for the predicted year, 2016.

	df	*p-value*
Glen Eden	350	0.208
Henderson	353	0.06
Pakuranga	352	0.40
Patumahoe	353	0.15
Penrose	351	0.15
Takapuna	350	0.23

Figure 4.9 shows the residual plots of the Harmonic Regression for the Glen Eden site. From the residuals plot it can be concluded that some information has not been captured with this model that relies solely on Fourier terms as predictors. Using other predictors such as meteorological data might improve the results. The plots for the remaining sites are provided in Appendix B (2).

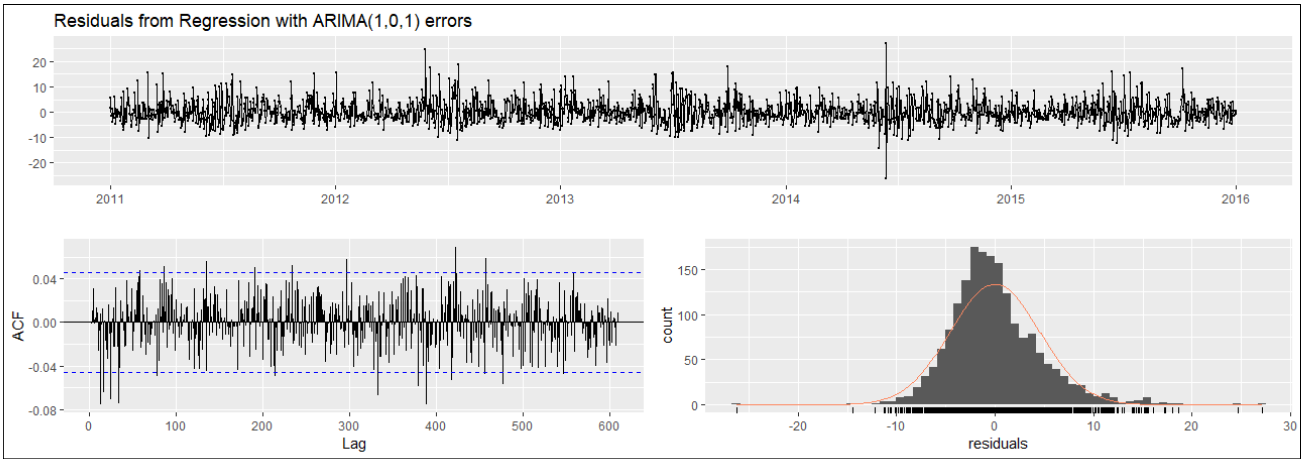


Figure 4.9: Plots of residuals from Regression with ARIMA (1,0,1) error, Glen Eden.

It can be observed, from the results in Table 4.10, that Harmonic regression using Fourier terms should forecast one day ahead with reasonable accuracy for the Auckland sites under investigation due to the fact that there are no correlated residuals (Hyndman, 2018). It can also be observed that the values of RMSE and MAPE provided by models is low in horizon 1. The forecast accuracy decreases in h=7 which could be due to the weekly seasonality period not being captured by the model. The accuracy improves for h=120 when compared with h=7 but marginally decreases at h=365. This could be due to harmonic regression terms that force the seasonal patterns to repeat periodically without changing.

Two error components RMSE and MAPE were computed to check the out-of-sample performances for the different forecasting horizons: h=1, 7, 120 and 365. Table 4.11 shows the comparison between the

RMSE and MAPE error components obtained from the site-specific Harmonic Regression models using the Fourier terms.

Table 4.11: Dynamic Regression measure error for different horizons.

	Glen Eden		**Henderson**		**Pakuranga**		**Patumahoe**		**Penrose**		**Takapuna**	
Horizon	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
1	0.83	7.32	0.33	2.93	0.19	1.60	4.30	42.95	2.67	16.49	3.44	22.10
7	3.03	29.39	3.23	35.6	2.82	26.21	6.12	92.16	4.78	50.19	4.51	36.46
120	3.53	27.57	3.24	27.35	3.99	24.80	5.16	39.96	5.90	69.85	3.74	23.23
365	5.22	32.90	4.49	29.63	6.01	30.57	5.32	46.26	6.05	45.12	3.39	28.13

4.9.2 TBATS

The TBATS models was applied to the PM$_{10}$ time series and an error analysis was performed. The TBATS model was applied to the training data set repeatedly. The obtained TABTS $(1,\ \{1,1\},\ -,\ \{<7,\ 3>,\ <\ 365.25,\ 3>\})$ model for the Glen Eden represents $w = 1$ (no Box-Cox transformation), the order of ARMA error is (1, 1), with no damping parameter. The number of harmonics is $k_1 = 3$, $k_2 = 5$. The total number of original seasonal values were 12, calculated as $(2 \times (3 + 3))$. Decomposing the PM$_{10}$ time series allows inferences to be drawn about patterns of change over time, leading to physically interpretable latent processes underlying the data. Decomposition of Glen Eden PM$_{10}$ data by the TBATS model is represented in Figure 4.10. Figures of remaining sites are provided in Appendix B (3).

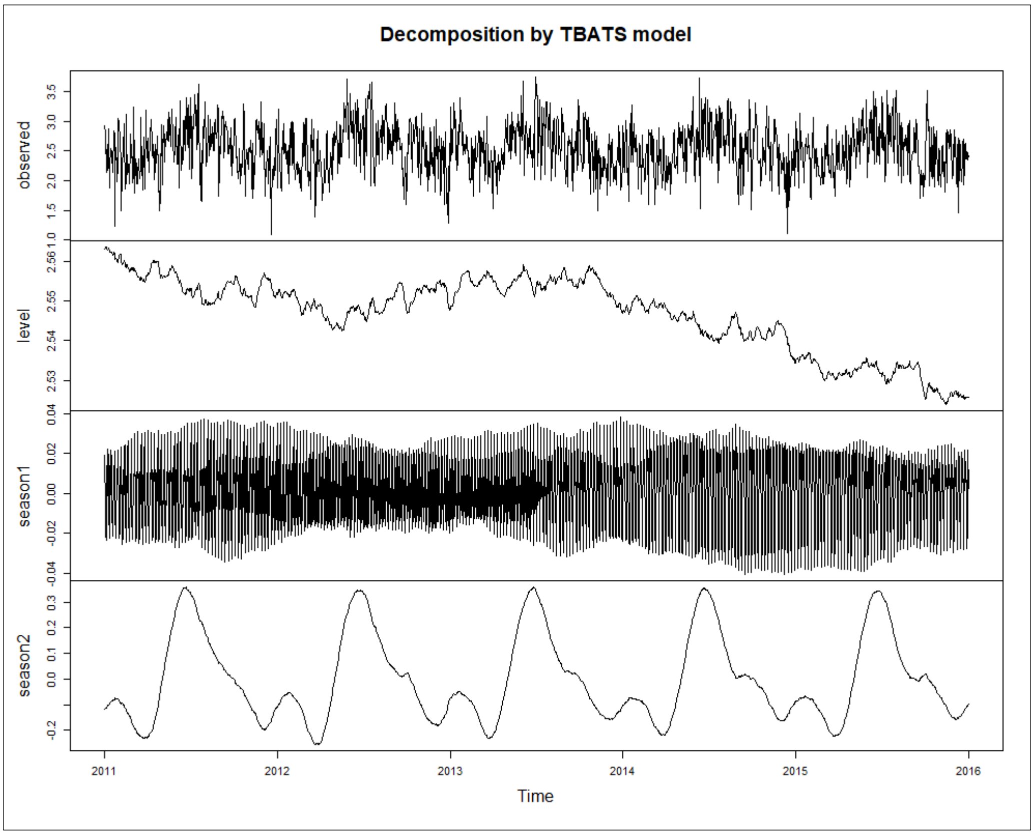


Figure 4.10: Decomposition from TBATS model, Glen Eden.

The decomposition generated from the TBATS $(1, \{1,1\}, -, \{<7, 3>, <\ 365.25, 3>\})$ is divided into four parts including observed data, trend component, and two seasonality components namely season$_1$ and season$_2$ representing weekly and yearly seasonality, respectively (Figure 4.10). The pattern of weekly seasonal component changes with time, while pattern of yearly seasonal component stays relatively more stable. The obtained TBATS models are provided in Table 4.12.

Table 4.12: Site specific TBATS models.

	TBATS Models	initial seasonal values
Glen Eden	TBATS (1, {1,1}, -, {<7, 3>, < 365.25, 3>})	12
Henderson	TBATS (1, {2,1}, -, {<7, 2>, < 365.25, 6>})	16
Pakuranga	TBATS (0.99, {4,0}, -, {<7, 3>, < 365.25, 3>})	12
Patumahoe	TBATS (1, {3,1}, -, {<7, 1>, < 365.25, 4>})	10
Penrose	TBATS (1, {4,0}, -, {<7, 2>, < 365.25, 5>})	14
Takapuna	TBATS (1, {4,0}, -, {<7, 2>, < 365.25, 7>})	18

TBATS models were also applied on training data to generate forecast values for four different forecasting horizons as h=1, 7, 120 and 365. Two error components RMSE and MAPE were computed to check the test performances for these forecasting horizons. Table 4.13 shows the RMSE and MAPE metrics obtained from site specific TBATS models for different horizons. The observed value of RMSE and MAPE provided by TBATS model for the shorter forecasting horizon is lower than the RMSE value for the longer forecasting period for all sites. The TBATS model accommodated non-integer seasonality as well as a lesser number of initial parameter estimation.

Table 4.13: TBATS measure error for different horizons.

	Glen Eden		Henderson		Pakuranga		Patumahoe		Penrose		Takapuna	
Horizon	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE	RMSE	MAPE
1	0.20	1.65	0.15	1.36	0.003	0.02	3.45	34.48	3.30	20.36	3.07	20.33
7	3.09	30.87	3.19	35.39	2.83	27.07	4.84	71.32	3.79	38.07	4.07	35.99
120	3.50	27.78	3.18	25.37	4.04	29.59	5.02	33.38	5.92	63.46	3.99	25.33
365	4.96	33.80	4.48	28.53	5.58	33.30	5.56	41.64	6.44	42.02	5.34	28.58

4.10 Conclusion

Complex seasonality is often present in long time series including Auckland's daily PM_{10} concentration. As traditional forecasting methods fail to handle such complex seasonality, new forecasting methods like Dynamic harmonic regression and TBATS become more valuable. In this Chapter, both Dynamic harmonic regression and TBATS methods have been applied to Auckland's daily PM_{10} time series. The analysis exposed both weekly, and yearly multiple seasonality in the Auckland PM_{10} time series. Two error components RMSE and MAPE were studied at different forecasting horizon to compare the forecasting accuracies of these two models. The TBATS model with a lesser number of parameters was found to be a more accurate technique for forecasting of this data set with minimum error when there is the presence of complex seasonality. This could be because seasonality can change slowly over time in a TBATS model. Performance of TBATS however was subject to the forecasting horizon itself. Considering different forecasting horizons, the result can be explained by the ability of the TBATS model containing trigonometric terms to accommodate non-integer seasonality as well as a lesser number of initial estimate parameters. Thus, it is concluded that TBATS is a better technique, when compared with the other time series models considered in this research, for forecasting PM_{10} with complex seasonality at least for the available training and test data set used here. Future work in this area should include examining the effect of other predictors such as meteorological parameters on PM_{10} concentration forecasts.

Chapter 5 STATISTICAL ESTIMATION AND MODEL BUILDING

In this Chapter two different statistical model building approaches, that were applied to provide statistically rigorous descriptions of the effect of meteorological variables and trends on Auckland's PM_{10} concentration, are discussed. The model building stages are described and insights into the features of the developed models are provided. The assumptions made when building the models are outlined and reasonability of violation of the assumptions are discussed.

5.1 Introduction

In Chapter 4 trend detection approaches were explored to detect the trend and serial autocorrelation features of the PM_{10} concentration time series, *without* taking meteorological effects into account. To examine this in more depth the work reported in this Chapter looks at Generalized Additive Models (GAMs) and Generalized Additive Mixed Models (GAMMs). These methods are alternative analytic approaches to the inclusion of autocorrelation and they use nonparametric local smoothing methods (Scarrott et al., 2009). Both GAM and GAMMs bypass the requirement of specifying a parametric form for seasonal trends and improve the level of robustness against model misspecification. The ability to smooth concurrent input variables in order to simulate nonlinear relationships in statistical models (Hastie & Tibshirani, 1986), makes GAMs a standard analytic tool in time-series studies for many environmental problems (Belušić et al., 2015). GAM has been found to be exceptionally beneficial for handling the complex nonlinearity associated with air pollution data (Belusic et al., 2015; Shuang et al., 2017; Wu & Zhang, 2019). However, there is very little work in the literature that looks at GAM as a method for modeling PM_{10}.

Performance of GAMs was found to be similar to ANN models for predicting Ozone concentration 8-hours in advance (Schlink et al., 2003). In work more similar to the work presented in this thesis, GAMs was employed to model PM_{10}, $PM_{2.5}$ and coarse PM concentration using different meteorological variables (Aldrin & Haff, 2005). The authors noted that both PM_{10} and $PM_{2.5}$ concentrations increased with decreasing temperature. However, PM_{10} concentration decreased with an increase in relative humidity whereas an increase in RH resulted in increase in $PM_{2.5}$ concentration. In another study, GAM analysis demonstrated that the largest influence on air quality in Melbourne, Australia was attributable to local-scale meteorological conditions (Pearce, Beringer, Nicholls, Hyndman, & Tapper, 2011).

It has only been in recent years, that researchers have tried to use GAMMs to explain air pollution and its relationship with epidemiological problems (W. Li et al., 2018). As a result, there is a lack of literature reporting the use of GAMMs for studying the short-term health effects of air pollutants, and in particular of PM_{10} and PM contributors such as traffic and meteorology. One of the few papers using a GAMM model to assess the impact of both meteorology factors and traffic density on ultrafine particulate matter (UFP) and $PM_{2.5}$ concentrations is that of Zwack et al. (2011). They examined these phenomena in Brooklyn, New York (USA). A more recent paper details a hybrid model containing both a dispersion model and GAMM to estimate contributions of traffic to daily $PM_{2.5}$ concentration in China (Fang et al., 2016). Clearly while the few papers using GAM and GAMMs to model PM promote the use of these methods further work is warranted in order to evaluate their general usefulness across different geographical areas.

In this Chapter the influence of meteorology on the distribution of daily PM_{10} concentration is modelled using both GAMs and GAMMs. If consistency in results from these two complex models occurs, then it will provide a degree of confidence in any conclusions drawn. These models will be fitted to the meteorological dataset explored in Chapter 3.

5.2 Generalized Additive Models (GAMs)

GAMs are an extension of Generalized Linear Models (GLMs) where the linear predictor incorporates smoothing of the predictor to allow for nonlinear relationships between the predictor and the target variable.

The nonparametric function of l has a structure of (Belusic et al., 2015):

$$\log(y_i) = s_0 + \sum_{j=1}^{n} s_j(x_{ij}) + \varepsilon_i \qquad \text{Eq. 5.1}$$

Where

i varies from 1 to n and n is the number of observations,

j is number of predictors in the model,

y_i is the i^{th} PM_{10} concentration,

s_0 is the overall mean of the response,

$s_j(x_{ij})$ is the smoothing function of the i-th value of covariate *and*

ε_i is the i^{th} residual.

In generalised models the difference between the observations and the fitted values is measured using deviance which is considered to be equivalent to variance in a linear regression (Murase et al., 2009).

A GAM model can find and describe the relationships between complex variables using smoothing functions (Green & Silverman, 1993), or locally weighted smoothers (LOWESS) (Cleveland et al., 1990; Cleveland & Devlin, 1988). The use of splines for non- and semi-parametric modeling to smooth curves, surfaces and non-linear covariate effects is well established (Clifford, 2013). The simplicity of splines, their flexibility to include penalties to tune the amount of smoothness, and include periodic bases to avoid non-smooth joins in periodic data (Eilers & Marx, 2010) has led to their implementation as 'basis functions' for GAM (Hart et al., 2009; Li et al., 2012; Q. Li et al., 2018).

The choice of smoothing term in practice does not usually result in any significant change to the final GAM outcome (Scarrott et al., 2009). In contrast, the choice of smoothing parameter is critical. In a study by Wood (2017) penalized thin plate splines are used as basis functions, as they retain the above-mentioned advantage of using splines and are computationally efficient. Another study has looked at using both a Self-Organizing Map (SOM) and GAM in an attempt to study the influence of synoptic-scale circulations on some air pollutants including PM_{10} (Pearce, Beringer, Nicholls, Hyndman, Uotila, et al., 2011). In their study, the authors used a sub-set of their data to develop the GAM model because, according to the authors, GAM was not practical to implement on their complete datasets.

In a GAM model, concurvity can be presented between smooths of space or time covariates (Wood, 2013). A major statistical concern has been raised in the literature in regards to the suitability of GAMs concurvity as it can cause overestimation of the GAM model parameters and underestimation of their variances (He et al., 2006; Ramsay et al., 2003).

5.3 Generalized Additive Mixed Models (GAMMs) with Autocorrelated Errors

The main drawback with GAMs is the assumption of independence between the observations of response. Spatially dependent or environmental data may be autocorrelated and using models that ignore this dependence can lead to inaccurate parameter estimates (the standard errors of the estimated regression coefficients) and inadequate quantification of uncertainty (Latimer et al., 2006). The results of the explanatory analysis in Chapter 3 showed a correlation between PM_{10} concentration and previous days' lag. This observed autocorrelation may be attributed to the meteorological driving factors that are also correlated with prior days. Therefore, is reasonable to assume that some correlations exist amongst observations from the same feature. According to Lin and Zhang (1999), the over-dispersion and correlation issues observed in GAM and GLMM can be overcome by adding random effects to the additive predictors.

Generalized Additive Mixed Models (GAMMs) (Lin & Zhang, 1999) is extension of GAM that contains both fixed and random effects hence the term mixed model.

"*As an innovative and outstanding model in the modern era, the GAMM seems not to be widely applied to most areas because of the inconvenience of software support*" (Chien, 2009, p. 30).

5.3.1 Smoothing Term Selection Methods

Shrinkage methods have received notable interest for variable selection in GAMs and are considered to be a valid alternative to more traditional subset selection and stepwise methods. Shrinkage approaches have been shown to be more stable when compared with stepwise and selection methods as they do not depend on the path chosen through the variable space (Marra & Wood, 2011). Furthermore, variable selection in shrinkage procedures is achieved in one step as opposed to the multiple steps required by subset selection and stepwise procedures (Hesterberg et al., 2008). The shrinkage approach has been proposed as a mechanism for smooth component selection in GAMs and operates by adjusting the smoothing penalty. The level of shrinkage component is set in such a way that it has an insignificant contribution to the model penalization and only activates when the term is effectively 'completely smooth' according to the conventional penalty. Another approach is to construct an additional penalty for each smoothing function instead of changing the original smoothing penalty. This penalty penalizes only completely smooth functions. A term will be eliminated if all the smoothing parameters have a tendency to infinity (Wood, 2017a).

Prediction error criteria or likelihood-based methods can be used to choose the proper smoothness of each term. Using Generalized Cross Validation (GCV) saves computational time/effort as it can be calculated without preforming cross-validation. Lower values of the GCV score indicate better fitting models. GAMs fitted using GCV smoothness selection can suffer from under-smoothing and may present local minima that can trap the minimisation algorithms (Wood, 2017b). Alternatively using REML of maximum likelihood (ML) has been shown to be much more robust to under smoothing, but at computational expense (Wood et al., 2016).

5.4 Experiments and Results

5.4.1 Modeling Process

GAM and GAMM models were used to explain the variation of log transform of the response variable (PM_{10} concentration) using meteorological and trend terms. These model then were used to estimate the log transform of PM_{10} concentration based on selected meteorological and temporal (Julian Day and DOW) variables.

The use of log transformation is sensible for various reasons. The exploratory analysis results in Chapter 3 revealed the negative exponential impact of some meteorological variables (e.g., temperature) on PM_{10} concentration. Using a log transform will make the relationship closer to linear and make model fitting simpler. The log transformation also changes the additive meteorological effects in linear models to multiplicative effects on the transformed scale. In addition, classical statistical inferences are usually based on the assumption that observations are normally distributed (Scarrott et al., 2009). The log transformation of air pollutants is also used widely in literature for similar reasons (Chaloulakou et al., 2003; Schlink et al., 2003)

A regression model was fitted to the available datasets and tested for heteroscedasticity of residuals. Figure 5.1 shows plots of residuals vs fitted values and standardised residuals on the Y axis for Penrose's linear regression model. To interpret the plot distribution of points throughout the range of X axis values and the shape of red line are observed. Homoscedasticity is indicated if flat red line and a completely random, equal distribution of points. The plots in shows a slightly curved red line and decrease in residuals as the fitted Y values increase are observed in Figure 5.1 (a) indicating presence of heteroscedasticity in Penrose PM_{10} concentration.

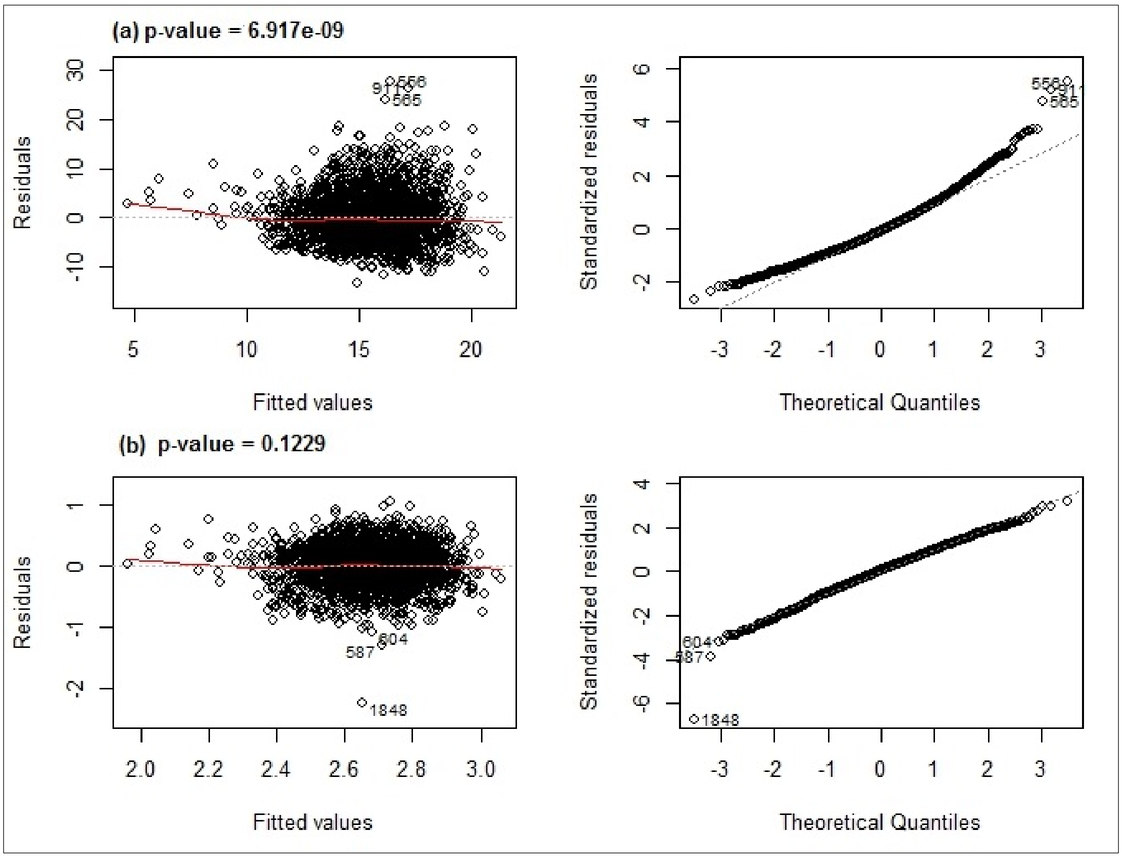


Figure 5.1: Linear regression plots of Penrose PM_{10} on (a) original scale and (b) log transformed scale.

The results of both the Breush-Pagan test (Breusch & Pagan, 1979) and the Non-constant Variance Score (NCV) test (Cook & Weisberg, 1983) gave a *p*-value for Penrose that is less than the 0.05 significance level (Table 5.1) therefore we fail to reject the null hypothesis that the variance of the residuals is constant, thereby confirming the graphical inference observed in Figure 5.1(a). Log transformation of PM_{10} data ensures homoscedasticity and that predicted values are positive after back transforming on the original scale. The result of a linear regression model fitted to the log transformed Penrose PM_{10} is presented in Figure 5.1(b). With a *p*-value of 0.1229, the null hypothesis is supported (that the variance of the residuals is constant) and therefore we can infer that the residuals are homoscedastic. The plots, for the log transformed model, also show a much flatter line for the standardized residuals (bottom-right plot) and more evenly distributed residuals (bottom-left plot) than the original untransformed scale model.

Table 5.1: *p*-value of heteroscedasticity tests on $PM_{10.}$

	***p*-value**	
	Breush-Pagan	**NCV**
Glen Eden	2.2e-16	3.32e-65
Henderson	2.12e-16	2.02e-36
Pakuranga	3.28e-16	2.78e-45
Patumahoe	1.02e-16	1.19e-18
Penrose	6.917e-09	2.14e-23
Takapuna	3.74e-11	1.56e-17

Normal, gamma and Poisson distribution are amongst the most frequently used distributions for modeling of air pollution data. Choosing the normal distribution yields constant variance in the response errors even in case of increments in PM_{10} concentration level. Acceptable approximation to normal and log normal distribution can be provided by gamma family (Scarrott et al., 2009). The effects of using gamma distribution methods instead of normal distributions were investigated in this study, however no significant changes in the results were observed. Given the reliability of the inference results in GAM and GAMM, it was decided to use the normal distribution models in the research reported herein. Smoothly varying trend will capture the variation in PM_{10} concentrations due to anthropogenic sources (Scarrott et al., 2009). Modeling these smooth functional forms can be done through different statistical methods. The degree of smoothness can be decided from the data itself (Scarrott et al., 2009). The only assumption taken in this study's nonparametric approach is that the functional form is smooth in nature.

5.4.1.1 Concurvity Check

To ensure that a GAM was appropriate for modeling of Auckland's PM_{10} concentration, a concurvity check was performed using Wood's method (Wood, 2013). In this method a relative index of concurvity, bounded between 0 and 1, is used to measure the degree of identifiability of the covariates. In this method an index near to 1 indicates a total lack of identifiability while 0 indicates that there is no concurvity issue and the covariates are identifiable. The concurvity values of the models in this study were found to be within the acceptable range (see Table 5.2).

5.4.1.2 Smoothing Term Selection

In this thesis's research, the penalized approach is implemented using the *select* argument in R's msgv package (Wood, 2017b). In practice the upper limit on the Degrees of Freedom (DF) coupled with a smooth term are set using the knots (k or k-1). The choice of k is important, and the default is arbitrary. However, the recommendation is to set k to be large enough to contain the envisioned dimensionality of the underlying function that is being recovered from the data (Simon N Wood, 2018). Additionally, the GCV score of a fitted GAM is used for smoothness selection to avoid the computational cost of ML. To overcome the issue of local minima using GCV extreme smoothing parameters were used as initial values in optimization and big leaps in smoothing parameters during optimization was avoided, as suggested by Wood (2018).

5.4.1.3 Modeling the Short-Term Effect of Meteorological and Temporal Variables on PM$_{10}$ Concentration Using GAM

In the GAM modeling procedure, a GAM model is constructed fitting the predictors into formula (1) for each station area using the GAM function available in package mgcv (Wood, 2017b) in R. The GAM can be mathematically expressed as follows:

$$g(\mu_t) = \beta_0 + f_1(x_{i,1}) + f_2(x_{i,2}) + f_3(Rain_{i,1}) + \ldots + f_p(DOW) + f_{p+1}(JD) \qquad \text{Eq. 5.2}$$

Where

n is the observation number day,

$g(\mu_i)$ is the "link" function (here, a log link is used),

$x_{i,j}$ are meteorological and lag$_1$ and lag$_2$ predictors,

$f_j(x_{i,j})$ is an initially unknown smooth function of $x_{i,j}$ obtained from a thin-plate spline basis set.

To adjust for seasonal cycles and other temporal trends, a smoothing function of days using natural splines with 28 DF as a base model is included. During the sensitivity analysis, different DFs may be replaced to find the best knot if necessary. This 28-day range covers the extent of temporal smoothing used in most published time-series studies. A second time predictor represented as a smooth function $f_p(DOW, k = 7)$ for the day of the week was included in the model in a similar to manner to that reported in literature (Belusic et al., 2015; Bertaccini et al., 2012). The significance of coefficients for the DOW term in each GAM model is presented in section 6.4.2. The EDF value for the smoothing of DOW suggests whether DOW should be used as a linear parametric term or not. That is evaluated by plotting the fitted smoothed model as reported later in section 6.4.1. The selected meteorological terms by GAMs are statistically significant at the $\alpha = 0.001$ level.

In this research, the GAMM building process undertaken and reported in this Chapter was designed to encapsulate the First-Order Auto-Regressive (AR(1)) error structure into the GAM model given in Eq. 5.3 and replace the GAM with a GAMM model with auto-correlated errors. The meteorological terms were included as fixed effects. An AR(1) with a site specific estimated autocorrelation was included in the model to represent for the random effect. Hence, the estimated smooth trend from the GAMM captures less variability than the GAM with no AR(1) component.

5.4.2 Model Validation Method

Performance of the GAMs and the robustness of the results for the functional relationships between the meteorological predictors and PM_{10} was evaluated by performing a 10-fold cross-validation (Stone, 1974) for each GAM. Using the "CVgam" function in the R package the cross validation returns two measures namely *GAMscale* and *CV-mse-GAM*. The *GAMscale* is the mean of the squares of the errors of the original GAM fits. The *CV-mse-GAM* is the mean of the squares of the errors calculated for the 10% of data not included in the fit for each of the 10 cross validation GAMs. The *CV-mse-GAM* is expected to be larger than the *GAMscale*. However, large differences between these values would suggest that the GAMs are over-fitting the data, as the performance is much poorer on the data not included in the fit during the cross-validation (McVey et al., 2018). In addition to *CV-mse-GAM* and *GAMscale*, several other metrics were calculated to assess model performance, their tendency toward over/under prediction and overfitting. These metrics are: Root Mean Square Error (RMSE), Coefficient of determination (R^2), Fractional Bias (FB) used as measure of mean relative bias, Factor of 2 (FAC2) used as fraction of data satisfying the FB (Legates & Maccabe, 1999). Index of Agreement (IOA) (Willmott et al., 2011) is used to calculate the model's prediction error degree and for comparing the performance of models. In the case of inconsistency in the results using these measures a majority rules approach will be adopted.

The site specific **s**tandard diagnostic plots of GAM are presented in Figure 5.13 to Figure 5.18. The plots are interpreted according to Wood's guidelines (Wood, 2017c).

5.4.3 Auckland PM_{10} GAM Models

PM_{10} for each of the six stations was modelled independently using the model given by Eq. 5.4 using both meteorological and temporal (DOW and Julian day) variables. Although Aldrin and Haff (2005) collinearity test was exclusively based on correlation coefficients, in this study concurvity of each term with the rest of the GAM model is calculated. For all variables, the estimated values were smaller than one, with a maximum value of 0.5 found between relative humidity and solar radiation in Takapuna (Figure 5.2). Therefore, it was assumed that the meteorological and temporal variables are not collinear, and that a GAM method could therefore be used.

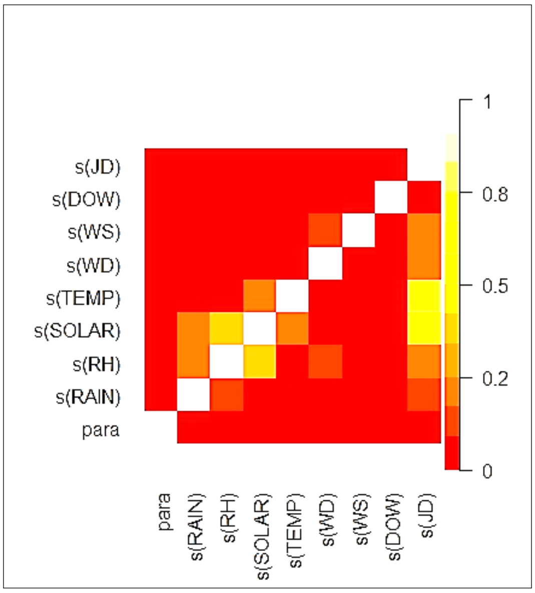


Figure 5.2: Concurvity plot of Takapuna's GAM model.

Three related indices of concurvity, ranging between 0 (different) and 1 (totally identical) were computed. In all the three indices a smooth term, *f*, in the model is decomposed into a *g* part (that lies entirely in the space of one or more other terms in the model) and a remainder part that is completely within the term's own space. If a large part of *f* is made up from *g* then there is a concurvity problem.

> "*The indices used are all based on the square of* $||g||/||f||$, *that is the ratio of the squared Euclidean norms of the vectors of f and g evaluated at the observed covariate values*." (S. N. Wood, 2018)

Table 5.2 shows the three indices of concurvity test for Takapuna. The value of the first row (worst, Table 5.2) is a pessimistic measure (largest value that the square of ||g||/||f|| could take) showing the worst case possible irrespective of data. Temperature has the highest value (0.91) in this row. The second row contains the value of $(||g||/||f||)^2$ according to the estimated coefficients which in some cases could somewhat underestimate the possibility of concurvity. The third row measures the extent that the g basis can explain the f basis and does have the pessimism or potential for over-optimism of the other two. The estimated results show a weak concurvity between the meteorological terms for all sites, and therefore are not shown here as they are nearly identical. Table 5.3 shows the parametric component (para) and the estimated pairwise concurvity measures between each smooth term for Takapuna's GAM model. The moderate concurvity found between time and two meteorological terms (solar and temperature) are likely due to the seasonal trends of these two terms.

Table 5.2: Concurvity indices for Takapuna's GAM model (*para* is the parametric component and *JD* is the Julien Day).

	para	Rain	RH	Solar	Temp	WD	WS	DOW	*JD*
worst	4.7e-19	0.56	0.718	0.813	0.909	0.406	0.533	0.052	0.9
observed	4.7e-19	0.289	0.668	0.608	0.439	0.253	0.247	0.038	0.3
estimate	4.7e-19	0.286	0.584	0.699	0.797	0.310	0.405	0.033	0.0

Table 5.3: Concurvity results of Takapuna GAM model (R: s(RAIN), RH: s(RH), S: s(SOLAR), T: s(TEMP), WD: s(WD), WS: s(WS), DOW: s(DOW), JD: (JD)).

	para	Rain	RH	Solar	Temp	WD	WS	DOW	*JD*
para	1	1.9e-24	7.1e-26	3.3e-31	9.5e-25	3.2e-31	1.0e-24	1.1e-31	9.9e-25
Rain	1.0e-19	1	2.2e-01	1.8e-01	2.1e-02	3.8e-02	6.8e-02	8.2e-03	9.0e-03
RH	2.4e-23	1.3e-01	1	3.3e-01	1.6e-02	6.2e-02	1.9e-02	3.6e-03	2.9e-03
Solar	2.3e-28	7.2e-02	3.4e-01	1	2.1e-01	5.5e-02	2.7e-02	1.4e-03	3.4e-03
Temp	3.0e-22	1.1e-02	2.0e-02	2.0e-01	1	7.1e-02	4.0e-02	3.3e-03	7.5e-03
WD	1.2e-28	3.6e-02	1.4e-01	6.0e-02	8.2e-02	1	8.8e-02	4.0e-03	6.8e-03
WS	2.1e-22	2.9e-02	3.3e-02	7.3e-02	6.3e-02	3.8e-02	1	6.9e-03	5.3e-03
DOW	1.0e-27	3.8e-03	2.7e-03	1.5e-03	3.2e-04	1.1e-03	1.1e-03	1	4.1e-05
JD	3.5e-19	1.3e-01	2.0e-01	5.3e-01	7.3e-01	1.8e-01	1.7e-01	1.1e-03	1

The association of PM_{10} with the meteorological and temporal factors in the GAM modeling was examined by individually extracting the partial responses of PM_{10} to these factors. The estimated partial response functions for the site-specific GAM models are presented in Figure 5.3- Figure 5.12. The partial effects of each predictor on PM_{10} concentration are shown by the curves. The dashed lines show the 95% confidence intervals. The input variables value is shown by the thick marks on the X axis. The error bars are wider for fewer observations. The vertical scales show the original scale. Estimated smooth relationships which have been penalized out of the model are blank (e.g., Pakuranga in Figure 5.3).

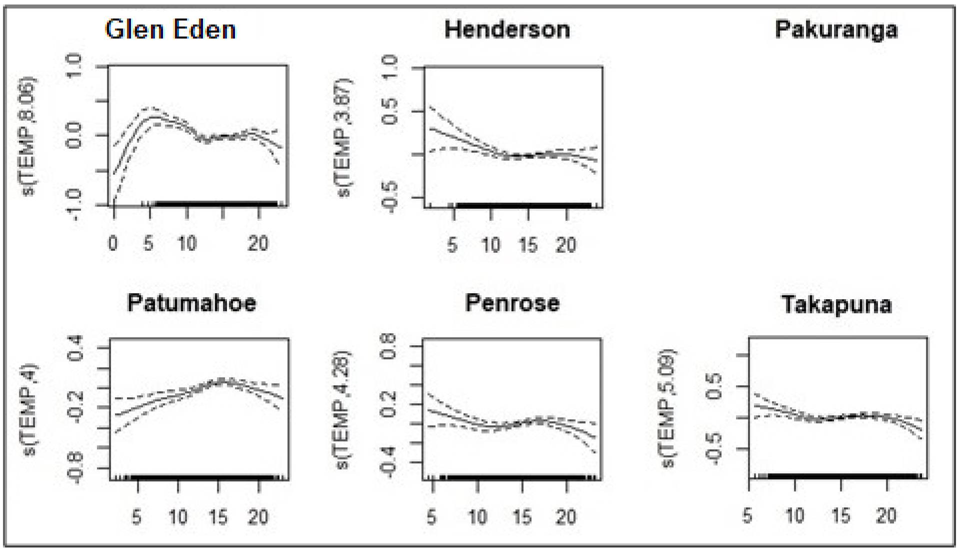


Figure 5.3: Estimated smooth relationships with 95% confidence intervals between PM_{10} and temperature. For Pakuranga, blank plot, temperature has been eliminated from the model.

As depicted in Figure 5.3, the partial response is non-linear, with an apparent decline when temperatures increase from 5 °C to 15 °C. All other predictors are constant for the urban sites where temperature was selected as a significant covariant by GAM. In contrast, a trend of an increase in partial response with an increase in temperature is observed for Patumahoe, a rural site, within the same temperature range.

The effect of RH on PM_{10} concentration is obvious for Glen Eden, Henderson, Penrose, and Takapuna where concentrations initially increase with rising RH until a RH of around 65% is reached. A decrease in PM_{10} concentration beyond this 65% RH threshold is observed for Henderson and Penrose. This decrease at over 65% RH could be the result of the wet deposition of pollutants, in which PM_{10} reacts with water vapor ($H_2O(g)$) in the atmosphere. A similar observation was also found by Klaić et al. (2012) for residential Zagreb. As illustrated in Figure 5.4, the positive impact on PM_{10} concentration is clear for extremely high RH values (≥ 80%) at both Glen Eden and Takapuna.

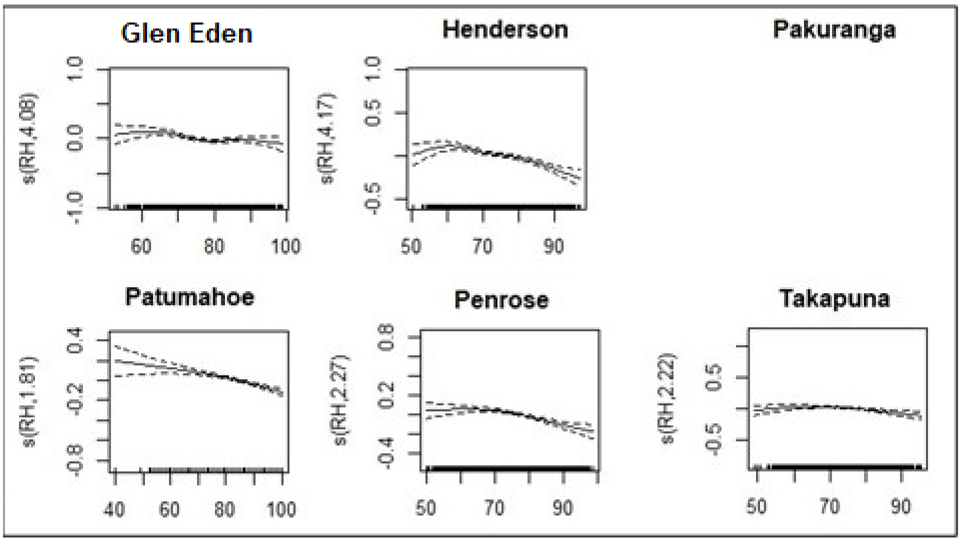


Figure 5.4: Estimated smooth relationships with 95% CI are between PM_{10} and RH. It should be noted that RH was not selected as a smoothing term in Pakuranga by GAM.

All sites except for Pakuranga exhibited the highest PM_{10} concentrations for wind directions of around 210º (SSW) and lowest concentrations at around 110º (ESE) (Figure 5.5). As discussed in Chapter 3, Penrose with industrial background is located on the southeast of Auckland close to the Southern motorway and is subject to pollutant transport by southeasterly winds.

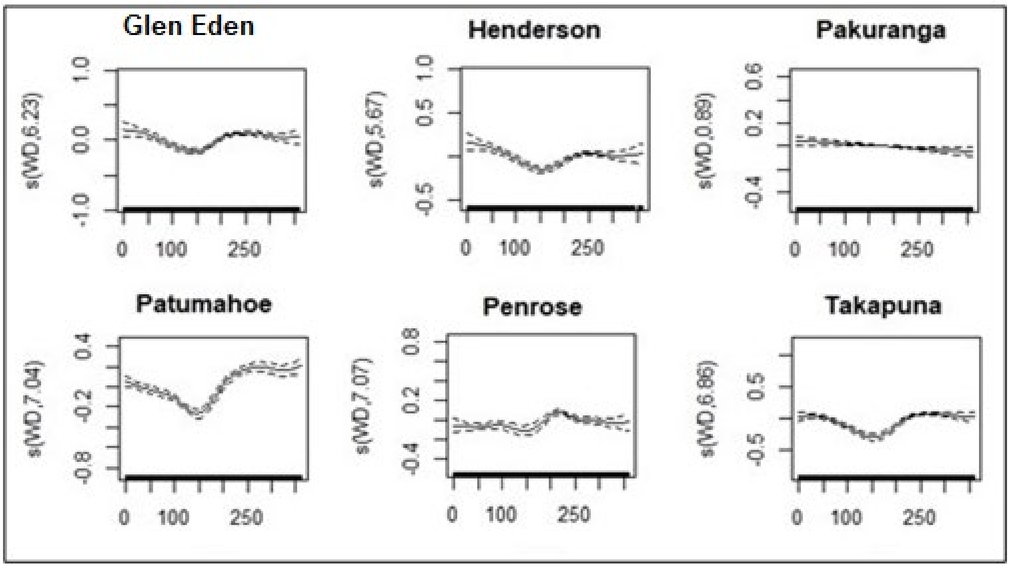


Figure 5.5: Estimated smooth relationships with 95% CI between PM_{10} and wind direction.

The decreasing effect of north and northeasterly flows on PM_{10} concentrations is obvious in all sites. Furthermore, the same decreasing effect was also found for easterly flows. The west and south westerly patterns were found to have a large impact on the concentrations of PM_{10}, (in agreement with Chapter 3/wind polar results). When considering the position of the Penrose site with respect to the southern motorway, one would expect the maximum pollutant concentrations to be associated with traffic emissions. Similarly, westerly winds should therefore contribute to an increase in concentrations from road traffic in Henderson.

Pollutant concentrations and wind speeds for Pakuranga exhibit mainly a negative correlation clearly showing the effect of local ventilation. Maximum PM_{10} concentrations occur as wind speeds increase from 2 ms^{-1} in the remaining sites demonstrating that PM_{10} transportation to the sites via advection dominates the local ventilation. Similar results were also found for western winds and their influence on CO_2 over Zagreb (Belusic et al., 2015). This hypothesis was also further confirmed by the analysis of the relationships between pollutant concentrations and both wind speed and direction which is presented in Chapter 3.3.1.1.

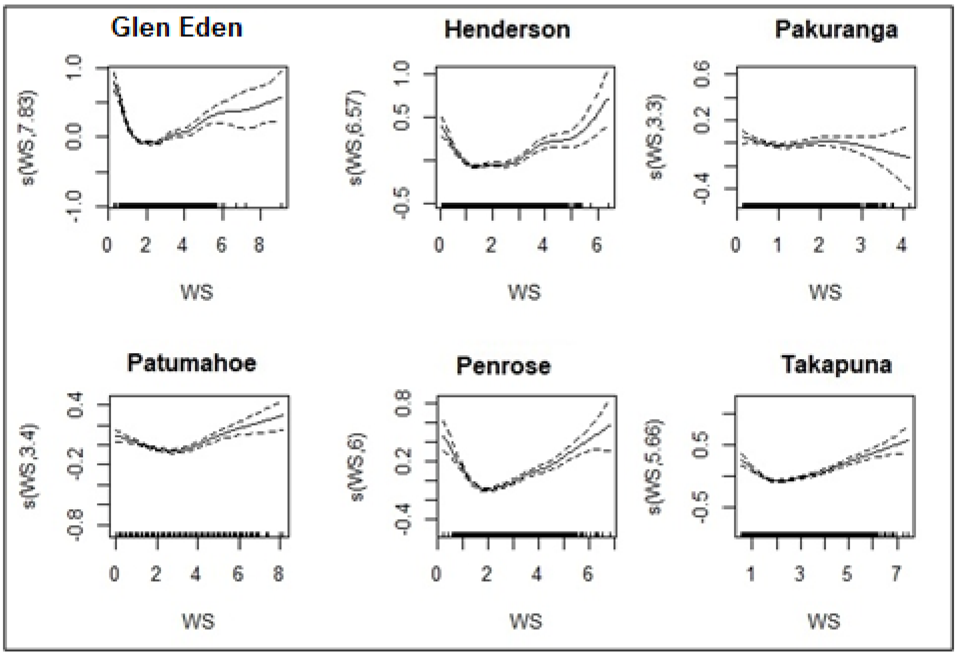


Figure 5.6: Estimated smooth relationships with 95% CI's between PM_{10} and wind speed.

The nonlinear relationship between rainfall and PM_{10} captured by the GAM model is illustrated in Figure 5.7. At weaker rainfall intensities, PM_{10} concentrations decrease in Glen Eden and Patumahoe – this trend could be due to pollutant foraging by rain drops. In contrast, larger rainfall intensities corresponded to elevated concentrations. Similar findings were also reported by Aldrin and Haff (2005). It is notable that, periods of high rainfall intensity (rainfall of more than 0.4 mm per day) occurred less frequently than those of weaker intensity during the study period. The confidence intervals given in Figure 5.7 are sufficiently large, suggesting that the curvature in the relationship between rain and PM_{10} concentration is true. Takapuna, however, did not exhibit a clear pattern possibly due to the consistently low daily mean rainfall values. GAM eliminated rainfall as a factor in the models for Henderson and

Pakuranga. This also illustrates the localized (micro scale) effect of weather within Auckland which poses challenges for environmental modeling.

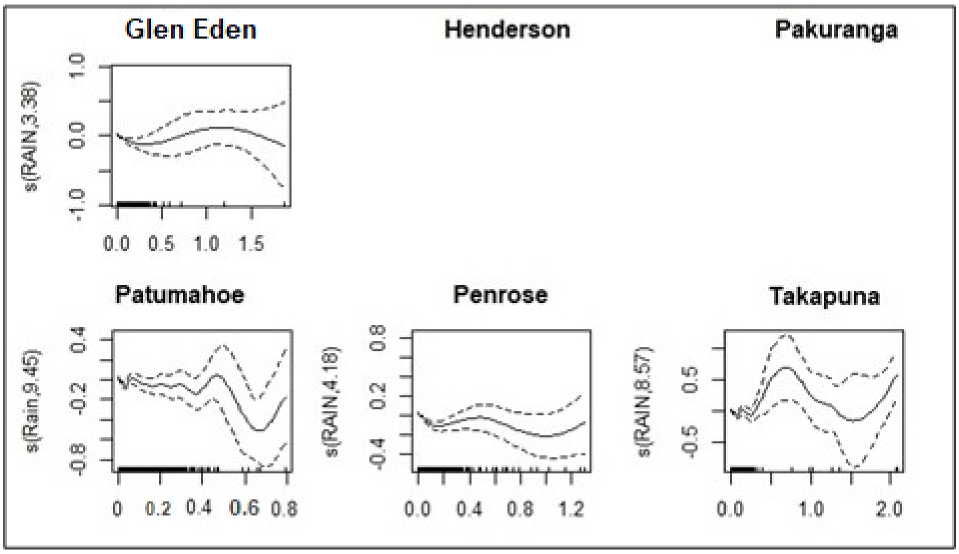


Figure 5.7: Estimated smooth relationships with 95% CI's between PM_{10} and rainfall.

During weekdays, PM_{10} concentrations show a slight increase across all the available sites, probably due to people travelling to work and school (Figure 5.8). PM_{10} concentrations were almost unchanged from Monday (2) to Friday (6). PM_{10} concentrations declined during weekends and this finding agreed with reports by Chaloulakou et al. (2003). The only PM_{10} levels increasing on a weekend day were those on Sunday at Patumahoe. Patumahoe is rural area and is less subject to the effects of commuter traffic.

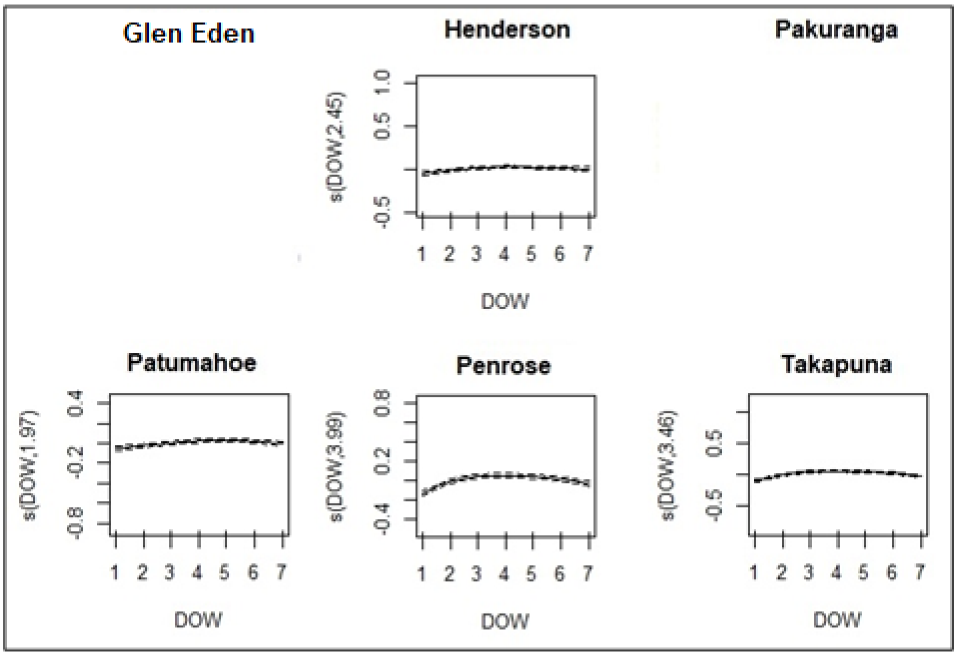


Figure 5.8: Estimated smooth relationships with 95% CI's between PM_{10} and the day of week (Sunday (1) – Saturday (7)).

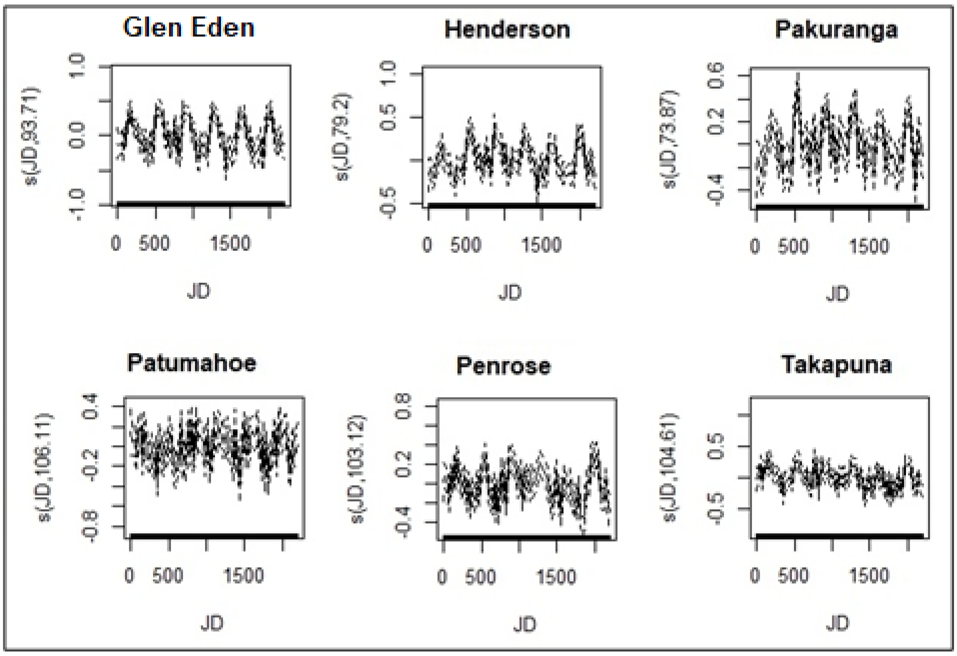


Figure 5.9: Estimated smooth relationships with 95% CI between PM_{10} and *JD.*

Note for the s(*JD*) trend model estimated over the whole year, the partial effect plots (Figure 5.9) show a distinct seasonality as higher PM_{10} are observed during cold seasons. Analysis of the day number curve showed the highest PM_{10} concentration during winter and a minimum level during summer. These higher concentrations can be explained by the increase in emissions associated with fuel consumption and wood burners during winter (Chapter 3).

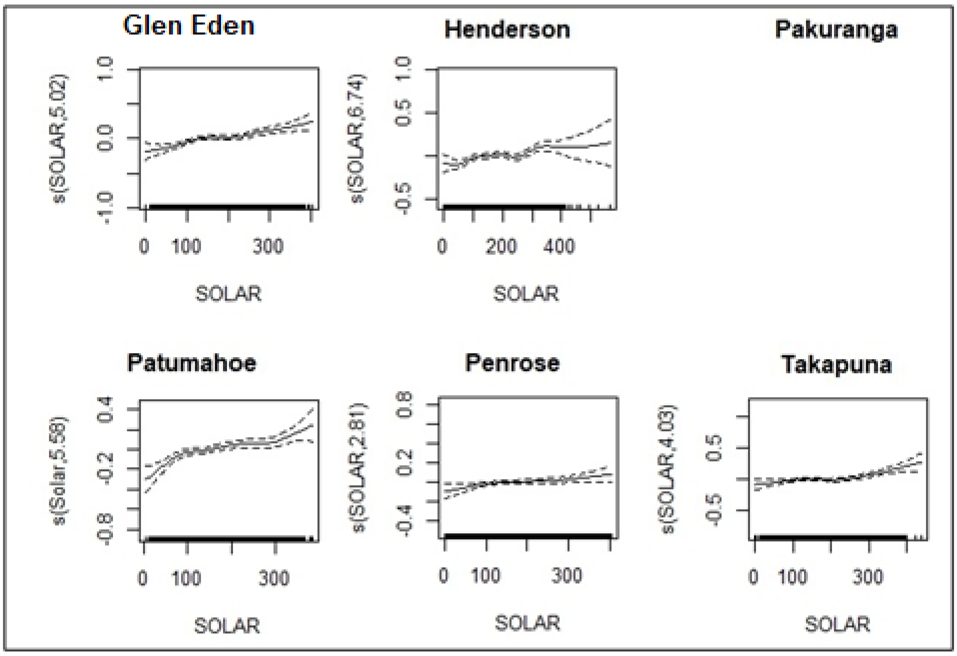


Figure 5.10: Estimated smooth relationships with 95% CI's between PM_{10} and solar radiation.

It was expected, based on prior studies (for example see (Afzali et al., 2014) and (Mao et al., 2018)), that solar radiation would have a negative correlation with $PM_{10.}$ However, a slight increase is observed in Figure 5.10 as solar radiation increases to 100. The curve levels off when solar radiation reaches around 200. A more obvious second increase is observed as the solar radiation rises above 300. This positive correlation may be explained by the windy conditions that are normal for Auckland during sunny days (warm season) and the slight positive correlation with wind speed and PM_{10} concentration during some warmer seasons (as discussed in Chapter 4's section 4.3.2).

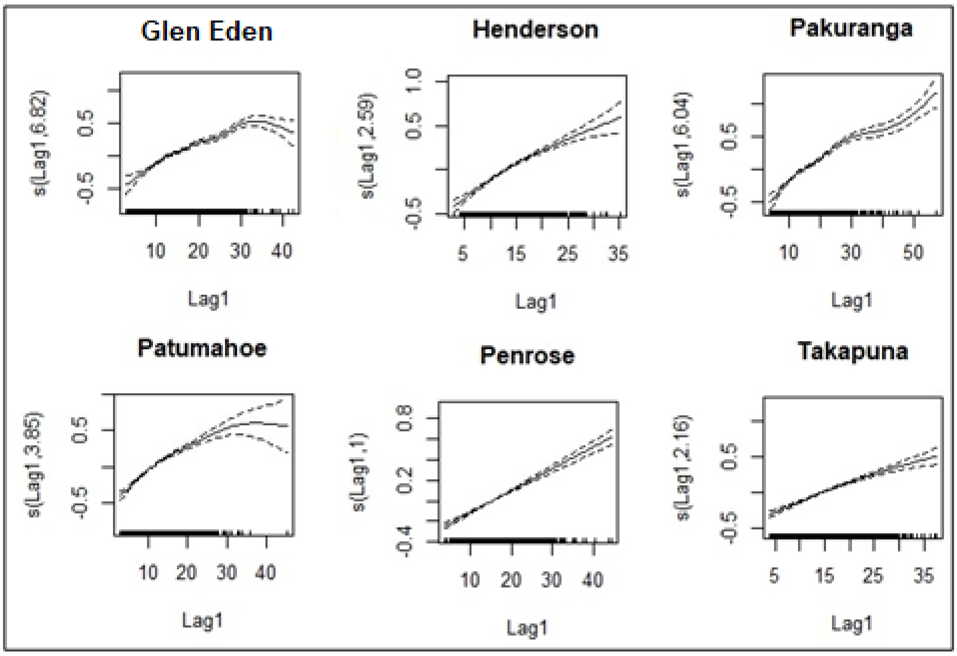


Figure 5.11: Estimated smooth relationships with 95% CI's between PM_{10} and PM_{10} lag_1

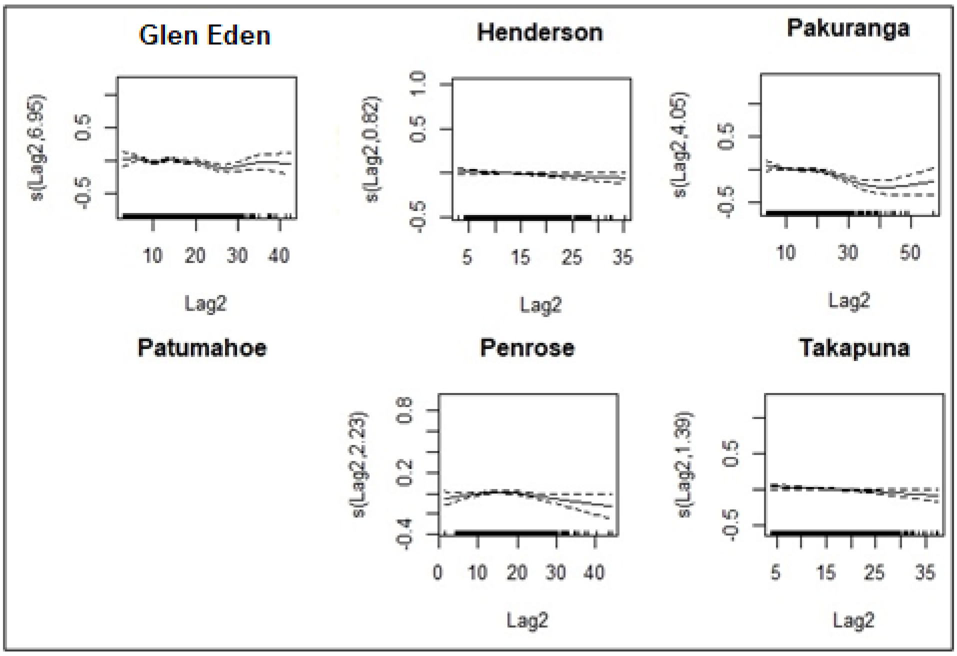


Figure 5.12: Estimated smooth relationships with 95% CI's between PM_{10} and PM_{10} lag_2

5.4.3.1 Model Evaluation

In all sites the normal *QQ*-plot (A) is close to a straight line which supports the reasonable distributional assumption. The variance (B) is virtually unchanged as the mean rises. The histogram of residuals (C) seems consistent with normality. A positive linear relation with a fair deal of scatter is presented in response versus fitted values (D).

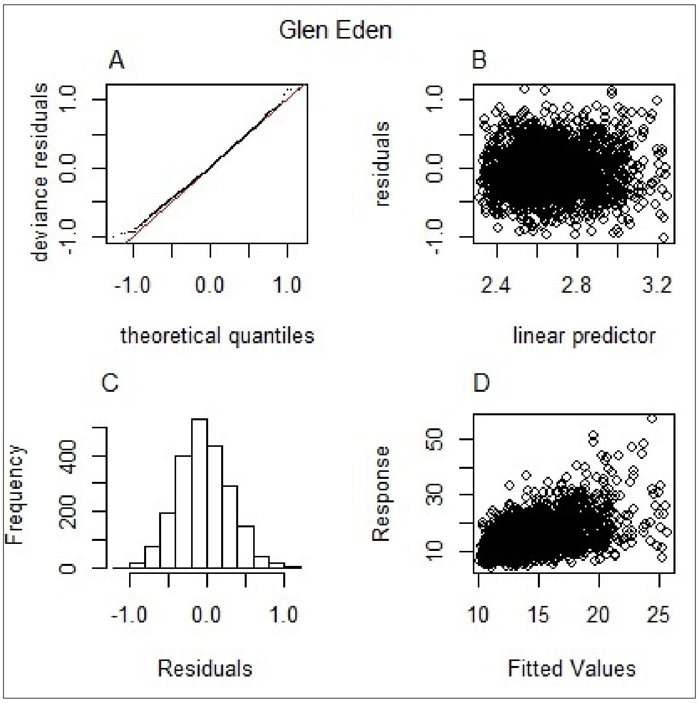


Figure 5.13: GAM standard diagnostic plots, Glen Eden.

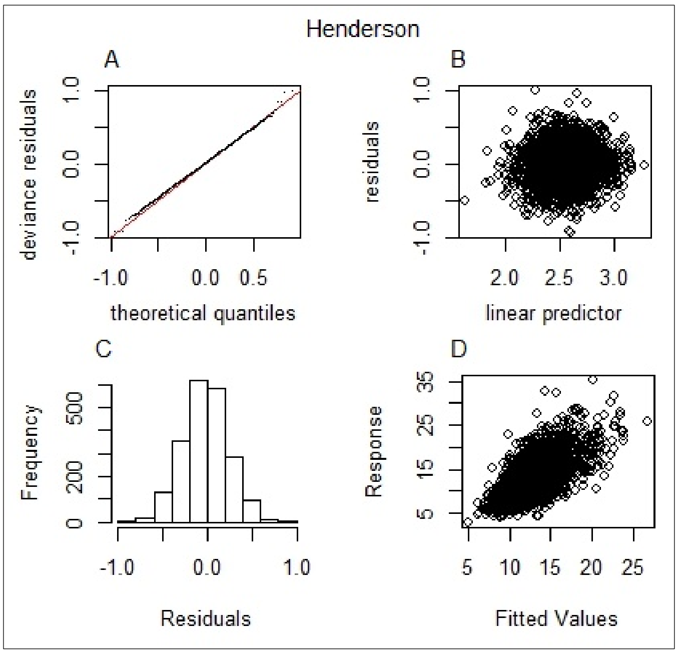


Figure 5.14: GAM standard diagnostic plots, Henderson**.**

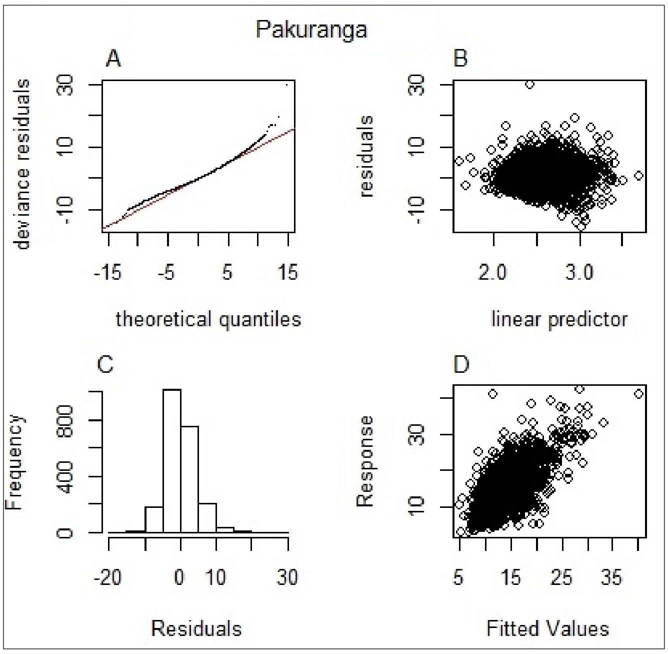


Figure 5.15: GAM standard diagnostic plots, Pakuranga**.**

The Figure 5.16 plots indicate a poor fit as residuals do not follow a normal distribution.

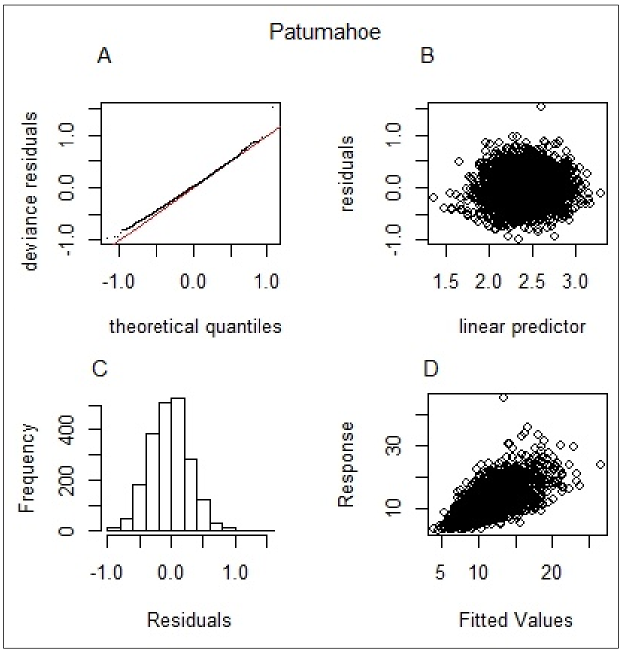


Figure 5.16: GAM standard diagnostic plots, Patumahoe.

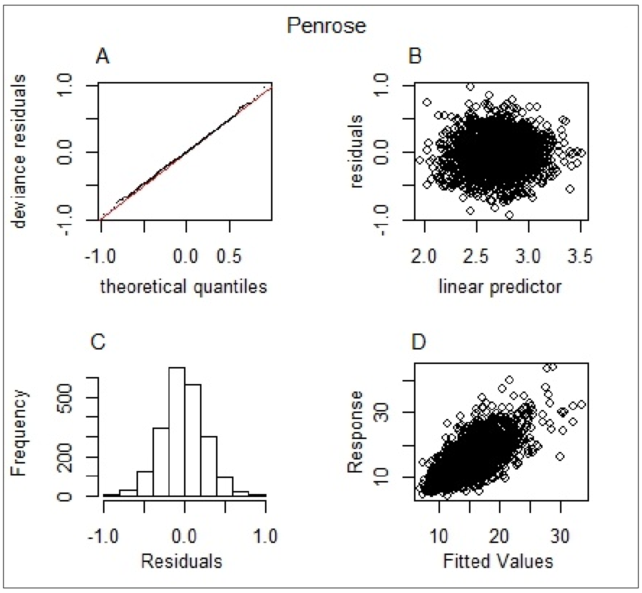


Figure 5.17: GAM standard diagnostic plots, Penrose.

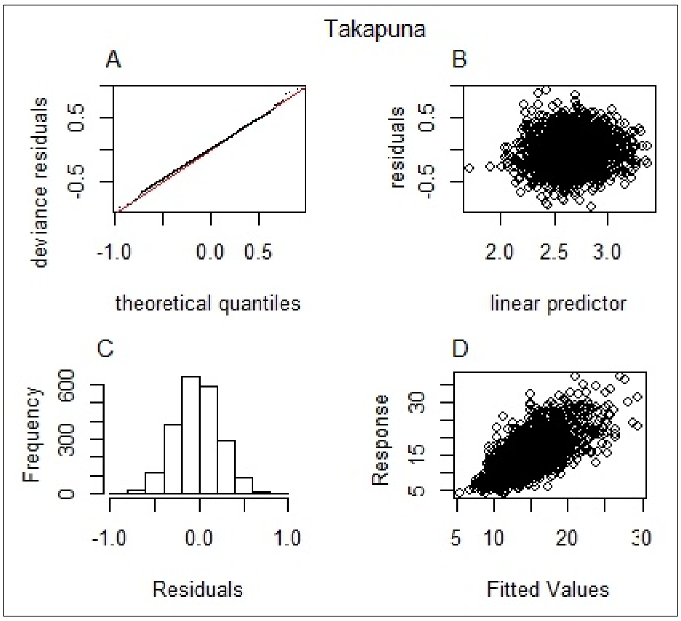


Figure 5.18: GAM standard diagnostic plots, Takapuna.

The residual plots for all sites, except Pakuranga, indicate a good fit as the histogram of residuals follow a normal distribution. The majority of residuals cluster around zero as evidenced by the diagnostic plots (Figures 5.11-5.18). Table 5.4 depicts the result of fitted basis GAMs for each individual site and for the entire dataset. The results of the GAMs show that the selected terms explained 47.1% of the deviance in PM_{10} at Glen Eden and Patumahoe. The GAM relating meteorological and lag variables to the total daily average PM_{10} for Pakuranga explained only 23% of the PM_{10} deviance, and generally has a much poorer fit. The result is comparable with finding of Pearce et.al., (2011) with GAM explaining 41.4%

of PM_{10} variation and Hart et.al., (2009) with 49% of PM_{10} deviance explained. The AIC value obtained is used for comparison of the selected terms and should not be used to evaluate model performance.

Table 5.4: Statistical evaluation of the GAM model on the log scaled PM_{10} concentrations for the entire dataset.

	MAE	MSE	RMSE	MAPE	AIC	R^2 (adj)	Deviance explained $^{(\%)}$
Glen Eden	0.25	0.10	0.31	0.11	12629.59	0.44	**47.1**
Henderson	0.21	0.07	0.26	0.09	11125.09	0.38	45.01
Pakuranga	0.27	0.12	0.35	0.11	13458.16	0.20	23.00
Patumahoe	0.24	0.10	0.31	0.11	11495.77	0.38	42.80
Penrose	0.21	0.07	0.26	0.08	11978.17	0.42	46.12
Takapuna	0.20	0.07	0.26	0.08	11887.70	0.40	44.10

Over fitting in the GAMs and the robustness of the results of the functional relationships between the predictors and PM_{10} were tested through 10-fold cross-validation for each GAM. The result of cross validation of the fitted GAMs on the test data for all sites is presented in Table 5.5. A comparison of the *GAM-scale* and the *MSE* values in all GAM fits shows *MSE* is larger than the *GAM-scale* value as expected. However, both values were found to be very close to each other suggesting that the models are not over-fitting. The negative values of Fractional Bias (FB), for Henderson, Pakuranga, Patumahoe and Penrose suggest a tendency toward over prediction whereas the positive values indicate under-prediction. However, the calculated values of FB are small and close to zero confirming that there is no systematic tendency to over or under prediction. Patumahoe shows the highest Pearson coefficient and IOA of 0.69 and 0.64, respectively.

Table 5.5: Statistical evaluation of the GAM model on the log scale for test set.

	Glen Eden	**Henderson**	**Pakuranga**	**Patumahoe**	**Penrose**	**Takapuna**
MAE	0.26	0.22	0.28	0.26	0.22	0.22
MSE	0.11	0.08	0.13	0.11	0.08	0.08
RMSE	0.33	0.28	0.35	0.33	0.28	0.28
MAPE	0.11	0.09	0.11	0.12	0.09	0.09
GAM scale	0.10	0.07	0.12	0.10	0.07	0.07
FB	2.06E-08	-4.72E-09	-8.26E-08	-5.70E-08	6.08E-08	-6.00E-08
$R^2_{(adj)}$	0.43	0.40	0.15	0.47	0.44	0.42
IOA	0.63	0.62	0.58	**0.64**	0.63	0.62

The values of the model performance metrics (Table 5.5) and the graphical comparison (scatter plots, Figure 5.19) give a level of confidence in the models' performances. Figure 5.19 illustrates the site-specific scatter plots of the predicted versus observed values of PM_{10} concentrations. The center line (1:1) is given to assist with comparison of the GAM with the ideal linear model. The dashed lines show the within FAC2 region, the top line is 0.5:1 and bottom is the 2:1 line. The closer the FAC2 values to one indicates the closer match between observed and modelled values and therefore better model performance (Barmpadimos et al., 2011); (Sayegh et al., 2014). The added Pearson correlation coefficients (r) show the strength of the linear relationship between the observed and predicted values. For low concentrations, the values fall within the FAC2 region, whereas for higher concentrations scattering is more evident.

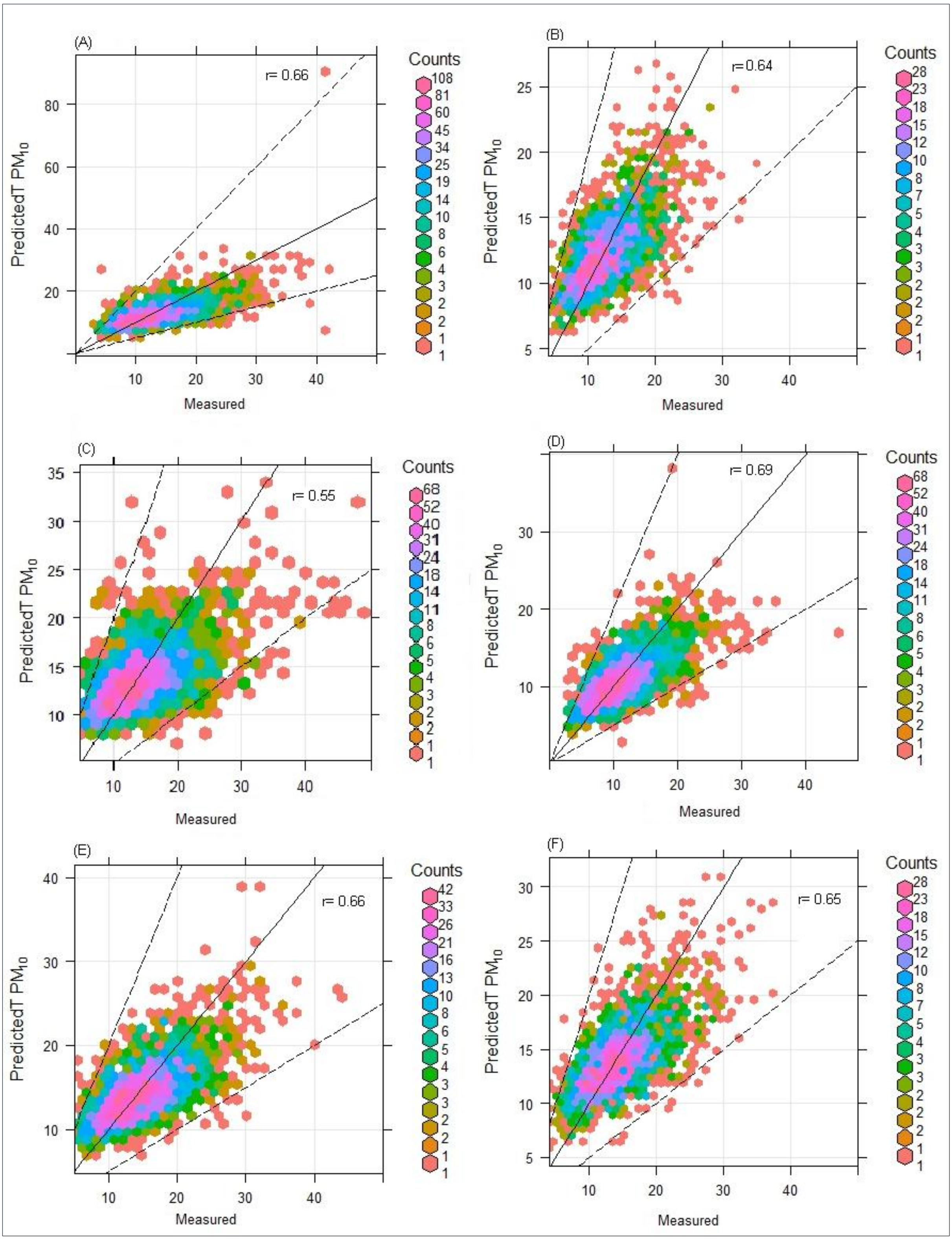


Figure 5.19: Scatter plot of observed and predicted PM$_{10}$ concentrations using cross-validation. Glen Eden (A), Henderson (B), Pakuranga (C), Patumahoe (D), Penrose (E) and Takapuna (F).

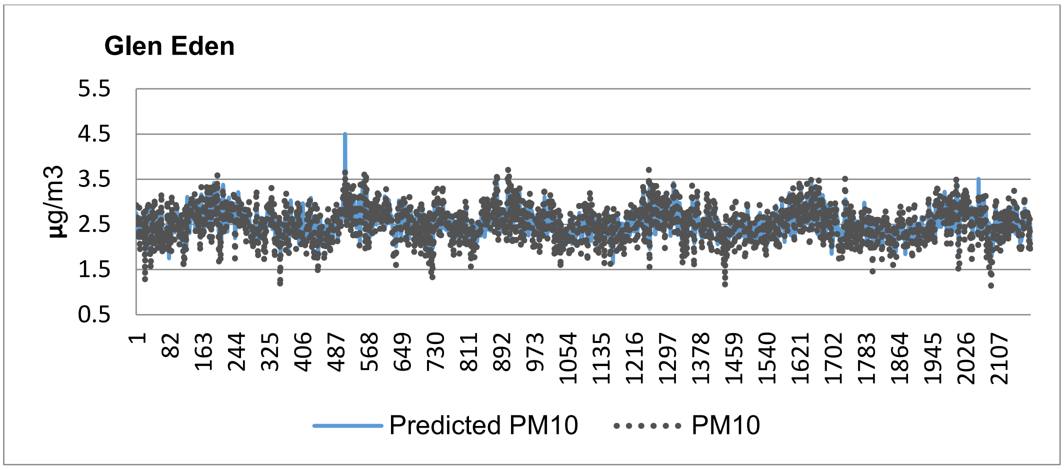


Figure 5.20: Observed and predicted PM_{10} concentrations of each model versus time on test dataset, Glen Eden.

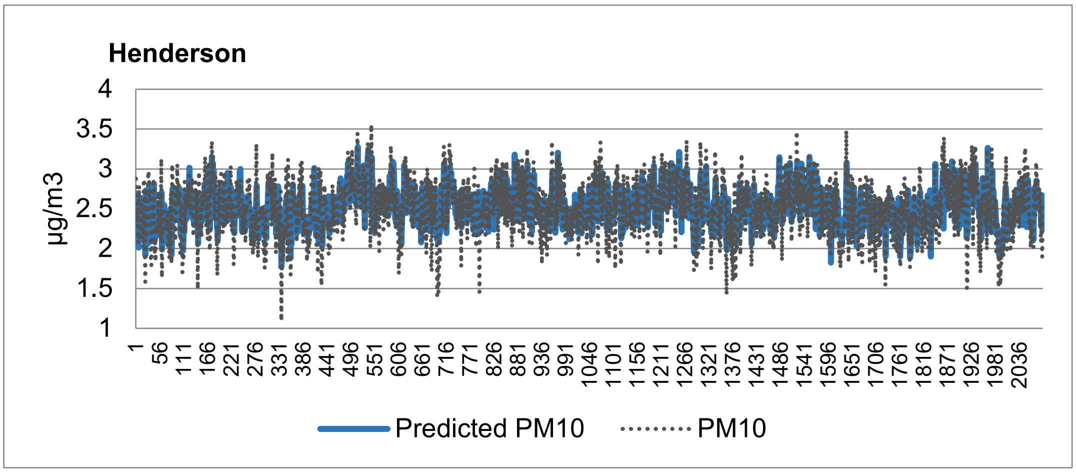


Figure 5.21: Observed and predicted PM_{10} concentrations of each model versus time on test dataset, Henderson.

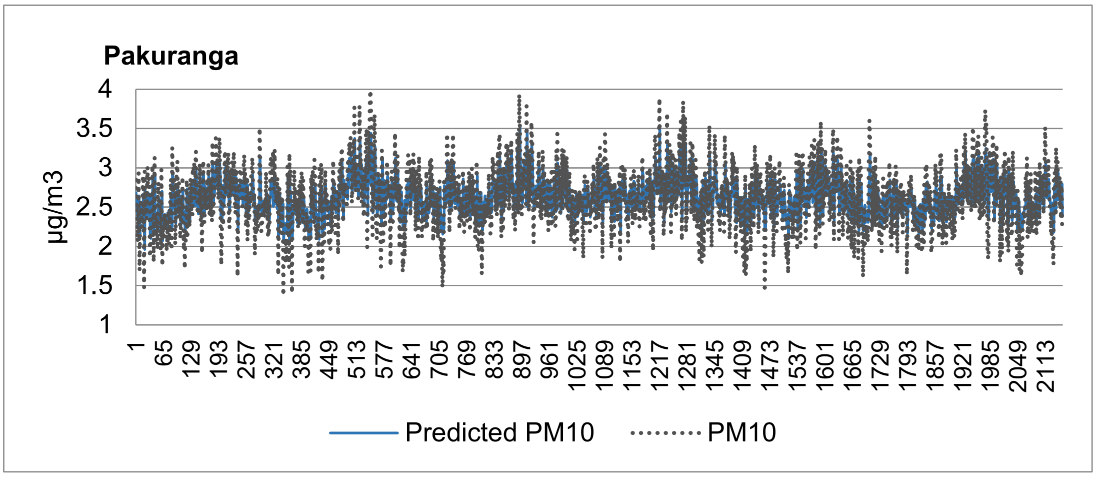


Figure 5.22: Observed and predicted PM_{10} concentrations of each model versus time on test dataset, Pakuranga.

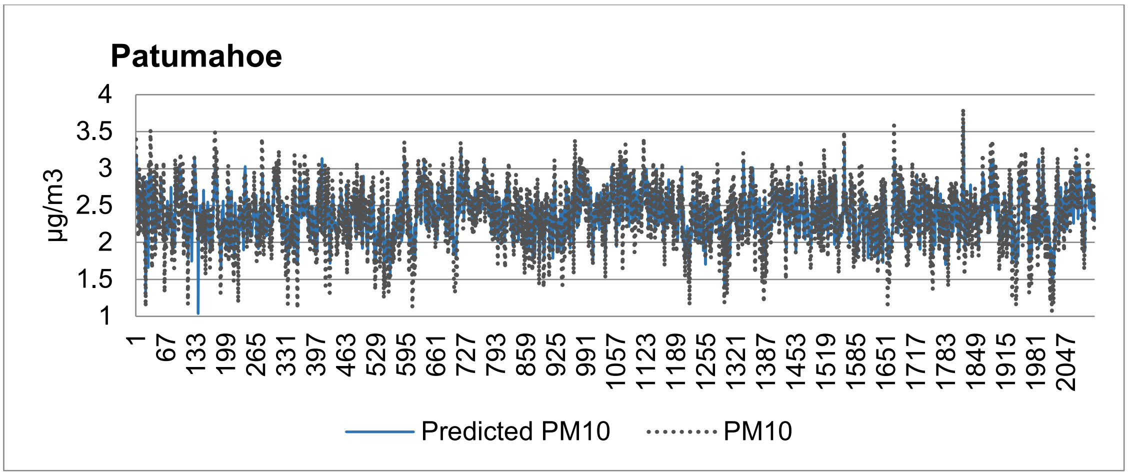


Figure 5.23: Observed and predicted PM_{10} concentrations of each model versus time on test dataset, Patumahoe.

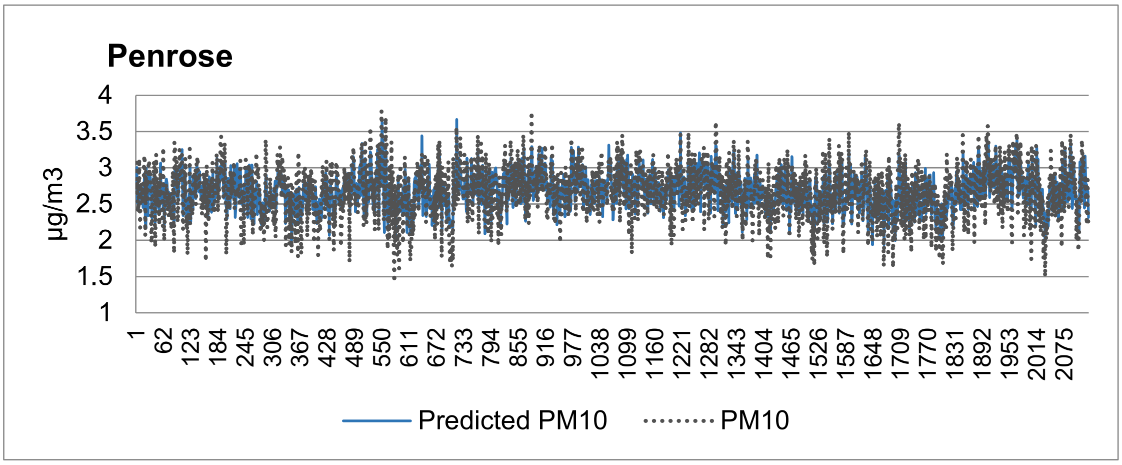


Figure 5.24: Observed and predicted PM_{10} concentrations of each model versus time on test dataset, Penrose.

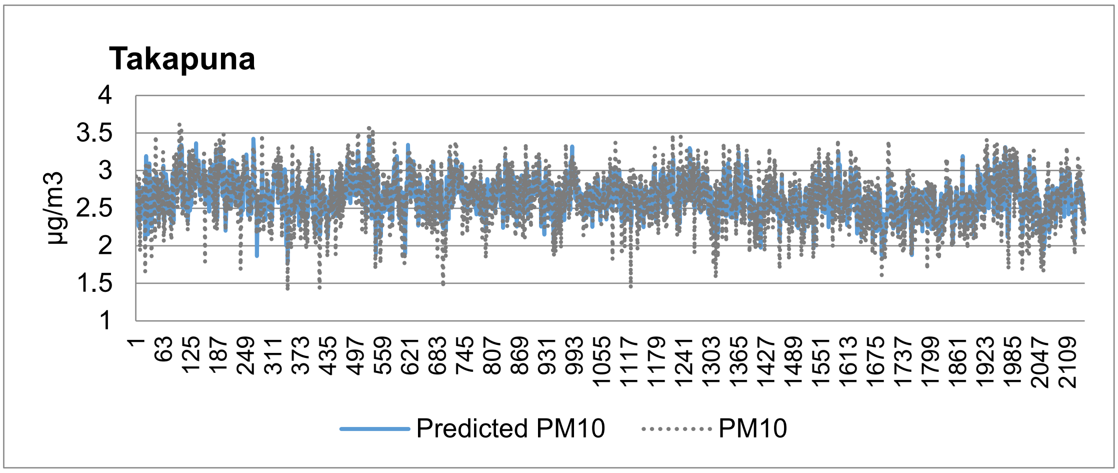


Figure 5.25: Observed and predicted PM_{10} concentrations of each model versus time on test dataset, Takapuna.

Figure 5.20 to Figure 5.25 show observed and predicted PM_{10} concentrations of each model on the test dataset depicting the under-prediction of PM_{10} at lower values.

In these GAMs it was assumed that the errors were independent and identically normally distributed for inference purposes.

5.4.3.2 Discussion

The ACF and PACF presented in Chapter 4 showed that the PM_{10} concentrations exhibit mild autocorrelation. The fitted GAM model assumed the errors are independent and not correlated. The residual autocorrelation plots (Figure 5.13- Figure 5.18) of the GAM outlined above suggests that there is a fair amount of positive autocorrelation (0.421) in Pakuranga at lag_1. This confirms our suspicion that there are patterns in the residuals of the model and that these patterns are likely affecting the model's outputs. The lag_1 value for the other sites is rather less significant ranging from 0.24-0.32. The lag_2 autocorrelation values are not significant with low values mainly between -0.036 and 0.86 as presented in Table 5.6.

Table 5.6: Estimated auto-correlation value for the residuals of GAM models.

	rho	
	Lag_1	**Lag_2**
Glen Eden	0.297	-0.012
Henderson	0.319	-0.002
Pakuranga	**0.421**	0.086
Patumahoe	0.344	0.0578
Penrose	0.242	0.002
Takapuna	0.272	-0.036

The slight observed correlation between neighboring days is probably due to the auto-correlated data generating process (midnight-midnight). In addition, there are some lagged meteorological effects (which are not currently encapsulated within these models) that might affect the PM_{10} concentrations on subsequent days. As the autocorrelation appears to be rather weak it was decided to neglect it in the models at this stage. However, to address this issue a GAMM model is applied where the GAMs are coupled with an AR(1) error model within the time series. The ACF plots (presented in Chapter 4) suggested that the AR(1) is likely sufficient to capture the remaining autocorrelation.

The partial effect results presented in section 5.4.2 showed differences between the effects of anthropogenic sources (crudely assimilated in the model by temporal variables) and those of atmospheric conditions, with the former.

To the author's knowledge, this is the first effort to use GAMs to model PM_{10} concentration in the Auckland airshed, where multiple sources contribute to the observed PM_{10} concentrations. Further work is required to quantify the contribution of traffic sources, being the second largest source for several air pollutants in Auckland's airshed. To do this improved monitoring is needed to provide higher quality emission data over the airshed.

5.4.4 GAMM Model Results and Discussion

In this study as in recent work investigating PM_{10} trends in Christchurch (Scarrott et al., 2009), the meteorological and lag PM_{10} terms in the model are included as fixed effects while the random effect component of the model is representative of the AR(1) structure and the smooth trend. The idea is that by modeling the time of year (*JD*) and autocorrelation properly, then we should be in a good position to establish whether there is a significant overall PM_{10} trend. Similar to the GAM (Section 5.4.2), the association of PM_{10} with meteorological and PM_{10} lag factors in the GAMMs was examined. The individual partial responses of PM_{10} to these terms were extracted and are presented in Figure 5.26 to Figure 5.30. The curves illustrate the partial effects of each predictor on PM_{10} concentration. The dashed lines show the 95% confidence intervals. The error bars are wider where there are fewer observations. The vertical scales show the original scale. Estimated smooth relationships which have been penalized out of the model are not presented, only those acceptable to the model are shown.

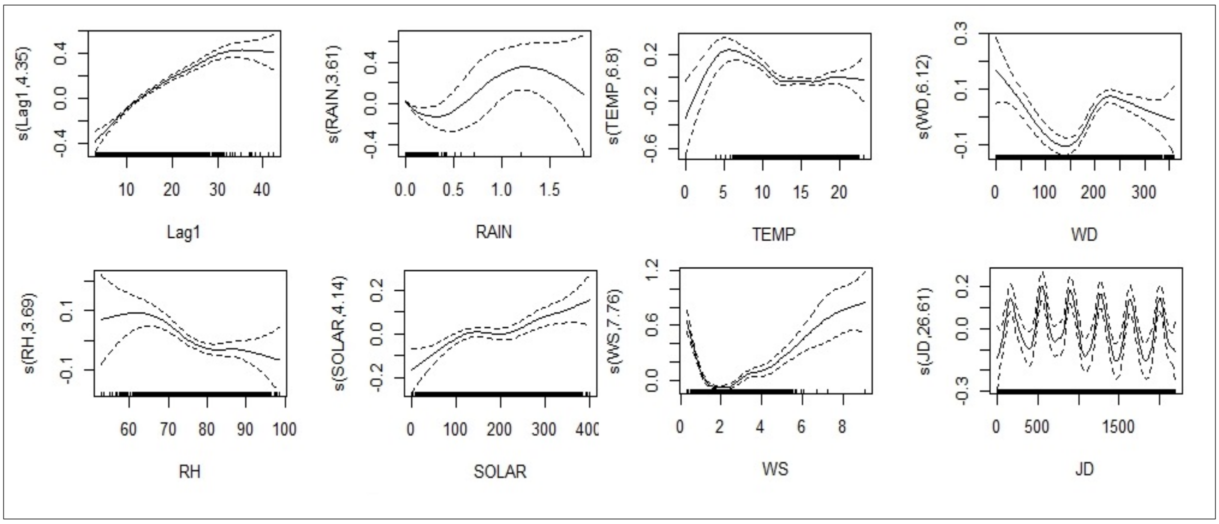


Figure 5.26: Partial responses of PM_{10}, Glen Eden**.**

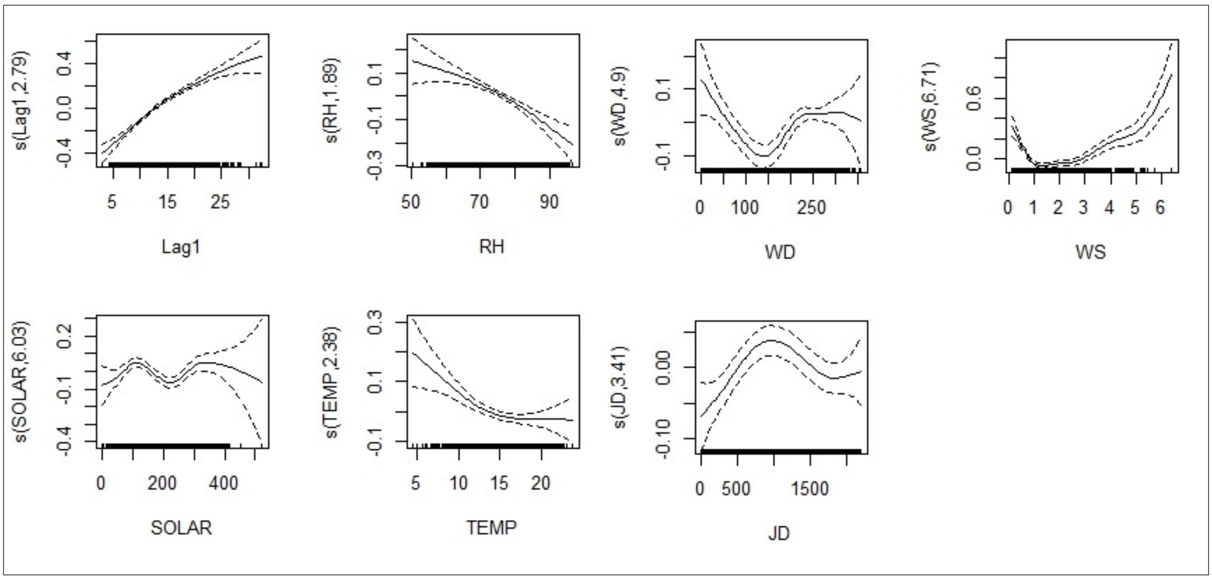


Figure 5.27: Partial responses of PM_{10}, Henderson.

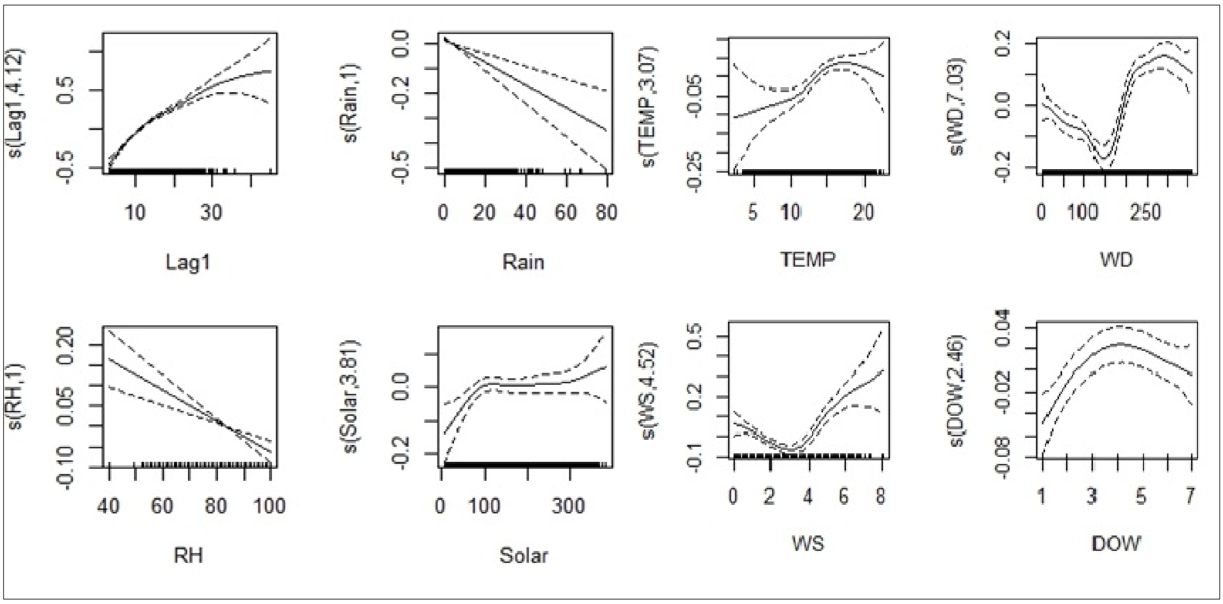


Figure 5.28: Partial responses of PM_{10}, Patumahoe.

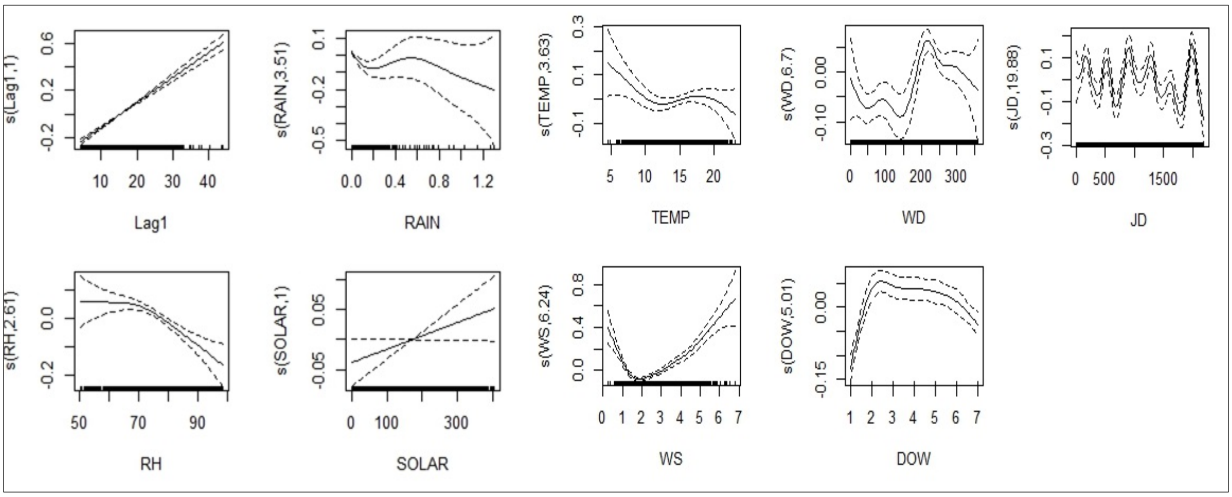


Figure 5.29: Partial responses of PM_{10}, Penrose**.**

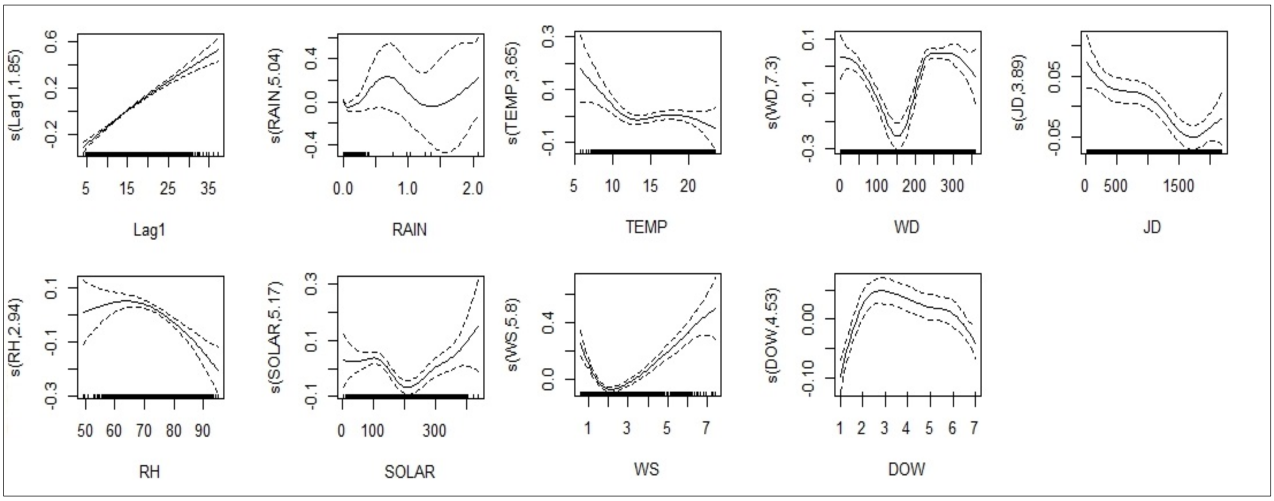


Figure 5.30: Partial responses of PM_{10}, Takapuna.

For all urban sites, a non-linear partial response is depicted with an apparent decline when temperatures increase from 5 °C to 15 °C. All other predictors are constant for the urban sites. In contrast, an increase in partial response is observed with an increase in temperature, within the same temperature range, for the rural Patumahoe site. The effect of RH on PM_{10} concentration is obvious for Glen Eden, Henderson, Penrose, and Takapuna where concentrations initially increase with rising RH up until about 75%. For higher RHs, above ~75%, PM_{10} starts to decline. A negative correlation between PM_{10} concentrations and wind speeds is observed for Pakuranga showing the effect of local ventilation. Maximum PM_{10} concentrations occur as wind speeds increase in the remaining sites showing the effect of stronger winds is the opposite of that recorded for lighter wind speeds; that is, pollution transport to the site via advection predominates over local ventilation. Similar results were also found for GAM. A nonlinear relationship between rainfall and PM_{10} is observed. At lower rainfall intensities, the presence of rainfall reduces PM_{10} concentrations in Glen Eden and Patumahoe. The large confidence intervals suggest that the curvature in the relationship between rainfall and PM$_{10}$ concentration is real. Takapuna exhibited no clear pattern, probably due to the low daily mean rainfall values. Like GAM, GAMM eliminated rainfall as a factor in the models for Henderson and Pakuranga.

5.4.4.1 Model Evaluation and Discussion

The site-specific standard diagnostic plots of the GAMM models are shown in Figure 5.31. A reasonable distributional assumption is suggested by the *QQ*-plot (A). Variance is nearly constant Plot (B), and the residuals histogram (C) seems consistent with normality. Response versus fitted values (D) confirms a positive linear relation with reasonable scatter.

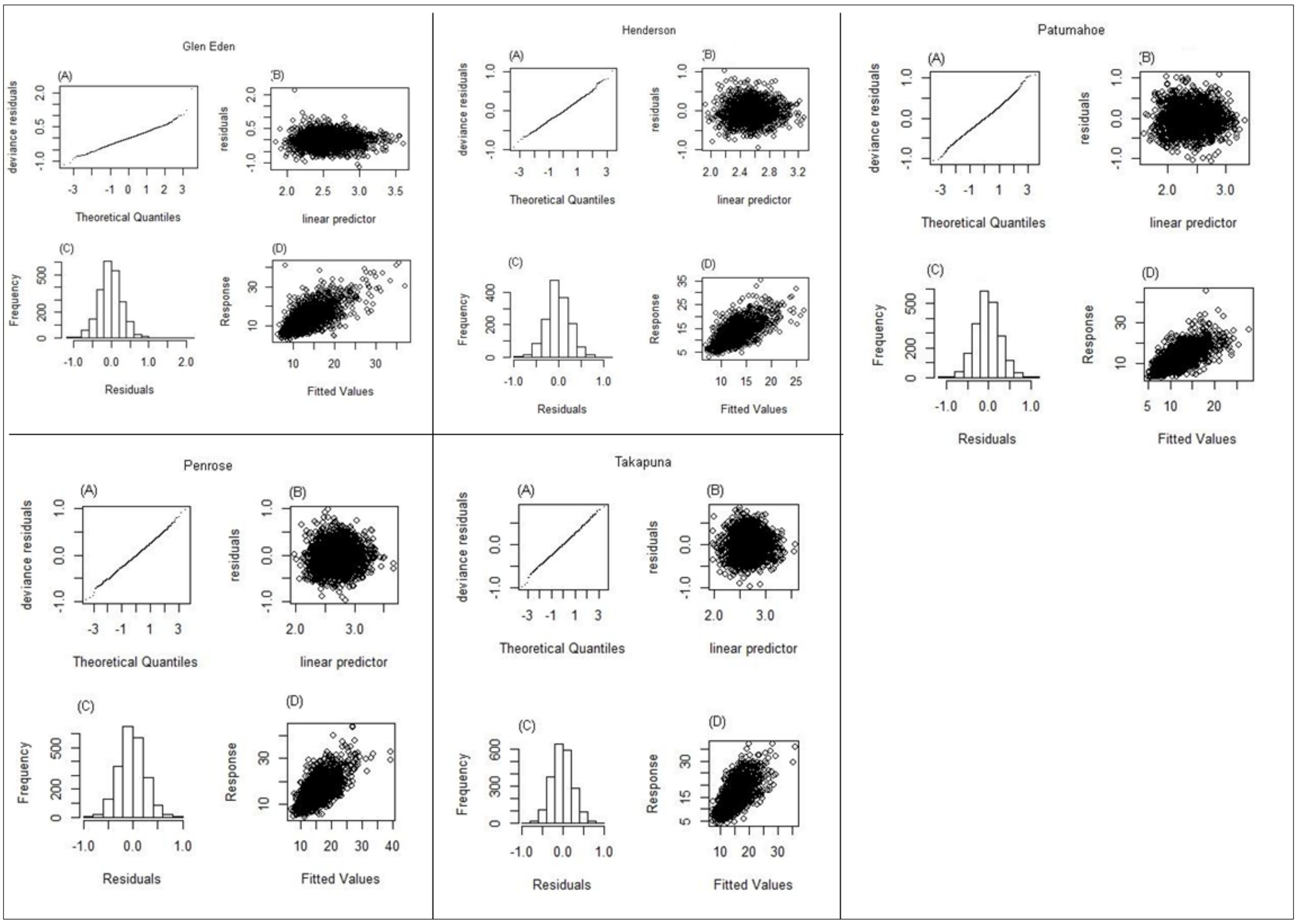


Figure 5.31: GAMM diagnostic plots.

Table 5.7 depicts the result of fitted GAMMs by site. Results shows the selected terms explained 50% of the variance in Glen Eden. The GAMM relating meteorological and lag variables to the total daily average PM$_{10}$ for Pakuranga explained only 37.4% of the PM$_{10}$ deviance.

Table 5.7: Statistical evaluation of the GAMM model on the log scaled PM$_{10}$ concentrations for the entire dataset.

	MAE	MSE	RMSE	MAPE	R^2 (adj)	Deviance explained $^{(\%)}$
Glen Eden	0.23	0.09	0.30	0.10	0.48	50.00
Henderson	0.2	0.06	0.25	0.08	0.43	44.80
Pakuranga	0.25	0.10	0.32	0.10	0.36	37.40
Patumahoe	0.23	0.08	0.29	0.10	0.48	50.70
Penrose	0.21	0.07	0.26	0.08	0.46	50.00
Takapuna	0.20	0.06	0.25	0.08	0.47	0.49

Performance of the GAMMs was evaluated using a 10-fold cross-validation. The statistical evaluation of the GAMMs, for all sites, is presented in Table 5.8 and shows the result of cross validation of the fitted GAMMs on each site's testing dataset. The positive values of FB indicate under-prediction. However, the calculated values of FB are too small and close to zero indicating that there is no systematic tendency to under prediction. Glen Eden shows the highest R^2 of 0.48.

Table 5.8: Statistical evaluation of the GAMM model on the log scale for test set.

	MAE	MSE	RMSE	MAPE	R^2 (adj)	FB
Glen Eden	0.24	0.09	0.31	0.10	0.48	1.99E-07
Henderson	0.21	0.07	0.26	0.09	0.40	-7.33E-08
Pakuranga	0.25	0.10	0.32	0.10	0.32	-1.81E-07
Patumahoe	0.23	0.09	0.30	0.11	0.47	-2.1E-07
Penrose	0.21	0.07	0.27	0.08	0.44	-1.9E-08
Takapuna	0.21	0.07	0.26	0.08	0.42	4.0E-05

5.5 Conclusion

In the GAM/GAMM models (normal family with log link) investigated the mean of the distribution of PM_{10} concentration varied in a log linear form with the meteorological and temporal terms. The GAMM model included an AR(1) with a site-specific estimated autocorrelation and explained some of the smooth variation in the PM_{10} concentrations in urban sites. The GAMM's estimated smooth trend captured less variability when compared to that of the GAM. The terms selected by the GAM were verified again for the GAMM with an AR(1) component. However, some of the terms that were identified as insignificant in the GAMs, were selected for the GAMMs.

The predictive ability of the GAMM models, developed in this doctoral research, for Auckland's PM_{10} was lower ($0.36 < R^2 <0.48$) than those reported for modeling PM_{10} in the Northeastern and Midwestern U.S. ($R^2 = 0.58$) (Yanosky et al., 2014). The GAM model in this research ($0.20 < R^2 < 0.44$) was comparable to the GAM model ($R^2 = 0.49$) for estimating annual average PM_{10} concentration in the U.S. (Yanosky et al., 2014) using rich dataset of both meteorological and geographical data. The R^2 values for hourly PM_{10} GAM models for a few different sites in Oslo by Aldrin and Haff (2005) were reported to have ranged between 0.48 and 0.72. In the Oslo models both meteorological and traffic volume variables were employed.

The relatively poor performance of the GAM model developed in this research when compared to those reported by Aldrin and Haff (2005) could be due to lack of geographical and emission information in Auckland PM_{10} models. This data was not available at a matching temporal resolution to provide analytical results for this study. However, in the future additional data on the emissions inventory and mechanism of conversion and deletions as well as distance to road and land use could be incorporated into the model to further investigate their impact on the Auckland PM_{10} concentrations.

Chapter 6 Spatio-Temporal Statistics and Models

This Chapter provides an overview of statistical Spatio-Temporal (S-T) modeling techniques for analysis and prediction of continuous S-T PM_{10} data over the Auckland urban area. Various approaches exist for building statistical models to facilitate this analytical process. These approaches include techniques for determining model parameters and estimating model outcomes in the form of occurrence predictions. This Chapter surveys these methods and conclude with the identification of the most suitable or fit-for-purpose approaches and models in the context of this research. S-T statistics is a vast field that it cannot be explored in its entirety. Therefore, in this study the focus is mainly on data with spatial reference as point, without exploring issues associated to the “change-of-support” problem or dealing with S-T point processes.

This Chapter is structured to reflect the way that statistical modeling in general and S-T modeling in particular are typically approached. The first stage of data exploration is presented, and the results reported in Section 6.2. S-T statistical models are then reviewed from an introductory point of view in Section 6.3. In Section 6.3.2 a deterministic approach on modeling PM_{10} concentration is discussed through Inverse Distance Weighting (IDW). IDW is the simplest method to implement S-T prediction. Spatial prediction experiments based on basic statistical models, those that do not account for S-T structure (linear regression with trend, GLM, and GAMs), are subsequently presented in Section 6.3.3. Fitting these simple statistical models to the PM_{10} data using the spatial properties of data helped to reveal potential patterns and check for any violation of assumptions.

In this Chapter, to facilitate connections between the theoretical basis of methods and the models developed, the experiments and results are provided at the end of each section after describing the relevant method. The models reported in this Chapter were developed using R and MATLAB R2020b.

The results presented in graphical format in this Chapter are for the colder months (May-Aug) (year-specific). The choice of colder months was due to the presence during these months of higher PM_{10} variation (discussed in Chapter 3). Since the interpretation of graphs are typical for these months, May 2015 was selected as an example as it exhibited more distinct color variations making it easier for the reader to interpret. The numerical results for the entire dataset are presented as a summary for all years (and months if applicable) in tabular format. The graphical results (graphs) are accessible via (Dropbox link) due to high volume of generated graphs (each experiment generated (6 years × 12 months = 72) graphs) and practically they could not be presented in this thesis in their entirely.

6.1 Introduction

While scientists have always been interested in S-T data (e.g., Kepler 1571-1630), it is only recently, in the last three decades, that statisticians have focused in depth on this area of statistics (Cressie & Christopher, 2011). Spatio-temporal statistical properties are ubiquitous in real world data and have been recognised as a statistical challenge because of the nature of spatial statistical dependence and its associated errors. The issue of spatial dependency was noted by Fisher, a pioneer in statistics:

> "*…After choosing the area we usually have no guidance beyond the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are further apart.*" (Fisher, 1935, p. 66).

To overcome errors related to spatial dependence Fisher (1935) introduced randomization into his experimental design. This randomization helped to avoid confounding plot and treatment effects.

The pollution process can be perceived as the realisations of random or stochastic processes in geostatistical terms, assuming PM_{10} is a partial realisation of a S-T random function (De Iaco & Posa, 2012), a random process (Z) of PM_{10} can be written as:

$$\mathbf{Z} = \mathbf{Z}\{\mathbf{z}(\mathbf{s},\mathbf{t}), (\mathbf{s},\mathbf{t}) \in \mathbf{D} \times \mathbf{T}\} \qquad \text{Eq. 6.1}$$

where

$D \in R^2$ and $T \in R$.

To accommodate both non-constant spatial and temporal trends, the S-T random function can be decomposed into a mean and a residual component (De Iaco & Posa, 2012):

$$\mathbf{Z}(\mathbf{s},\mathbf{t}) = \mathbf{m}(\mathbf{s},\mathbf{t}) + \mathbf{R}_{\mathbf{es}}(\mathbf{s},\mathbf{t}) \qquad \text{Eq. 6.2}$$

where

$m(s,t)$ is a space time deterministic mean model and

$R_{es}(s,t)$ is residual component with assumption of a stationary random field

The deterministic S-T mean component $m(s,t)$ can then be decomposed into a spatial trend $m_1(s)$ and temporal trend $m_2(t)$ (Kyriakidis & Journel, 1999):

$$\boldsymbol{m}(\boldsymbol{s},\boldsymbol{t}) = \boldsymbol{m}_1(\boldsymbol{s}) + \boldsymbol{m}_2(\boldsymbol{t}) \qquad \text{Eq. 6.3}$$

The estimated S-T mean component is then deducted from the full dataset to return the S-T residuals component $R_{es}(s, t)$ which is used in later analyses. Such deterministic trend functions of the S-T coordinates have been implemented in previous studies (Spadavecchia & Williams, 2009).

6.2 Exploratory Analysis of Spatio-Temporal Data

The work that follows in this section explores the empirical S-T means and covariance estimates of the Auckland PM$_{10}$ data set used in this thesis.

6.2.1 Empirical Spatial Means and Covariances

The empirical spatial mean $\hat{\mu}_{z,s}(s_i)$ for location s_i is formulated as (C. K. Wikle et al., 2019c):

$$\hat{\boldsymbol{\mu}}_{\mathbf{z,s}}(\mathbf{s_i}) = \frac{\mathbf{1}}{\mathbf{T}}\sum_{\mathbf{j=z}}^{\mathbf{T}} \mathbf{Z}(\mathbf{s_i}; \mathbf{t_j}) \qquad \text{Eq. 6.4}$$

where

$Z(s_i; t_j)$ are the PM$_{10}$ observations,

$\{s_i : i = 1, . . . , m\}$ are spatial locations, and

$\{t_j : j = 1, . . . , T\}$ are observations time.

The spatial mean as an *m*-dimensional vector can be written as:

$$\hat{\boldsymbol{\mu}}_{\mathbf{z,s}}(\mathbf{s_i}) = \frac{\mathbf{1}}{\mathbf{T}}\sum_{\mathbf{j=z}}^{\mathbf{T}} \mathbf{Z}_{\mathbf{t_j}} \qquad \text{Eq. 6.5}$$

where

$$\mathbf{Z}_{\mathbf{t_j}} \equiv \left(\mathbf{Z}(\mathbf{s_1}; \mathbf{t_j}), \dots, \mathbf{Z}(\mathbf{s_m}; \mathbf{t_j})\right)' \qquad \text{Eq. 6.6}$$

The mean vector is a spatial quantity indexed by a relevant location. The number of observations at each location may not be the same due to missing or incorrect values. To calculate the average in such situations the mean vector should be calculated for each location separately:

$$\hat{\boldsymbol{\mu}}_{\mathbf{z,s}}(\mathbf{s_i}) = \frac{\mathbf{1}}{\mathbf{T_i}}\sum_{\mathbf{j=1}}^{\mathbf{T_i}} \mathbf{Z}(\mathbf{s_i}; \mathbf{t_j}) \qquad \text{Eq.6.7}$$

where T_i is the number of time points at location s_i.

The computed empirical spatial mean of PM_{10} concentration is calculated according to Eq.6.7 and the result is illustrated in Figure 6.1. A slight trend in the empirical spatial mean of PM_{10} with longitude can be observed Figure 6.1 (a), but no trend can be detected for latitude Figure 6.1 (b).

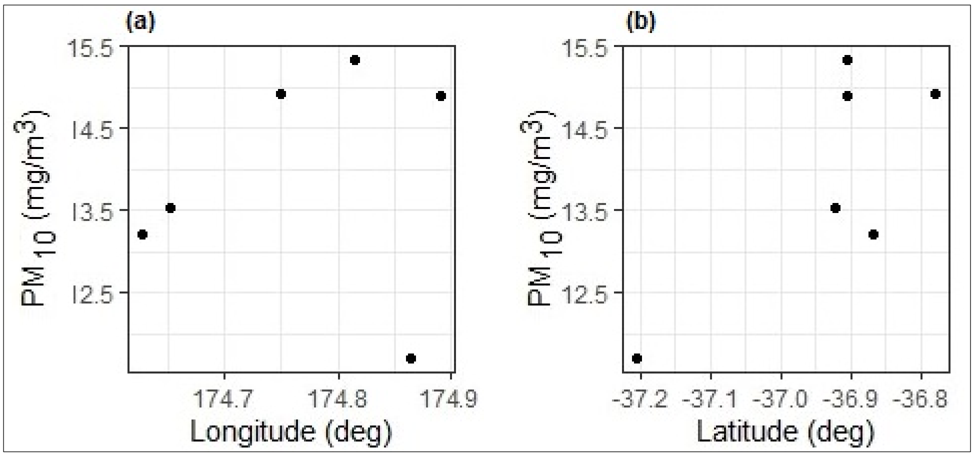


Figure 6.1: Empirical spatial mean of PM_{10} concentration in the Auckland study area as a function of (a) site longitude, (b) site latitude.

The empirical temporal mean, for time t_j , $\hat{\mu}_{z,t}(t_j)$ was calculated by taking the average across spaces follows:

$$\hat{\mu}_{z,t}(t_j) = \frac{1}{m}\sum_{i=1}^{m} Z(s_i; t_j) \qquad \text{Eq. 6.8}$$

The empirical temporal mean for Auckland's PM_{10} was computed according to Eq. 6.8. The plot shown in Figure 6.2 can be used in order to visualise the seasonal variation of PM_{10} concentration over the study area. The plot shows time series of the Auckland PM_{10} data set averaged across all available site's spatial locations where the blue line resembles an individual site, and the black line signifies the empirical temporal mean. A slight variation in PM_{10} concentration (higher during winter) can be seen. Variations in the seasonal pattern in each year were also noted as part of the time series analysis presented in Chapter 4.

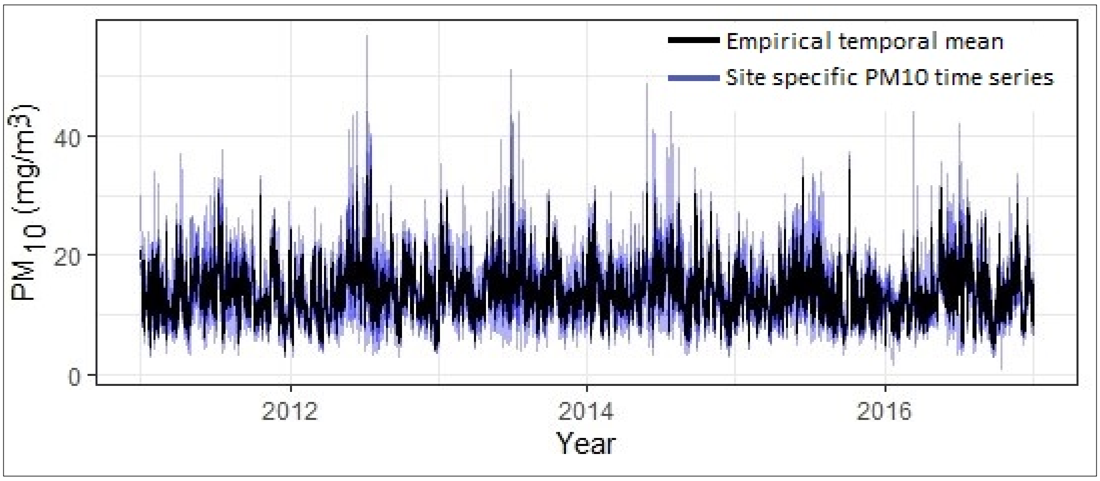


Figure 6.2: The empirical temporal mean (black line) and time series of PM10 (blue line). Each blue line corresponds to each site.

These empirical results reveal the existence of spatial and temporal co-variability in the PM_{10} concentration over the study area that must be considered in the S-T modeling of PM_{10}. In addition, the joint S-T dependence structure of the PM_{10} S-T process needs to be characterised in order to be able achieve the best predictive models possible (this aspect is discussed in Chapter 7).

6.3 Spatio-Temporal Statistical Models

6.3.1 Evaluating Spatio-Temporal Statistical Models

Model building is an iterative process in the sense that models are built around the data and/or a scientific assumption. The appropriateness of the model representation of the real world is then evaluated and modified accordingly if it is not. The model-evaluation approach taken in this Chapter involves the steps of model *checking*, model *validation*, and model *selection*. There is very limited discussion of the topic in the S-T modeling literature hence the citations included in this section are not extensive.

Model Checking: Models are evaluated to check the underlying assumptions and model output sensitivity to these assumptions and/or model choices. In this study, the S-T models are evaluated using statistical tests. In S-T regression models and S-T GLMs, analysis was first performed by evaluating the residuals. For additive Gaussian measurement error, the S-T residuals were evaluated. Quantitative summaries, such as with the PACF, Moran's Index, and S-T co-variogram, to evaluate residual temporal, spatial, and S-T dependence are also considered as measures of model performance. The changes in predictions made by the models were evaluated by altering the number of the basis functions or the degree of spatial dependence in an error distribution. For IDW, the optimal value was selected using cross-validation.

Model Validation: Spatio-temporal processes have a unidirectional time dependence and different degrees of spatial dependence that should be considered during S-T model evaluation. In this research the validation sets were obtained by leaving out data at random (random set). The random validation set was created by randomly sampling 30% of the data from the PM_{10} observations. To ensure that the S-T dependence structure is characterised in such way that it sufficiently fills in large gaps for S-T processes a second validation set was created by leaving out blocks of data (temporally sequential blocks) from the original data set (block set). In other words, a second validation set (block set) was created by leaving out a period of PM_{10} data prior to fitting the model. The indices of the observations, collected at seven arbitrary time points within each month, were selected and the data at those indices was removed.

All models presented in the following sections are fitted to the training data sets and their performance is validated using their related validation sets. In S-T models with complex dependence, the cross-validation scheme should respect the dependence structure. In this study the metrics suggested in literature (C. K. Wikle et al., 2019a) namely the bias, the Predictive Cross-Validation (PCV) and the Standardized Cross-Validation (SCV) measures (Kang et al., 2009), and the Continuous Ranked Probability Score (CRPS) (Zamo & Naveau, 2018) were considered for model validation. The most common scalar validation statistic for continuous-valued S-T processes is the Mean Squared Prediction Error (MSPE) as it is an empirical measure of expected squared error loss which, when minimized, results in the S-T kriging predictor. Therefore, in addition to the above-mentioned diagnostics MSPE was also calculated and used as validation statistics in this Chapter.

6.3.2 Deterministic Methods

Pure spatial interpolation is the process of forming a prediction of the values of unknown points using measurements at isolated points within the same area and forming a prediction surface. The theoretical basis of spatial interpolation is Tobler's First Law (TFL) of geography: "*Everything in space is related to every other thing but points close together are more likely to be similar than the points which are far apart*". (Tobler, 1970)

There has been an increasing interest in the use of S-T interpolation of ambient air quality levels in recent years. The primary strategy adopted by researchers, as identified by a survey of the literature, is to reduce the S-T interpolation task to a sequence of spatial interpolation snapshots (Li, 2008; Li & Revesz, 2004; Yu & Wang, 2013).

IDW was used by Krasnov et al. (2016) to analyze distributions of $PM_{2.5}$ and PM_{10} during the dust storm events in Israel. Li et al. (2016) reported on an extension of IDW used on $PM_{2.5}$ data within a space-time domain. Their extended method integrated space and time simultaneously by considering time as another dimension in space.

In this section the IDW method is described and applied to Auckland PM_{10}. In this case IDW is developed by considering nearby observations in both space and time with the aim of predicting the next day's PM_{10} concentration over the entire study area. This is one of the advantages of such an approach as most other methods produce a set of models were each model is specific and local to a specific monitoring site. Such approaches do not provide a holistic view of the PM^{10} situation over a wider area.

The IDW method is a deterministic exact interpolation model that follows Tobler's first law. To perform S-T prediction using IDW the data is averaged so that more weight is given to the nearest observations in space and time. Suppose we have S-T data such that for each time t_j there is m_j observations. The IDW predictor at location s_0 and time t_0 is calculated according to (C. K. Wikle et al., 2019c) :

$$\hat{\mathbf{Z}}(\mathbf{s_0}; \mathbf{t_0}) = \sum_{\mathbf{j=1}}^{\mathbf{T}} \sum_{\mathbf{i=1}}^{\mathbf{m_j}} \mathbf{w_{ij}}(\mathbf{s_0}; \mathbf{t_0})\, \mathbf{Z}(\mathbf{s_{ij}}; \mathbf{t_j}) \qquad \text{Eq. 6.9}$$

where

$$\mathbf{w_{ij}}(\mathbf{s_0}; \mathbf{t_0}) = \frac{\widetilde{\mathbf{w}}_{\mathbf{ij}}(\mathbf{s_0}; \mathbf{t_0})}{\sum_{\mathbf{k=1}}^{\mathbf{T}} \sum_{\mathbf{=1}}^{\mathbf{m_k}} \widetilde{\mathbf{w}}_{\mathbf{k}}(\mathbf{s_0}; \mathbf{t_0})} \qquad \text{Eq. 6.10}$$

and

$$\widetilde{\boldsymbol{w}}_{\boldsymbol{ij}}(\boldsymbol{s_0}; \boldsymbol{t_0}) = \frac{\mathbf{1}}{\boldsymbol{d}((\boldsymbol{s_{ij}}; \boldsymbol{t_j}), (\boldsymbol{s_0}; \boldsymbol{t_0}))^{\propto}} \qquad \text{Eq. 6.11}$$

where

$s_{ij}; t_j$ is the S-T location,

$(s_0; t_0)$ is the prediction location,

$d((s_{ij}; t_j), (s_0; t_0))$ is the distance between the S-T location and prediction location and the power coefficient $\propto$ is the smoothing parameter.

In deterministic interpolators, $\propto$ is usually selected through cross-validation. Measurement uncertainties are not considered in deterministic methods nor is the resultant prediction uncertainty. To overcome the problems associated with measurement uncertainty in exact interpolators, Cressie and Huang (1999) suggested to make the weights in Eq. 6.11 proportional to:

$$\frac{\mathbf{1}}{(\mathbf{d}(.,.)+\mathbf{C})^{\propto}} \qquad \text{Eq. 6.12}$$

where

$C > 0$ and

$d(.,.)$ is the distance.

The overall prediction performance can be evaluated using cross validation. Inexact interpolators, such as that of Cressie and Huang (1999), have the capability to apply a smoothing operation to remove any measurement errors by iteratively cross validating the output surfaces and thus minimising the Root Mean Square Prediction Error (RMSPE).

6.3.2.1 IDW Experiments and Results

Parameter Selection: The smoothing parameters for both IDW and the Gaussian kernel are chosen in a way that minimizes the Leave-One-Out Cross-Validation **(**LOOCV) MSPE scores. To perform LOOCV the pairwise distances between observed locations and the related kernel-weight matrix were calculated. For every observation, prediction at a left-out observation was performed by selecting rows and columns from the resulting matrix. The LOOCV score was calculated for a set of five plausible bandwidths. The IDW's optimal inverse-power (α) and the Gaussian kernel smoother's bandwidth (θ) were selected based on minimum LOOCV score. The lowest cross-validation score for IDW was lower than the lowest cross-validation score for the Gaussian kernel smoother (Table 6.1), suggesting that the IDW performs better than the Gaussian kernel smoother on PM_{10} data for all the years in the case study.

Table 6.1: Optimal power, bandwidth parameter and minimum LOOCV score for IDW and Gaussian Kernel Smoother.

Year	IDW		Kernel	
	CV-score	α	CV-score	θ
2011	10.76	1.72	10.98	0.34
2012	12.90	1.56	14.41	0.40
2013	12.37	1.56	13.38	0.45
2014	11.50	1.75	12.48	0.47
2015	7.65	1.87	8.35	0.35
2016	9.10	1.87	9.66	0.35

The scores obtained were then plotted as a function of α and θ for $PM_{10.}$ As an example, the results for 2016 are presented in Figure 6.3. The plots indicate that α = 1.87 and θ = 0.35 may give the best out-of-sample predictions for the 2016 dataset. The plots of remaining years are provided in Appendix C (1).

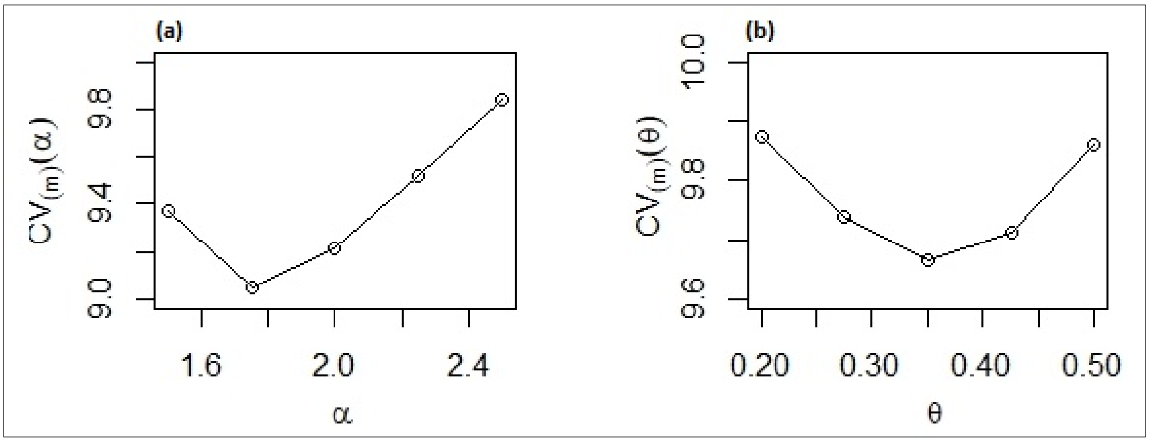


Figure 6.3: The LOOCV score on different range of α and θ for IDW prediction (a) and Gaussian kernel prediction (b) of PM_{10}, 2016.

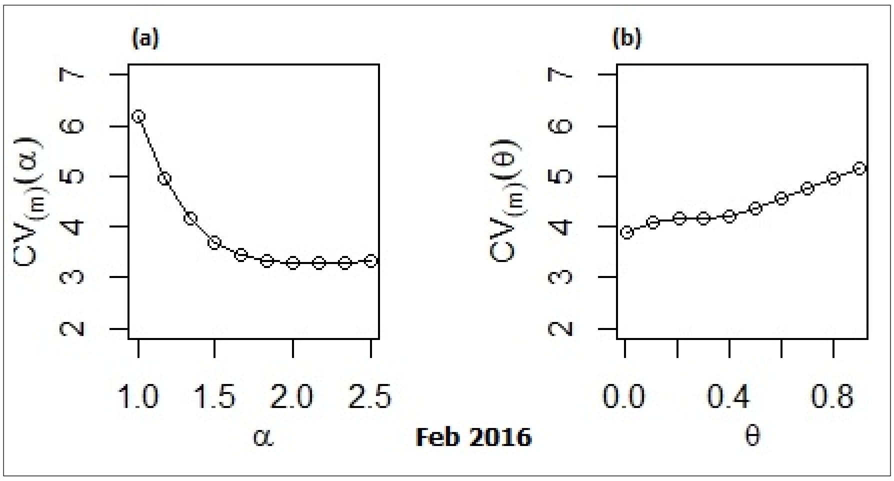


Figure 6.4: The LOOCV score on different range of α and θ for IDW prediction (a) and Gaussian kernel prediction (b) of PM_{10}, February 2016.

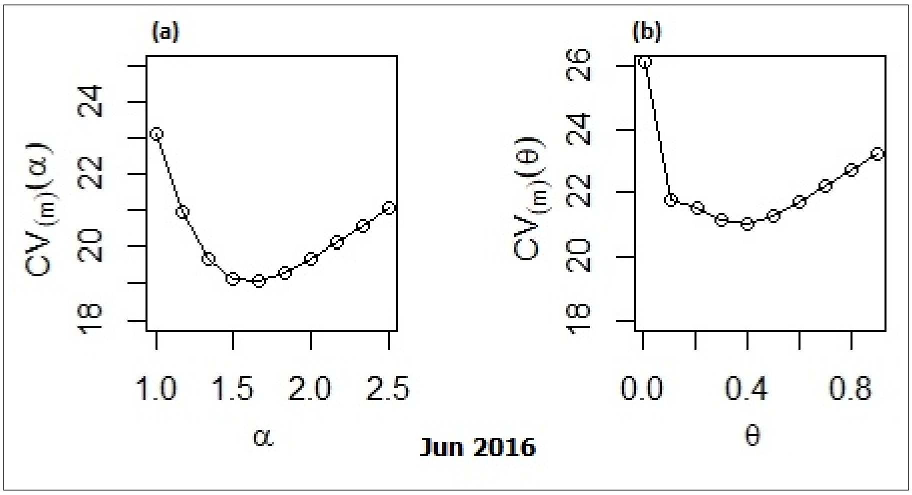


Figure 6.5: The LOOCV score on different range of α and θ for IDW prediction (a) and Gaussian kernel prediction (b) of PM_{10}, Jun 2016.

To perform a prediction of PM_{10} concentration over the study area a 10 by 10 three-dimensional S-T prediction grid in longitude, latitude and a sequence of 365 days regularly arranged in year was created. Due to lack of high spatial resolution for meteorological factors only spatial (latitude and longitude) and temporal variables (days) were used as input variables. Figure 6.6 to Figure 6.11 show the IDW predictions of PM_{10} concentration within the prediction grid enclosing the domain of interest for a random week spanning the temporal window of May to August for each year using the α parameter obtained via LOOCV.

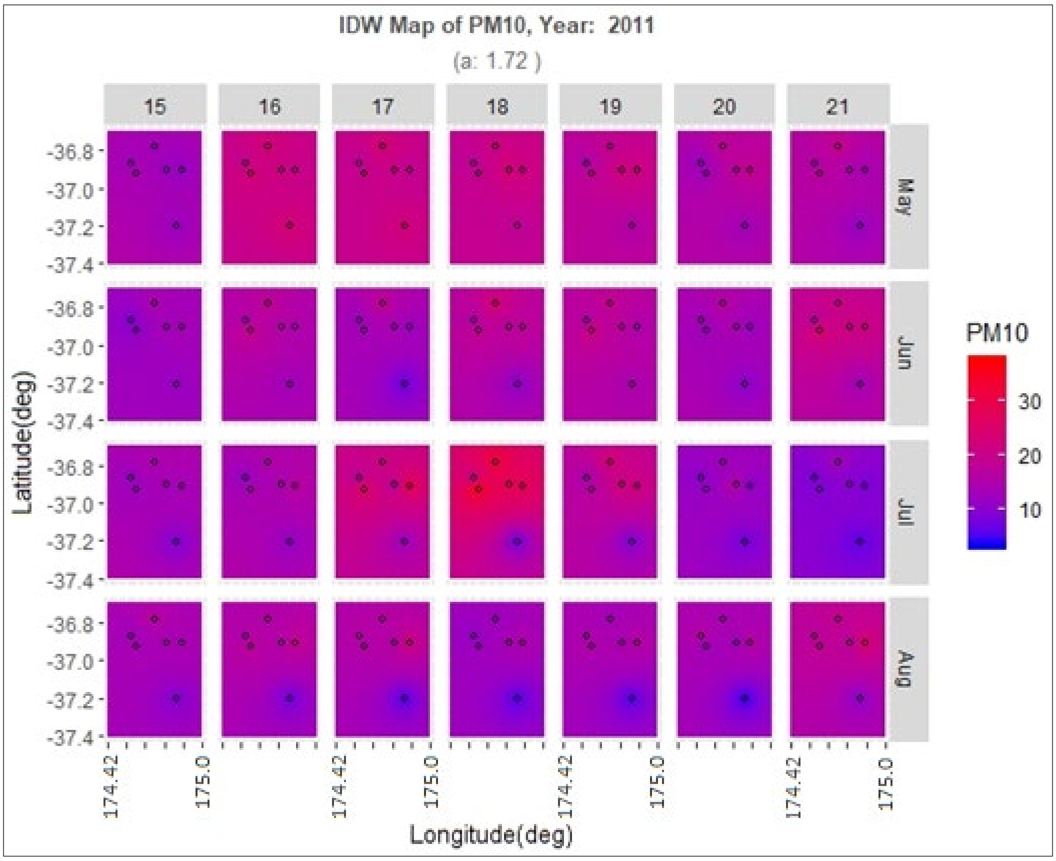


Figure 6.6: IDW prediction map of PM_{10} concentration, May-Aug 2011.

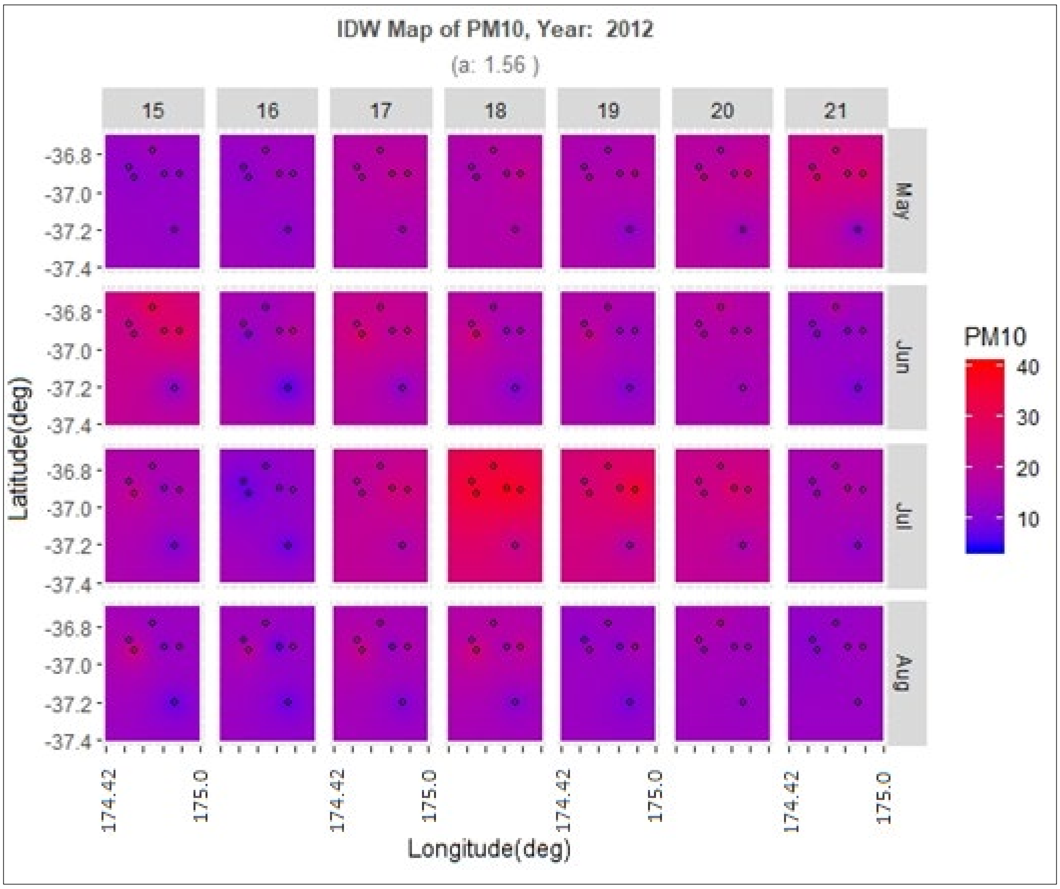


Figure 6.7: IDW prediction map of PM_{10} concentration, May-Aug 2012.

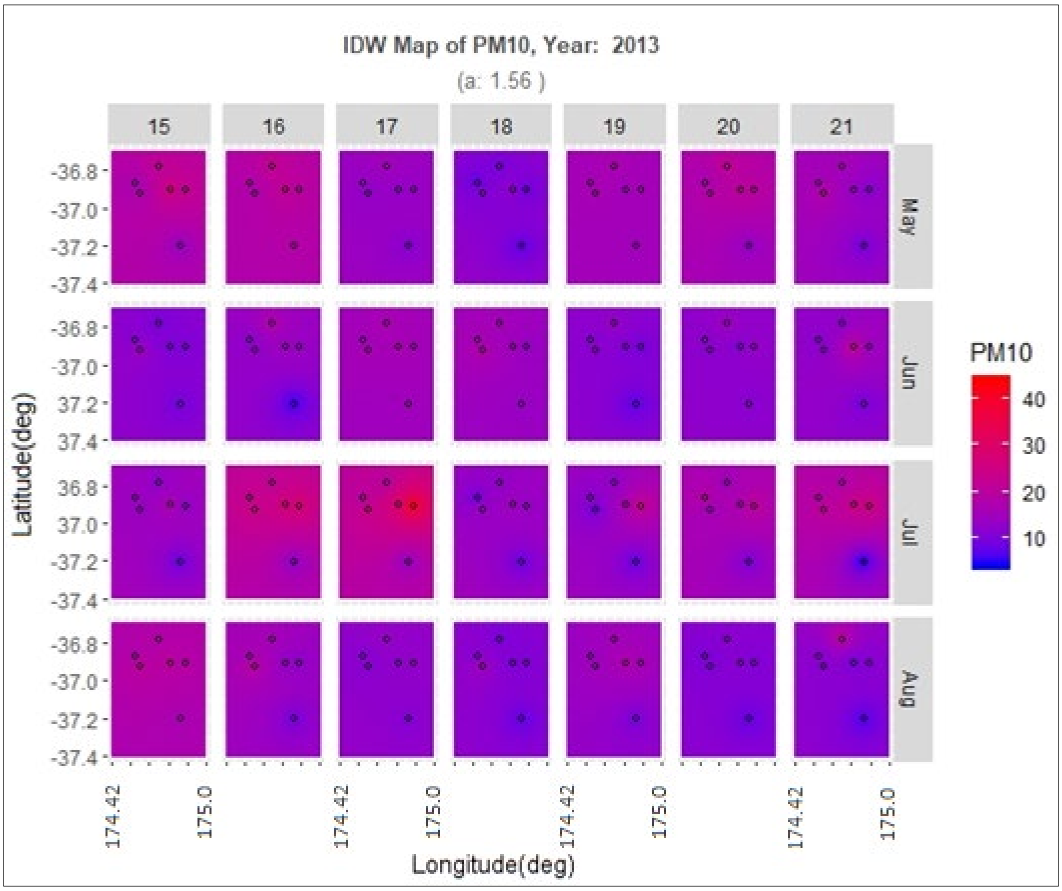


Figure 6.8: IDW prediction map of PM_{10} concentration, May-Aug 2013.

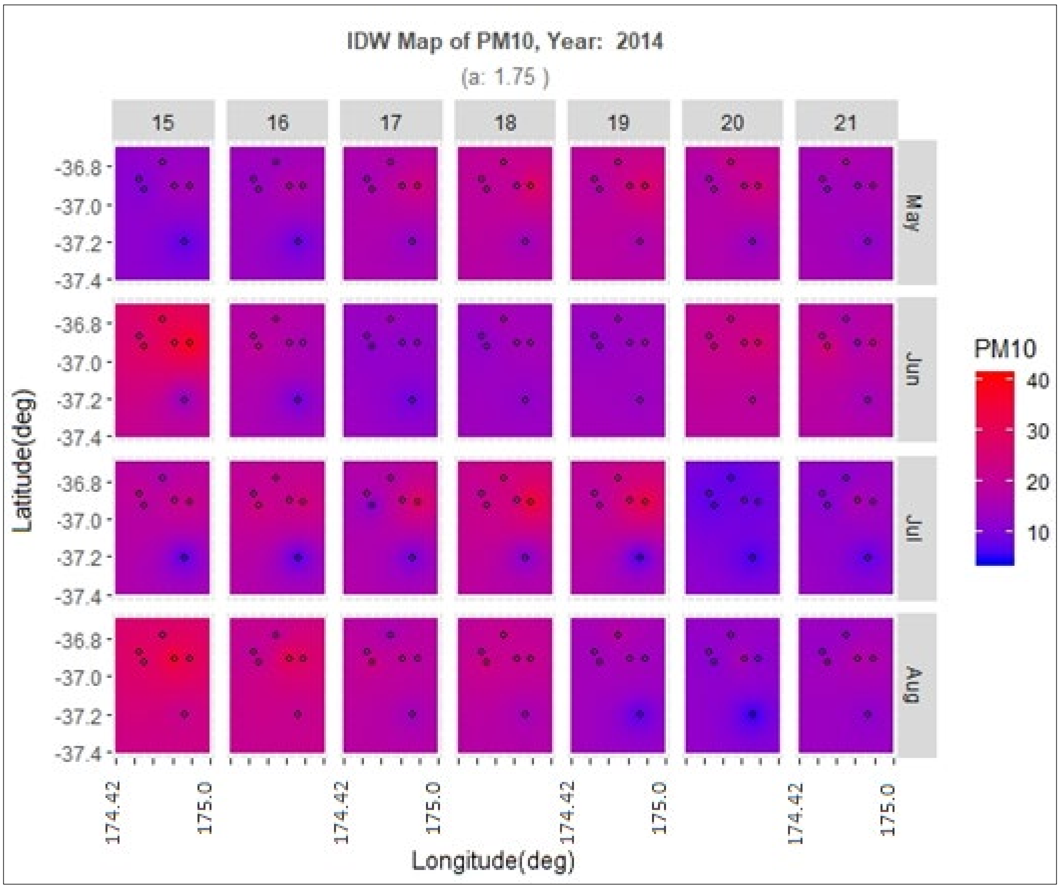


Figure 6.9: IDW prediction map of PM_{10} concentration, May-Aug 2014.

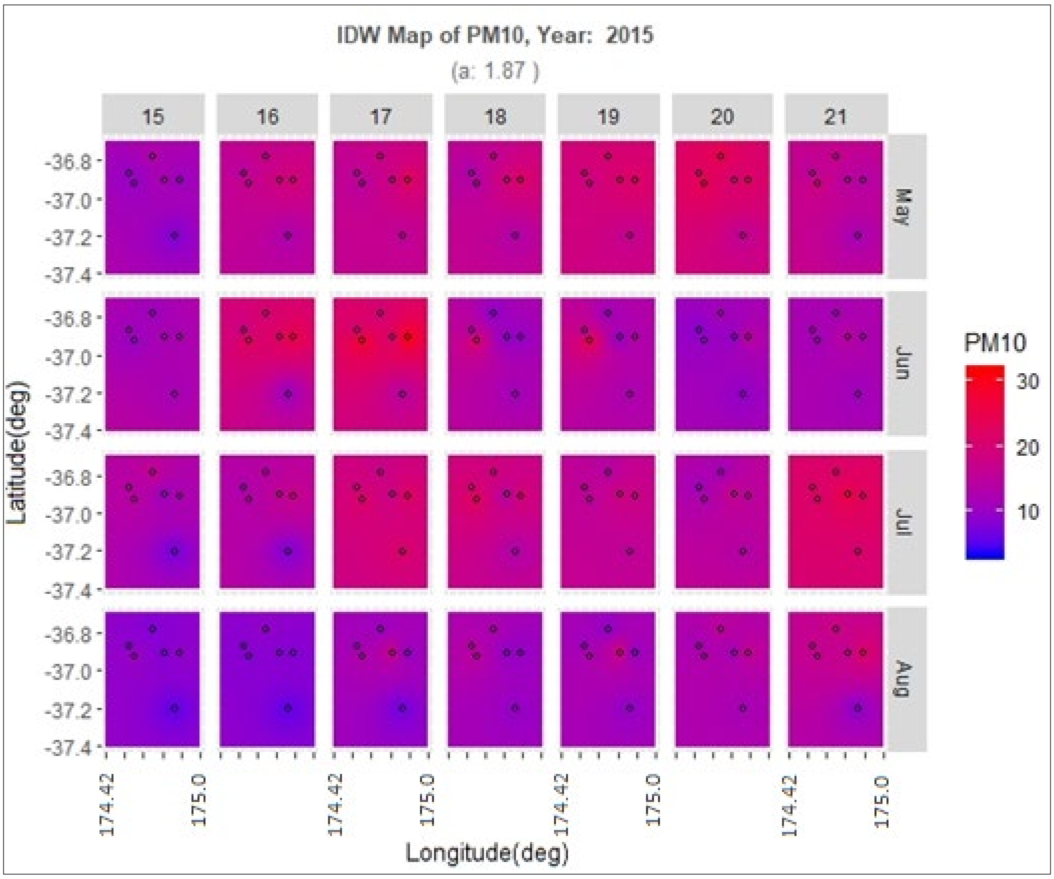


Figure 6.10: IDW prediction map of PM_{10} concentration, May-Aug 2015.

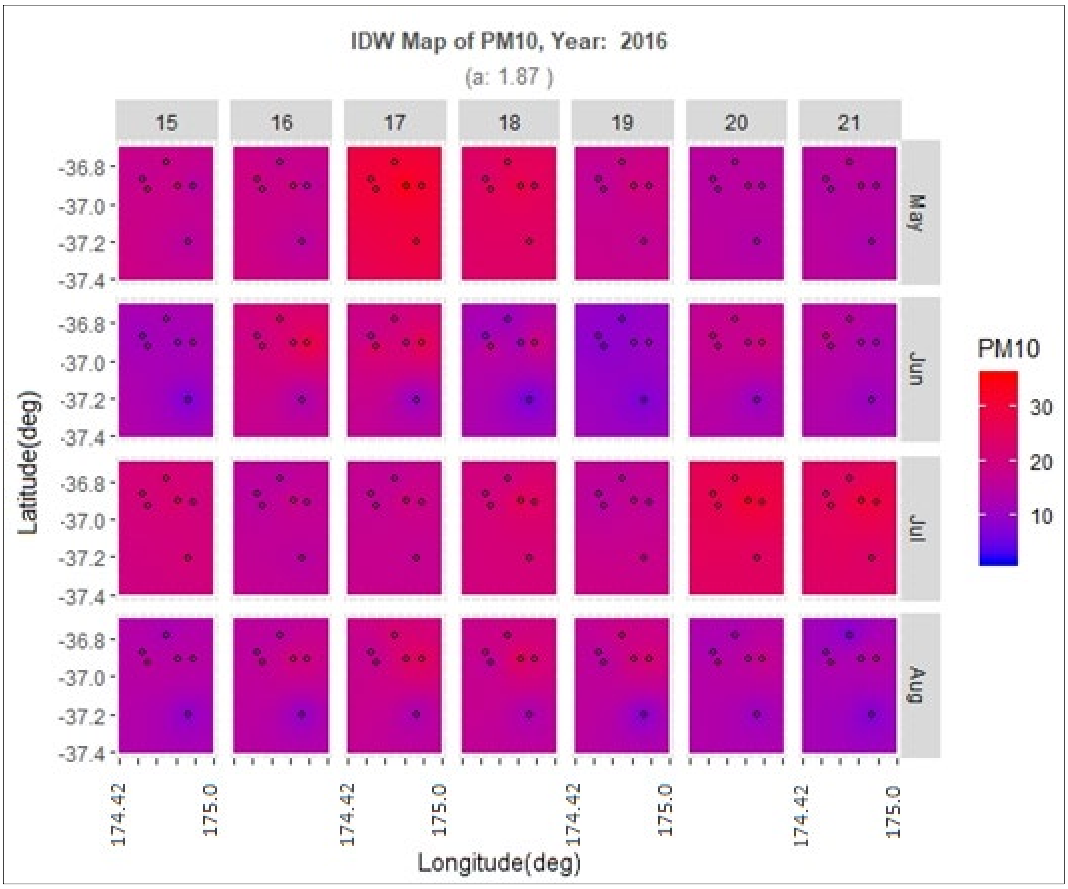


Figure 6.11: IDW prediction map of PM_{10} concentration, May-Aug 2016.

It can be observed from the IDW prediction maps that the predicted PM_{10} concentration on test sets (17 and 19) look smoother (less variation in predicted PM_{10} and hence less variation in predicted map color) than those on days for which we have data. Measurement uncertainty in the data and model-based estimates of the prediction uncertainty are not provided in traditional implementations of deterministic methods (C. Wikle et al., 2019). Therefore, prediction uncertainty of IDW on prediction of Auckland PM_{10} concentration was calculated using the estimates of the overall quality of predictions using cross-validation. The results for IDW performance in terms of MSPE score on the full dataset using LOOCV are presented in Table 6.2. The year specific optimal α-values used were selected the results reported in Table 6.1.

Table 6.2: The leave-one-out cross-validation score for IDW using the optimal α value obtained from Table 6.1.

Year	LOOCV Score
	MSPE (μg/m^{-3})
2011	10.75
2012	12.87
2013	12.44
2014	11.78
2015	9.64
2016	10.01

Model precision across years was varying between 9.64 (μg/m^{-3}) and 12.87 (μg/m^{-3}) though slightly better in 2015 at 9.64 (μg/m^{-3}) on average.

6.4 Regression (Trend-Surface) Estimation

In order to investigate if the regression errors are statistically dependent in space and time a descriptive approach is taken to study the influence of S-T covariates in a regression model on PM$_{10}$ concentration. Spatio-temporal data prediction is obtained using a basic statistical regression model based on the assumption that the trend terms can take into account S-T dependence. Such a regression model can be expressed as (C. K. Wikle et al., 2019a):

$$\mathbf{Z(s_i; t_j) = \beta_0 + \beta_1 X_1(s_i; t_j) + \cdots + \beta_k X_k(s_i; t_j) + e(s_i; t_j)} \qquad \text{Eq. 6.13}$$

where

β_0 is the intercept,

β_k (k > 0) is a regression coefficient associated with $X_k(s_i; t_j)$, the *k*th covariate at spatial location s_i and time t_j, and

errors are assumed to be *iid* such that $e(s_i; t_j) \sim$ *indep*. *N*(0, σ_e^2) for all $\{s_i; t_j\}$

where

data is available, and

N(μ, σ^2) corresponds to a normal distribution with mean μ and variance σ^2.

The covariates $X_k(s_i; t_j)$ can be descriptive features such as elevation that vary spatially but are temporally invariant; temporal trends that vary temporally but are spatially invariant; or other variables such as meteorological variables that are spatially and temporally varying. In addition, S-T basis functions can be used to reconstruct the observed data. In S-T statistics, the basis functions take their values over both the space and time dimensions. As noted previously, investigating the possibility of error dependency in space and/or time between observations is an important aspect of the statistical modeling of S-T data (RESSTE, 2017).

The potential relationship between time and space as a function of spatial and temporal lags can be explored by calculating the S-T covariogram/semivariogram from the residuals (C. K. Wikle et al., 2019b):

$$\hat{\mathbf{e}}(\mathbf{s_i}; \mathbf{t_j}) = \mathbf{Z}(\mathbf{s_i}; \mathbf{t_j}) - \hat{\mathbf{Z}}(\mathbf{s_i}; \mathbf{t_j}) \qquad \text{Eq. 6.14}$$

Apart from the above S-T covariogram/semivariogram diagnostics, statistical tests can be used to measure the spatial or temporal autocorrelation in residuals. The Durbin–Watson (1950) (DW) test is statistical test for measuring the temporal autocorrelation in the residuals. Moran’s Index (I_M) (Moran, 1950) can be applied into datasets in two-dimensional space to measure the spatial dependency between points. The Space-Time Index (STI) measure was proposed by Knox (1964) and Griffith (1981) to assess S-T dependency for areal regions that have a known adjacency structure. The Griffith (1981) formulation joins the I_M with the DW (Henebr, 1994). Alternatively, a S-T analog to the DW test based on the spatial semivariogram can be extended to the S-T setting (C. K. Wikle et al., 2019b):

$$\boldsymbol{F} = \frac{\hat{\boldsymbol{\gamma}}_{\mathbf{e}}(||\mathbf{h_1}||; \mathbf{t_1})}{\hat{\boldsymbol{\delta}}_{\mathbf{e}}^{2}} - \mathbf{1} \qquad \text{Eq. 6.15}$$

where

$\hat{\gamma}_e(||h_1||; t_1)$ is the estimation of empirical semivariogram at the smallest spatial $(||h_1||)$ and temporal t lags, and

$\hat{\delta}_e^2$ is the regression-error variance estimate.

The F statistic is calculated for the randomly permuted data in space and time. An F value that is below the 2.5th percentile or above the 97.5th percentile of these permutation samples is evaluated as “large” and the null hypothesis of S-T independence is rejected (at the 5% level of significance). A ‘large’ F therefore suggests that the data are dependent.

6.4.1 Regression Trend-Surface Estimation Experiments and Results

A linear model was fitted to the training set using longitude, latitude, day, and all the interactions between them, as well as the basis functions without interactions and meteorological data as covariates. The formula can be written as:

$$\mathbf{z} \sim (\mathbf{lon} + \mathbf{lat} + \mathbf{day})\hat{}\mathbf{2} \ + \cdot \qquad \text{Eq. 6.16}$$

where z is the PM_{10} value and the convenient notation "·" denotes the covariates.

Spatial trends were considered by allowing the covariates to correspond to the S-T coordinate, and/or their transformations and interactions in addition to meteorological variables.

Parameter Inference of Model: The ordinary least squares (OLS**)** parameter estimates and related standard errors from the OLS fit of the regression model are presented in Table 6.3. These estimates are based on the assumption that the errors are independent. The resultant standard errors suggest that the effect of choice of basis function was not significant in the model given all the other covariates involved in the model. It is also notable that the effect of latitude is not considered significant aligning with the findings of no latitudinal trend in the empirical spatial mean of PM_{10} (see Figure 6.1 in section 6.2.1). The interaction of the latitude by day is also not captured by the linear model.

The interaction between latitude and longitude was ignored in the model and this could possibly be confounding the results. Confounding factors can affect the inference and mean that the interpretation or significance of a model is substantially different if a significant variable is disregarded, or possibly an unnecessary variable is incorporated in the model. To examine the possibility of confounding the basis functions were removed from the model. The OLS parameter estimates, and related standard errors provided in Table 6.4 show the significance of latitude and longitude interaction after eliminating the basis functions from the model, indicating that latitude and longitude are indeed confounding factors for the model.

Spatial analysis of residuals: Having fitted the S-T LM model on both training sets, their related residuals were checked for S-Tl correlation. Spatio-temporally correlated residuals are indicators that the spatial and temporal variability in the data was not effectively captured by the model. The residuals plots showed that the residuals were strongly spatially correlated however the degree of spatial correlation changes daily, monthly, and annually.

The residual plots of fitted S-T LM on both block and random training sets for May 2015 are presented in Figure 6.12 and Figure 6.13 respectively showing strongly spatially correlated residuals.

Table 6.3: Estimated regression coefficients and the standard errors (SE) using OLS including basis function on full dataset and train sets (block and random).

	Full Dataset			Train set					
				Block			Random		
	$\hat{\beta}_{ols}$	$SE(\hat{\beta}_{ols})$		$\hat{\beta}_{ols}$	$SE(\hat{\beta}_{ols})$		$\hat{\beta}_{ols}$	$SE(\hat{\beta}_{ols})$	
(Intercept)	-3.85e+02	1.57e+02	*	-5.27e+02	1.94e+02	**	-4.08e+02	1.82e+02	*
Lon	2.27e+00	9.83e-01	*	3.27e+00	1.22e+00	**	2.22e+00	1.14e+00	.
Lat	-5.13e-03	1.30e+00		9.21e-01	1.62e+00		-8.32e-01	1.50e+00	
Day	-3.73e-01	1.14e-01	**	-2.70e-01	1.42e-01	.	-3.78e-01	1.30e-01	**
Lag_1	5.22e-01	8.60e-03	***	5.35e-01	1.07e-02	***	5.25e-01	1.01e-02	***
Lag_2	-1.97e-02	8.57e-03	*	-3.42e-02	1.06e-02	**	-1.59e-02	9.90e-03	
Temp	-1.37e-01	1.19e-02	***	-1.31e-01	1.47e-02	***	-1.37e-01	1.37e-02	***
Rain	-6.50e-02	1.25e-02	***	-7.01e-02	1.49e-02	***	-6.82e-02	1.42e-02	***
RH	-4.49e-02	5.47e-03	***	-3.50e-02	6.71e-03	***	-4.41e-02	6.29e-03	***
SR	2.08e-03	5.56e-04	***	1.95e-03	6.87e-04	**	1.47e-03	6.38e-04	*
WD	2.48e-03	4.55e-04	***	2.90e-03	5.65e-04	***	2.57e-03	5.21e-04	***
WS	4.33e-01	3.79e-02	***	4.03e-01	4.70e-02	***	5.14e-01	4.28e-02	***
B_1	1.75e+01	2.70e+01		1.31e+01	3.33e+01		3.65e+01	3.10e+01	
B_2	-3.07e+01	4.61e+01		-2.41e+01	5.67e+01		-8.33e+01	5.28e+01	
B_3	1.46e+01	2.43e+01		1.12e+01	2.99e+01		4.24e+01	2.79e+01	
Lon: Lat	NA	NA		NA	NA		NA	NA	
Lon: Day	2.13e-03	6.94e-04	**	1.42e-03	8.63e-04	.	2.15e-03	7.89e-04	**
Lat: Day	-1.48e-05	5.20e-04		-5.85e-04	6.41e-04		-3.60e-05	5.90e-04	
Adjusted R-squared	0.34			0.35			0.34		

Observations: Signif. (*p*-value) codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Table 6.4: Estimated regression coefficients and the standard errors (within parentheses) using OLS without basis function.

	Full Dataset			**Train set**					
				Block			Random		
	$\hat{\beta}_{ols}$	$SE(\hat{\beta}_{ols})$		$\hat{\beta}_{ols}$	$SE(\hat{\beta}_{ols})$		$\hat{\beta}_{ols}$	$SE(\hat{\beta}_{ols})$	
(Intercept)	-2.22e+05	-2.06e+05	***	-2.23e+05	5.61e+04	***	-.2000e+005	5.5120e+004	***
Lon	1.27e+03	1.18e+03	***	1.28e+03	3.21e+02	***	1.2590e+003	3.1530e+002	***
Lat	-6.02e+03	-5.56e+03	***	-6.04e+03	1.52e+03	***	-.9520e+003	1.4940e+003	***
Day	-3.76e-01	-2.71e-01	***	-3.82e-01	1.33e-01	**	-3.8220e-001	1.2950e-001	**
Lag_1	5.21e-01	5.35e-01	***	5.46e-01	1.01e-02	***	5.1570e-001	8.5270e-003	***
Lag_2	-1.97e-02	8.565e-03	*	-3.03e-02	1.01e-02	**	-1.594e-02	9.895e-03	
emp	-1.35e-01	-1.30e-01	***	-1.19e-01	1.40e-02	***	-1.3270e-001	1.3670e-002	***
Rain	-6.47e-02	-6.99e-02	***	-6.42e-02	1.49e-02	***	-6.7320e-002	1.4240e-002	***
RH	-4.61e-02	-3.61e-02	***	-4.61e-02	6.33e-03	***	-4.6200e-002	6.2330e-003	***
SR	2.00e-03	1.89e-03	***	2.20e-03	6.48e-04	***	-2.013e-003	6.453e-04	**
WD	2.53e-03	2.95e-03	***	3.41e-03	5.30e-04	***	2.6390e-003	5.1990e-004	***
WS	4.09e-01	3.82e-01	***	4.37e-01	4.03e-02	***	4.7170e-001	3.8970e-002	***
Lon: Lat	3.44e+01	3.82e-01	***	3.45e+01	8.70e+00	***	3.4060e+001	8.5430e+000	***
Lon: Day	2.15e-03	1.43e-03	**	2.18e-03	8.10e-04	**	1.855e-03	7.904e-004	*
Lat: Day	-1.73e-05	-5.96e-04		-3.81e-05	6.08e-04		-5.340e-04	5.915e-04	
Adjusted R-squared	0.34			0.35			0.36		

Observations: Signif. (*p*-value) codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

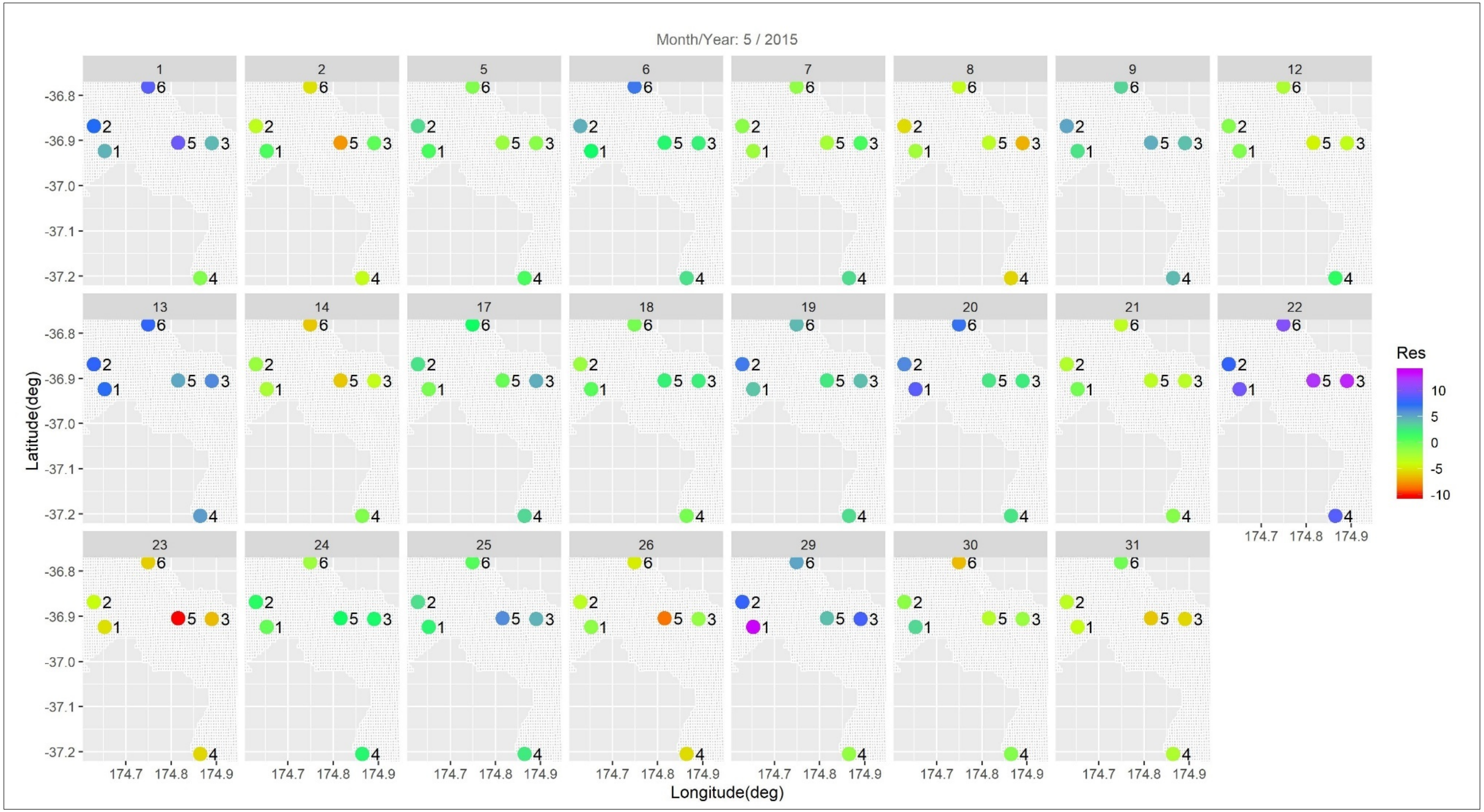


Figure 6.12: Spatial residuals from fitting the regression (trend) model to the train block set for May 2015.

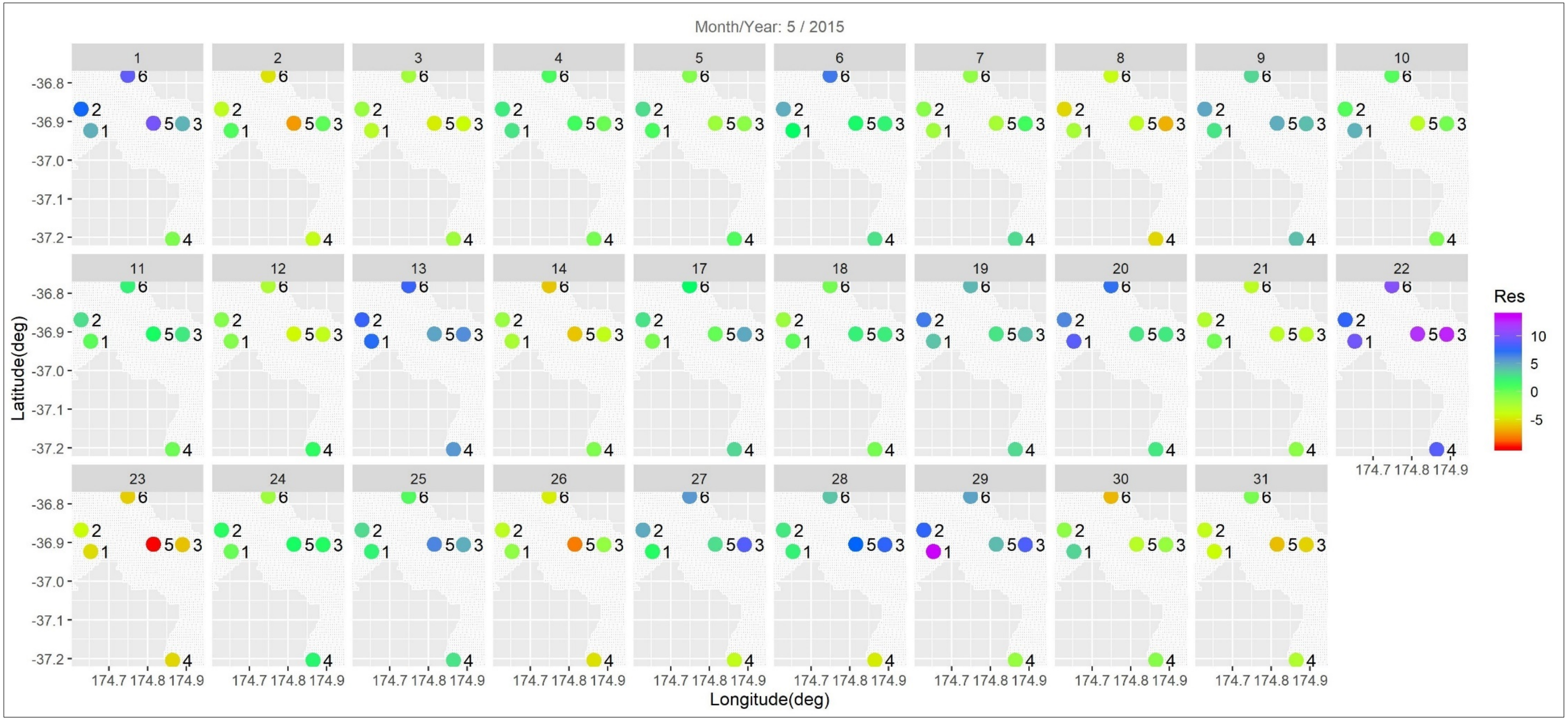


Figure 6.13: Spatial residuals from fitting the regression (trend) model to the train random set for May 2015.

In addition to visual inspection of plots, the Moran's I statistic was calculated to test for spatial dependence in the residuals on individual days. Details of the test are provided in Appendix C. The distances for each day were calculated to form a weight matrix. Moran's I test using the function Moran.I from the R package "ape" was used (Paradis & Schliep, 2019). The results of the Moran's I (I_M) test statistics, for all years, are low and therefore the H_0 of no spatial correlation at 5% can be rejected at each time point. As an example, the calculated *p*-value for May 2015 on block and random train sets were 1.44e-07 and 1.29e-06 respectively indicating spatial dependency between residuals as suggested by the plots (Figure 6.12 and Figure 6.13). These results imply that the S-T LM failed to capture the spatial variability in the data at each time point.

The performance of S-T LMs with respect to capturing the temporal correlation between the residuals were investigated using the DW test. The DW test for the residuals at every station was performed by grouping data using spatial coordinates and creating a nested data frame containing one row for each group. A function was defined to take the data frame associated with a single group and perform the DW test.

The autocorrelation in the residuals was then tested using the "dwtest" function available in R's "lmtest" package (Zeileis & Hothorn, 2002) after removing a temporal trend. The low Bonferroni correction value (0.06% and 0.03% for block and random sets); as proportion of *p*-values below the 5% significance level divided by the number of tests; provided evidence that there is not temporal autocorrelation in the residuals of the train sets. Table 6.5 presents, as an example, the results of the DW test on the residuals of both the block and random training sets for May 2015. The results imply that the DW H_0 of no temporal autocorrelation between the residuals should be accepted for May 2015.

Table 6.5: DW test on residuals of fitted S-T LM model on block and random train sets, May 2015.

	Bock set		**Random set**	
	DW statistics	*p*-value	DW statistics	*p*-value
Glen Eden	2.39	0.83	2.56	0.91
Henderson	2.85	0.98	2.91	0.99
Pakuranga	2.93	0.99	2.98	0.99
Patumahoe	2.60	0.93	2.50	0.88
Patumahoe	2.92	0.99	3.16	0.99
Takapuna	2.88	0.98	2.87	0.98
Bonferroni Value	0		0	
Alternative hypothesis (H_a) = "*true autocorrelation is greater than 0*"				

These temporal residuals were also plotted for visual inspection. The plots did not show a temporal correlation in the residuals. The residuals however were correlated between the stations meaning at the same time point the nearby stations' residuals were more similar than at other time points. This agrees with results of the tests for spatial correlation in the residuals. The PM_{10} residuals for May 2015 at all six stations are illustrated in Figure 6.14.

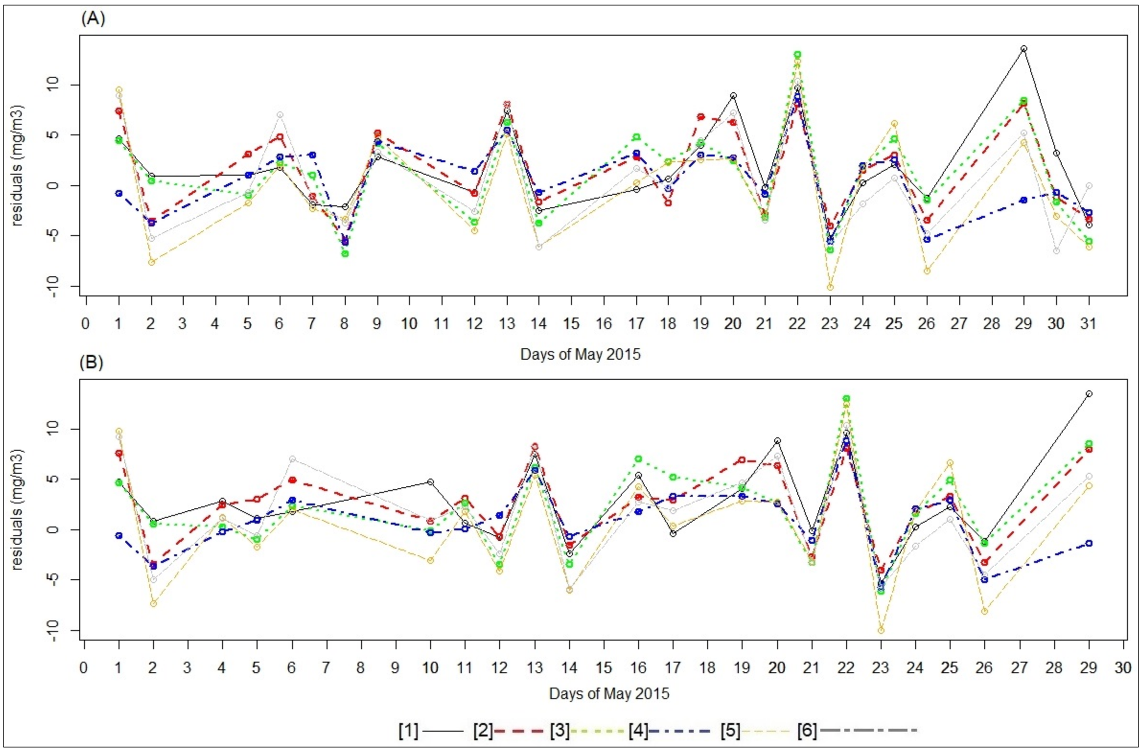


Figure 6.14: Temporal residuals from fitted T-S LM on (A) block and (B) random train sets, May 2015. In the legend, [1]– [6] correspond to Glen Eden; Henderson; Pakuranga; Patumahoe; Penrose and Takapuna stations, respectively.

Extending the I_M test to the S-T setting was a challenge in terms of appropriately scaling time to form a Euclidean distance through space and time with a sensible interpretation. One approach to overcome this challenge is to fit a dependence model that allows for scaling in time, and then to scale time by an estimate of the scaling factor prior to computing the Euclidean distance (C. K. Wikle et al., 2019b). This approach is reported in the next Chapter where a kriging model is developed that uses an anisotropic covariance function. In the method initially taken here however, time is not scaled and distances on the S-T domain are calculated. In this approach, the weights from the distances were computed and the diagonal was set to zero prior to calling the Moran.I function in R. The *p*-value calculated for S-T I_M was very small (max *p*-value = 0), strongly suggesting that there is a S-T dependency between residuals in the residual data for both training sets.

Further S-T analysis of residuals was performed by computing and visualizing the empirical S-T semivariogram of the residuals. To calculate the empirical S-T semivariogram spatial bins were set to 20 km and the cutoff point was set to consider points that are 50 km apart. Figure 6.15 shows the empirical S-T semivariogram of the original data set (A) and the residuals after fitting on the block training set (B) and the random training set (C) for May 2015. Comparison of the semivariogram plots in Figure 6.15 (A), (B) and (C) shows the original data set has a lower sill. This means that the time and spatial covariates, along with the rest of the covariates, explained little of the S-T variability in the data set. The highest variance for (A) is 30 (light yellow, top right hand of the figure), for (B) and (C) the highest variance value is just over 20.

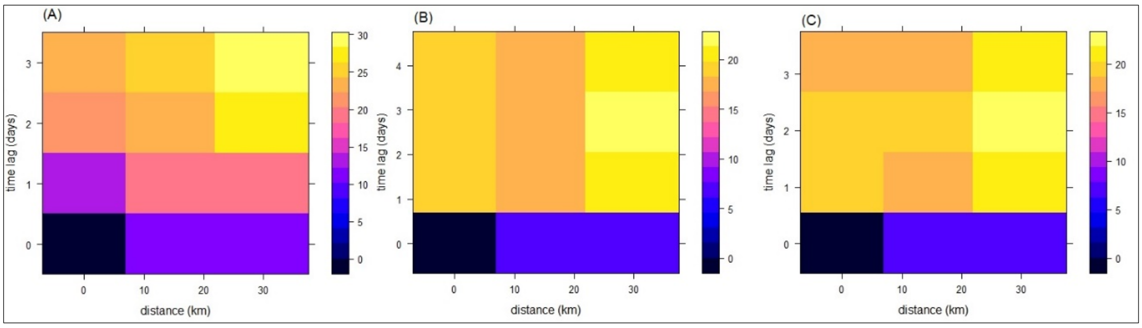


Figure 6.15: Empirical S-T semivariogram of daily PM_{10} (A) and Empirical S-T semivariogram of residuals after fitting the linear model on train sets: (B) block and (C) Random, May 2015.

Prediction from the constructed S-T LM was performed at the prediction locations where covariate values were available. For each test set the computed *p*-value from the spatio- temporal I_M test was low therefore the I_M test H_0 was rejected in favor of the H_a that the residuals values appear to be correlated with their S-T location. Prediction errors of the fitted S-T LM model on block and random test sets during May 2015 are presented in Figure 6.16 and Figure 6.17 respectively. The spatial correlation in the residuals is evident between the nearby stations; that is, at the same day, the residuals are more similar than at different time points.

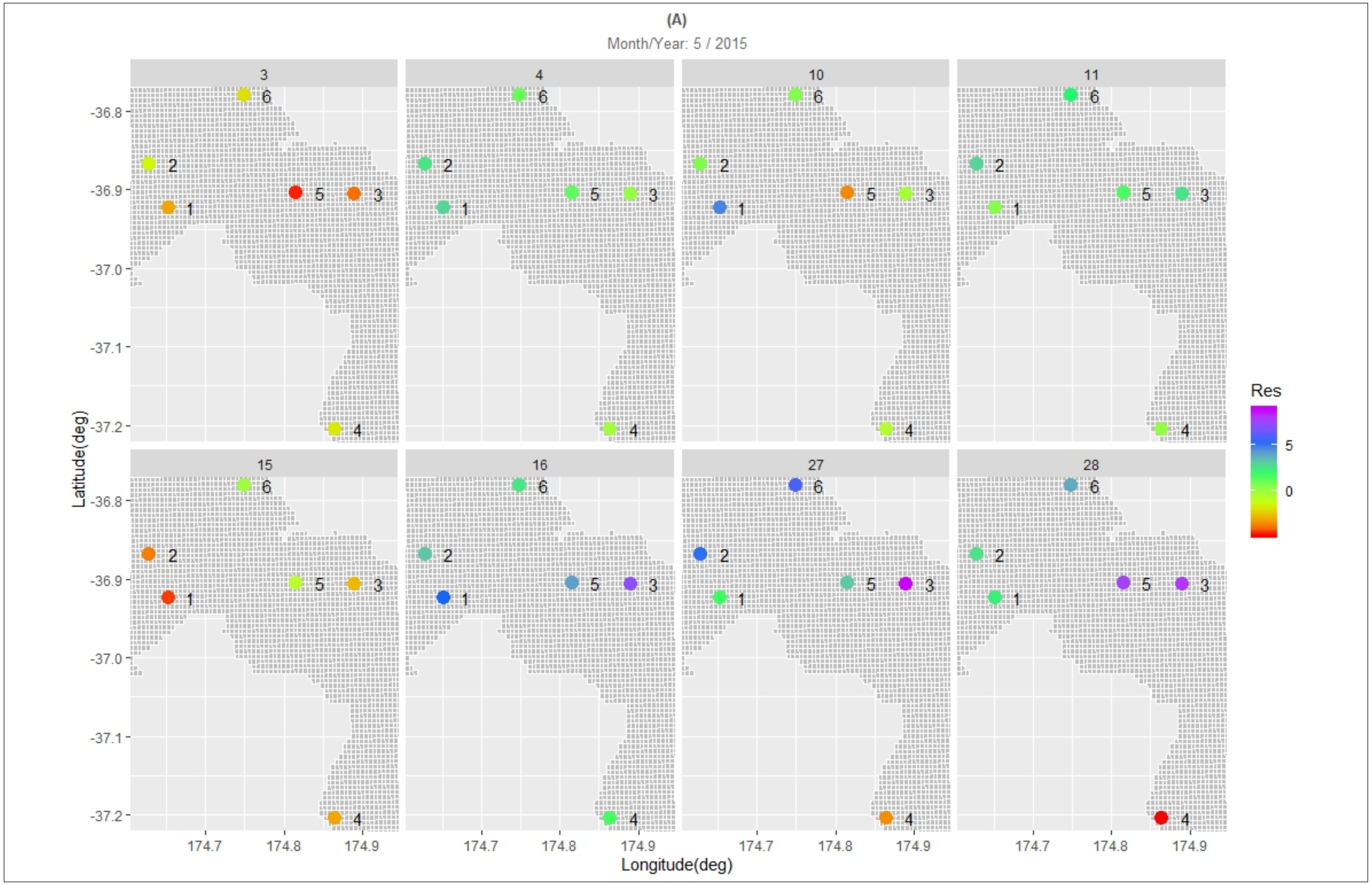


Figure 6.16: Residuals of predicted PM_{10} from fitted S-T LM on block test set.

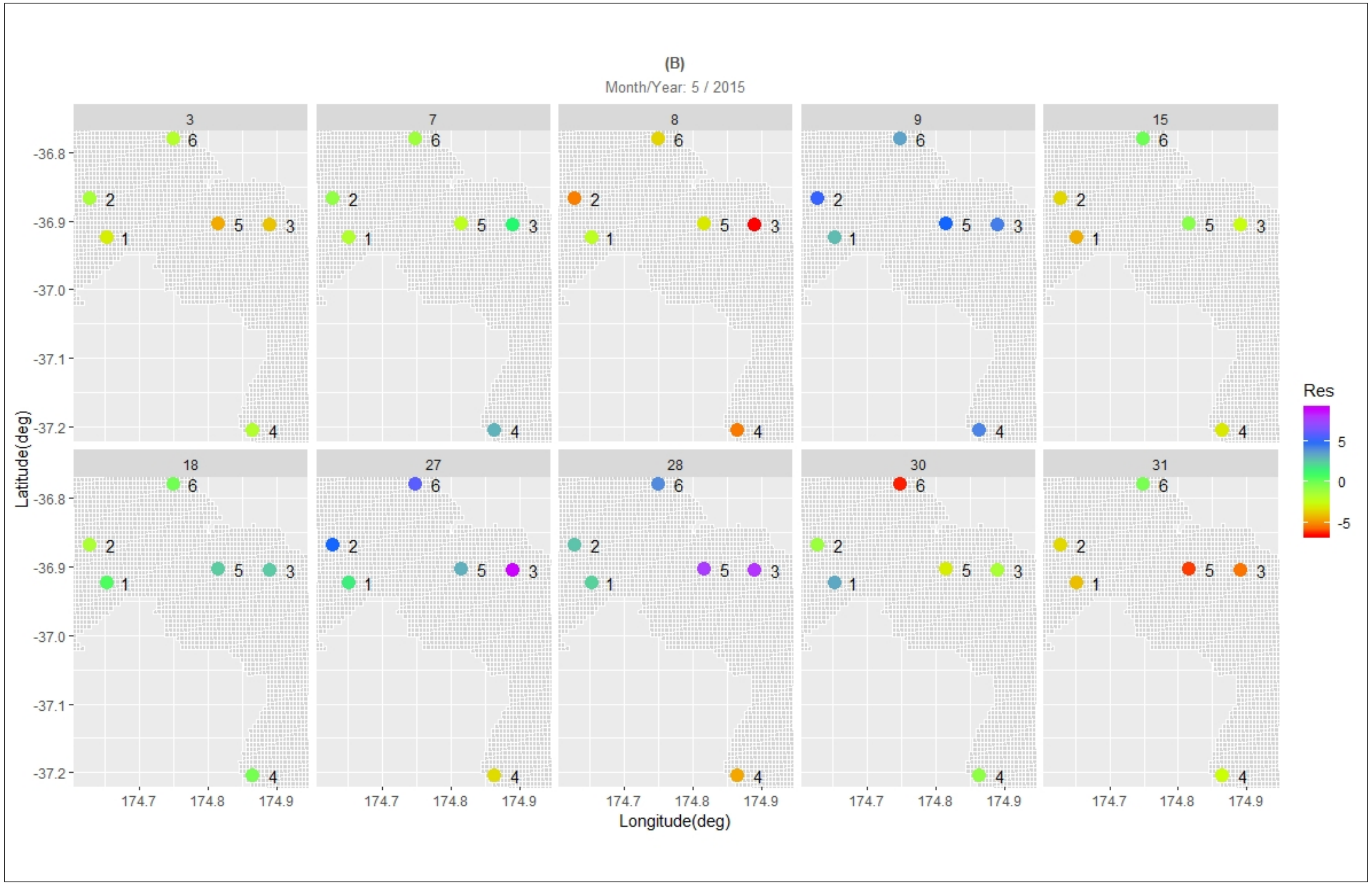


Figure 6.17: Residuals of predicted PM_{10} from fitted S-T LM on random test set.

Temporal correlation between the residuals was analyzed visually and by calculating the p-value of DW test. Temporal plot of residuals and the DW statistics by station for May 2015 are provided in Figure 6.18 and Table 6.6, respectively. The results showed no temporal autocorrelation between the residuals of fitted S-T LM tests on both sets as the Bonferroni correction value was close to 0 for both sets. It should be noted that the calculated p-value on block set was high except in the cases of the Pakuranga and Takapuna stations therefore the H_0 of no temporal autocorrelation between the residuals at 5% for these two stations is rejected. However, the low value of Bonferroni correction for both sets, suggests that the temporal autocorrelation in the residuals is not significant.

Table 6.6: DW test on residuals of fitted S-T LM model on block and random test sets, May 2015.

	Bock set		**Random set**	
	statistics	p-value	statistics	p-value
Glen Eden	2.27	0.66	2.07	0.55
Henderson	2.02	0.62	2.40	0.75
Pakuranga	0.89	**0.03**	1.62	0.26
Patumahoe	1.98	0.49	3.18	0.98
Penrose	1.18	0.11	1.42	0.16
Takapuna	0.95	**0.04**	2.09	0.56
Bonferroni Value	0.67		0	
H_A = *"true autocorrelation is greater than 0"*				

The plots show no temporal correlation between the residuals (except for Pakuranga and Takapuna stations on block sets) but there is ample spatial correlation suggested by the I_M test and the spatial residual plots.

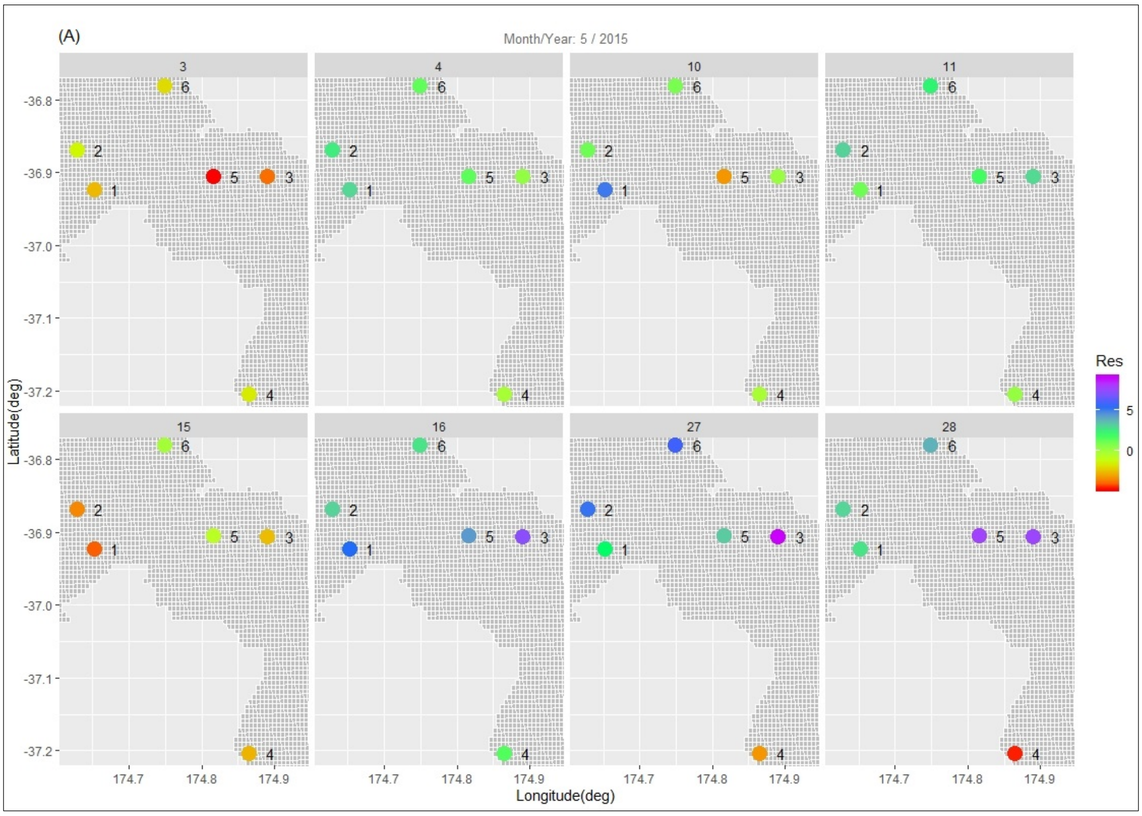


Figure 6.18: Temporal residuals from fitted T-S LM on (A) block and (B) random test sets, May 2015 where [1]:[6] corresponds to Glen Eden; Henderson; Pakuranga; Patumahoe; Penrose and Takapuna stations, respectively.

The p-value from the S-T I_M on residuals of predicted PM_{10} on both block and random test sets was low (1.44e-07 and 5.04e-11) therefore the H_0 of no spatial correlation between the residuals was rejected.

These results suggest that simple geographical and temporal trend terms and a linear model of these covariates is ***not*** capable of explaining all the observed S-T variability of such a complex environmental process. Thus, fitting the S-T LM model resulted in residuals that were spatio temporally correlated. This S-T dependency in residuals may indicate that a more refined S-T random-effects model should be considered for these data. Therefore, it is reasonable to conclude that the dependent errors in these S-T models represent the effects of confounding between the S-T dependent random errors and the covariates.

6.5 Generalized Linear Model (GLM)

A basic Generalized Linear Model (GLM) consists of a random and a systematic component. The random component assumes that observations are independent, and their distributions are from exponential family. The relationship between the mean response and the covariates is specified by the systematic component. This is achieved by transforming the mean response using a link function and then expressing it in terms of a linear function of the covariates (McCulloch et al., 2008). It is assumed that conditional on the presence of covariates (or functions of these in the case of GAMs), any location in the space-time domain S-T GLMs or S-T GAMs can be used for S-T prediction. Performance of the S-T GLMs and S-T GAMs in terms of sufficiently accommodating the dependence in the observations are subject to the data set and the available covariates. The next section reports on experiments in which S-T GLMs and GAMs are explored for the Auckland PM_{10} data.

6.5.1 S-T GLM Experiments and Results

A S-T GLM was fitted to daily PM_{10} data. The same class of covariates were used in the S-T GLM model as were employed in the S-T LM model presented in Section 6.3.3. To fit the S-T GLM function, both Gamma and Gaussian family models as well as the log link function were used. The results of these initial experiments resulted in the Gamma family ultimately being selected over Gaussian family as it had slightly lower AIC. The results of the S-T GLM model fit on the full dataset as well as both training data sets are provided in Table 6.7.

Similar conclusion to that formed for S-T LM can be drawn for S-T GLM including the fact that the basis function's covariates could not explain the interactions of latitude and longitude with PM_{10}. This lack of explanation suggests that the basis function covariates should be discarded from the model. Therefore, the models were rebuilt without the basis functions as covariates. The results of these experiments are provided in Table 6.8.

Spatial Analysis of Residuals on Test Sets: Prediction from the constructed S-T GLM was performed on both test sets (random and block). Prediction plots for the block and random sets are provided in Appendices G and H respectively. Figure 6.19 (page 205) and Figure 6.20 (page 206) shows the PM_{10} prediction error for May 2015 on (A) block and (B) random test sets.

Table 6.7: Estimated regression coefficients and the standard errors for the S-T GLM *with* the basis functions as covariates.

	Full Dataset			**Train set**					
				Block			Random		
	$\hat{\beta}_{GLM}$	$SE(\hat{\beta}_{GLM})$		$\hat{\beta}_{GLM}$	$SE(\hat{\beta}_{GLM})$		$\hat{\beta}_{GLM}$	$SE(\hat{\beta}_{GLM})$	
(Intercept)	1.12e+10	1.36e+10		4.81e+10	6.16e+10		-1.67e+11	1.83e+11	
Lon	-8.61e+07	1.05e+08		-2.29e+08	3.12e+08		1.20e+09	1.31e+09	
Lat	-1.06e+08	1.29e+08		2.21e+08	7.48e+08		1.15e+09	1.26e+09	
Day	-3.06e-02	8.22e-03	***	-2.99e-02	9.54e-03	**	-2.99e-02	9.46e-03	**
Lag$_1$	3.50e-02	6.20e-04	***	3.69e-02	7.23e-04	***	3.60e-02	7.26e-04	***
Lag$_2$	-2.09e-04	6.18e-04		-9.55e-04	7.25e-04		1.43e-04	7.21e-04	
Temp	-9.30e-03	8.58e-04	***	-8.31e-03	1.01e-03	***	-9.27e-03	1.00e-03	***
Rain	-6.53e-03	9.03e-04	***	-6.57e-03	1.07e-03	***	-6.00e-03	1.02e-03	***
RH	-3.59e-03	3.94e-04	***	-3.46e-03	4.58e-04	***	-3.11e-03	4.69e-04	***
SR	1.38e-04	4.01e-05	***	1.47e-04	4.66e-05	**	1.44e-04	4.72e-05	**
WD	2.09e-04	3.28e-05	***	2.74e-04	3.82e-05	***	2.09e-04	3.85e-05	***
WS	3.16e-02	2.73e-03	***	3.38e-02	3.18e-03	***	3.18e-02	3.20e-03	***
B$_1$	1.09e+10	1.32e+10		-1.18e+09	2.36e+10		-3.47e+10	3.79e+10	
B$_2$	-1.20e+09	1.46e+09		1.97e+10	1.47e+11		1.74e+10	1.90e+10	
B$_3$	1.91e+08	2.32e+08		-1.04e+10	7.78e+10		-7.73e+09	8.44e+09	
B$_5$	-7.65e+08	9.32e+08		-9.33e+08	7.31e+09		NA	NA	
B$_9$	NA	NA		1.85e+08	6.11e+08		7.59e+08	8.29e+08	
Lon: Lat	NA	NA		NA	NA		NA	NA	
Lon: Day	1.76e-04	5.01e-05	***	1.70e-04	5.82e-05	**	1.64e-04	5.76e-05	**
Lat: Day	1.80e-06	3.75e-05		-5.11e-06	4.36e-05		-3.50e-05	4.31e-05	
AIC	74367			54931			53305		

Observations: Signif. (*p*-value) codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Table 6.8: Estimated regression coefficients and standard errors for the S-T GLM *without* the basis functions as covariates.

	Full Dataset			**Train set**					
				Block			Random		
	$\hat{\beta}_{GLM}$	$SE(\hat{\beta}_{GLM})$		$\hat{\beta}_{GLM}$	$SE(\hat{\beta}_{GLM})$		$\hat{\beta}_{GLM}$	$SE(\hat{\beta}_{GLM})$	
(Intercept)	-1.55e+04	3.46e+03	***	-1.54e+04	4.02e+03	***	-1.50e+04	4.04e+03	***
Lon	8.89e+01	1.98e+01	***	8.80e+01	2.30e+01	***	8.59e+01	2.31e+01	***
Lat	-4.20e+02	9.38e+01	***	-4.16e+02	1.09e+02	***	-4.06e+02	1.09e+02	***
Day	-3.10e-02	8.20e-03	***	-3.04e-02	9.52e-03	**	-3.04e-02	9.43e-03	**
Lag$_1$	3.50e-02	6.18e-04	***	3.68e-02	7.21e-04	***	3.60e-02	7.23e-04	***
Lag$_2$	-1.97e-04	6.16e-04		-9.47e-04	7.23e-04		1.66e-04	7.19e-04	
Temp	-9.15e-03	8.52e-04	***	-8.10e-03	1.00e-03	***	-9.09e-03	9.98e-04	***
Rain	-6.54e-03	9.00e-04	***	-6.56e-03	1.07e-03	***	-6.00e-03	1.02e-03	***
RH	-3.70e-03	3.90e-04	***	-3.59e-03	4.53e-04	***	-3.27e-03	4.63e-04	***
SR	1.31e-04	3.99e-05	**	1.39e-04	4.64e-05	**	1.34e-04	4.69e-05	**
WD	2.15e-04	3.26e-05	***	2.79e-04	3.80e-05	***	2.16e-04	3.83e-05	***
WS	2.99e-02	2.48e-03	***	3.16e-02	2.88e-03	***	2.96e-02	2.90e-03	***
Lon: Lat	2.41e+00	5.37e-01	***	2.38e+00	6.23e-01	***	2.33e+00	6.26e-01	***
Lon: Day	1.78e-04	5.00e-05	***	1.73e-04	5.80e-05	**	1.67e-04	5.74e-05	**
Lat: Day	2.73e-06	3.74e-05		-3.74e-06	4.35e-05		-3.37e-05	4.30e-05	
AIC	74365			54926			53303		

Observations: Signif. (*p*-value) codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

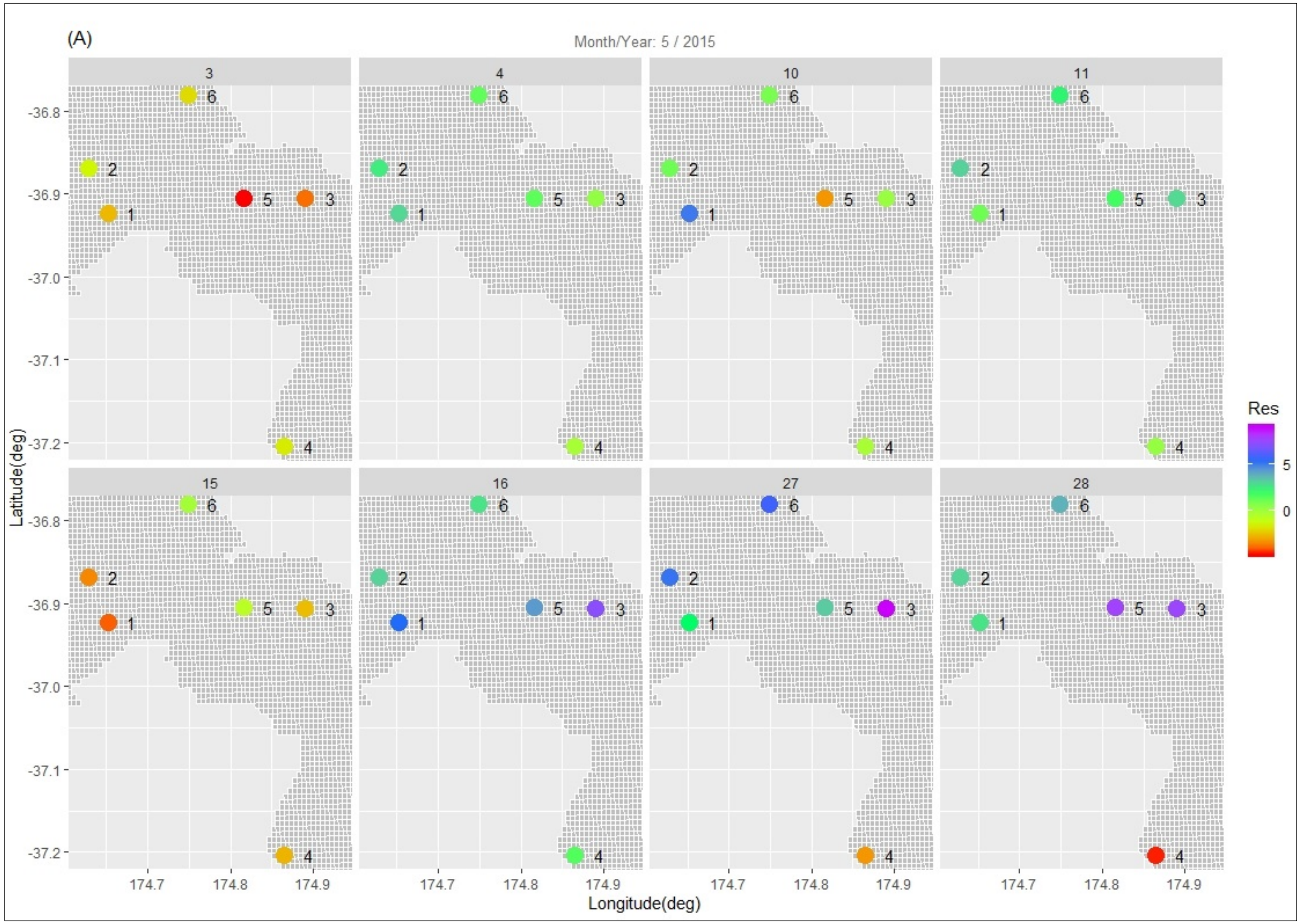


Figure 6.19: Residuals of predicted PM_{10} from fitted S-T GLM on block test set.

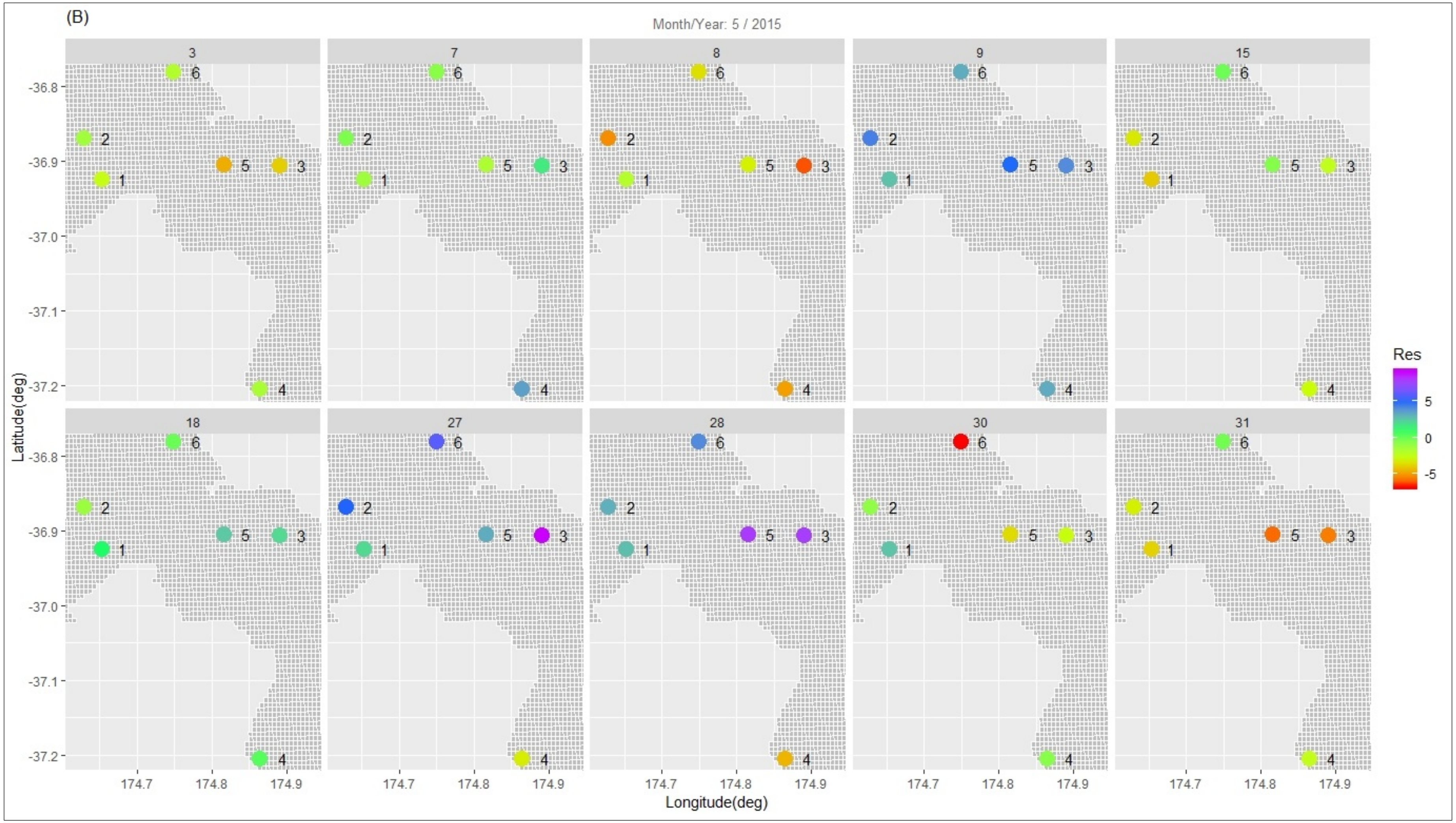


Figure 6.20: Residuals of predicted PM_{10} from fitted S-T GLM on random test set.

Spatial correlation of the residuals of the fitted S-T GLMs on both test sets was tested using the I_M test on the spatial residuals. The calculated *p*-value for I_M test was 9.57e-08 and 1.26e-11 for block and random test sets respectively. The results (Table 6.9) show that at the 5% level of significance, the *p*-value is statistically significant. Therefore, the null hypothesis of no spatial correlation between residuals is rejected. This finding is in agreement with the visualizations presented in Figure 6.19 and Figure 6.20.

To investigate the temporal dependency between the residuals visually and statistically S-T plots of residuals and DW were created and analysed in the same manner as for the S-T LM model. The results showed ample spatial and low temporal dependence between the residuals in both sets. Temporal plots of the residuals for May 2015 are provided in Figure 6.21 as an example. Since the difference of the residuals of S-T LM and S-T GLM were close, their temporal plot should look very similar. Therefore, the residual plots of both S-T LM and S-T GLM on block and random sets were produced to check this assumption. In the plots the thicker lines indicate the S-T GLM residuals.

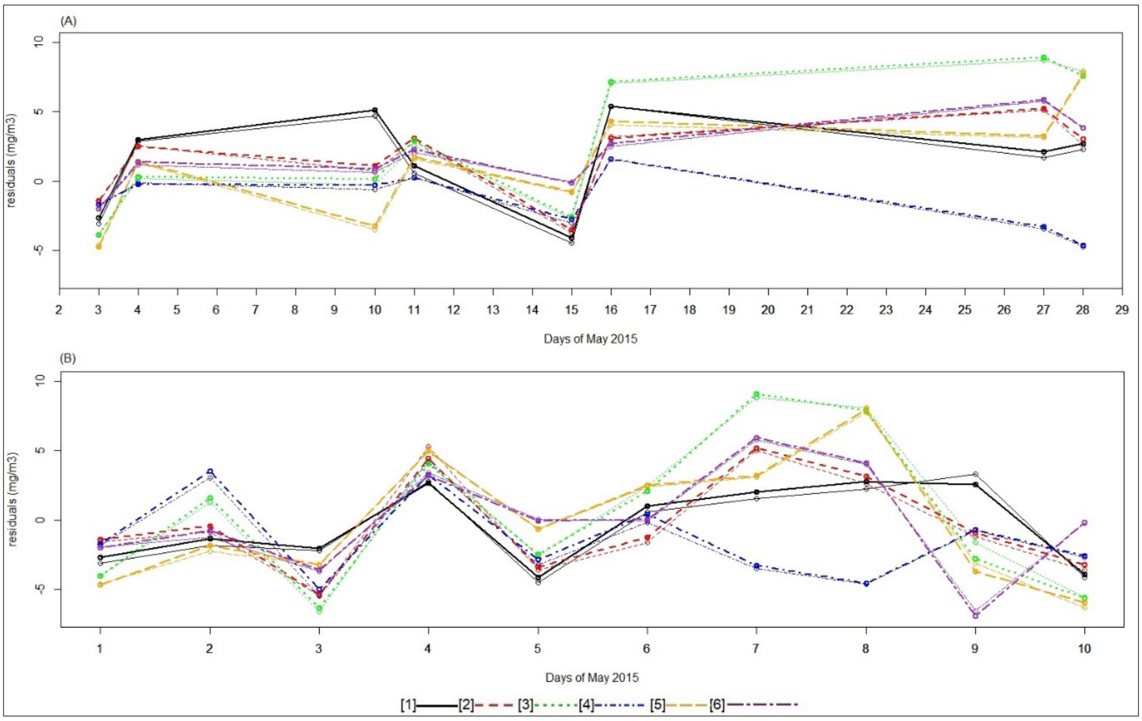


Figure 6.21: Temporal residuals of fitted S-T LM and S-T GLM on (A) block and (B) test sets. Thicker lines show the residuals of S-T GLM. In the legend, [1]– [6] correspond to Glen Eden; Henderson; Pakuranga; Patumahoe; Penrose and Takapuna stations, respectively.

The plots show no temporal correlation between the residuals (except for Pakuranga and Takapuna stations on block sets) but ample spatial correlation as suggested by the I_M test and the spatial residual plots (Figure 6.16 (page 198), Figure 6.17 (page 199), Figure 6.19 (page 205), and Figure 6.20 (page 206)).

The DW test was carried out on the residuals from the fitted S-T GLM on both test sets. The null hypothesis (H_0) of no temporal dependency between the residuals was accepted at 5%. The low value of Bonferroni correction for both sets suggests that the temporal autocorrelation in the residuals is not significant. The plot shows no temporal correlation between the residuals (except for Pakuranga and Takapuna stations on block sets) but ample spatial correlation as suggested by the I_M test and the spatial residual plots.

Table 6.9: DW test on residuals of fitted S-T GLM model on block and random test sets, May 2015.

	Bock set		***Random set***	
	statistics	*p*-value	statistics	*p*-value
Glen Eden	2.23	0.64	1.97	0.48
Henderson	2.12	0.57	2.95	0.69
Pakuranga	0.94	0.42	1.63	0.27
Patumahoe	1.84	0.40	3.13	0.98
Penrose	1.93	0.10	1.50	0.19
Takapuna	1.014	0.06	2.13	0.58
Bonferroni Value	0.01		0	
H_a = *"true autocorrelation is greater than 0"*				

The *p*-value from the S-T I_M on residuals of predicted PM_{10} on both block and random test sets was low except for two months in the block set (Mar.2016 *p*-value 0.80 and Aug. 2011 *p*-value 0.15) and one month in the random set (Nov.2014 *p*-value:0.65). The calculated *p*-value for May 2015 on block and random sets were 9.60e-08 and 1.25e-11 respectively and as a result the H_0 of no spatial correlation between the residuals is rejected. These results suggest that S-T GLM did not explain all the observed S-T variability of the PM_{10} concentration even though it captured the temporal correlation for Pakuranga station which was missed by S-T LM.

The empirical semivariogram of the residuals for May 2015 is presented in Figure 6.22. A clear decreasing in spatial dependency as distance increases was noted in plot (A) (c.f. plot (B)). Both plots appear noisy in terms of temporal dependency especially in (B) this may be due to the fact that for training set C, which is depicted in plot (B), the time points are randomly selected.

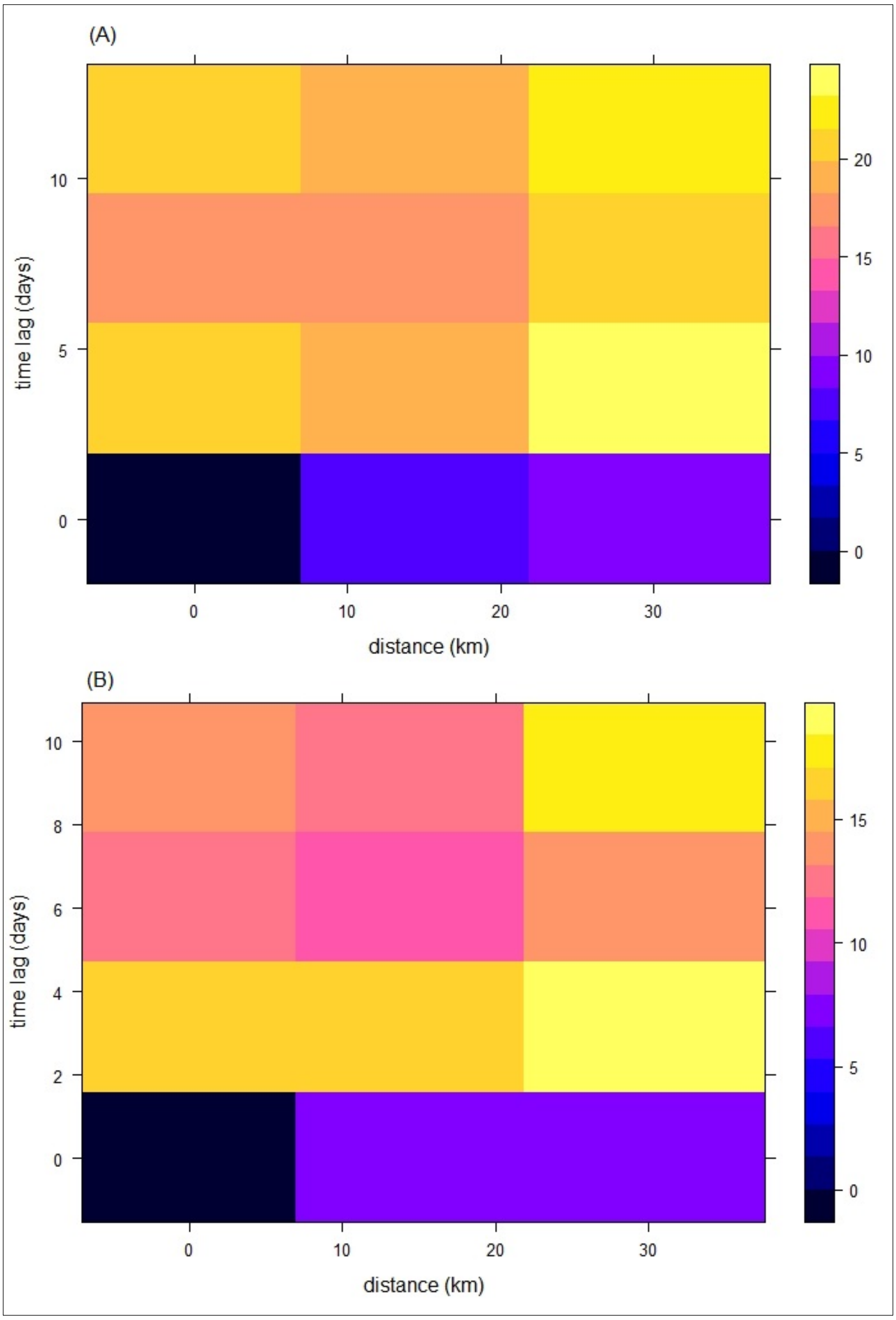


Figure 6.22: Empirical S-T semivariogram of residuals after fitting the S-T GLM model on (A) block and (B) random test sets, May 2015.

The developed S-T GLM model was a linear mixed model with random effects by day. However, presence of S-T dependency in the residuals showed **this model could *not* describe the complex inter-relationships among the covariates and possible nonlinearities between them and PM$_{10}$ concentration**. Hence, it was concluded that a more sophisticated S-T random-effects model was needed, at least in the context of this study, for modeling PM$_{10}$. The nonlinear smoother components of GAMs can be viewed as random effects for estimation purposes, so the next logical step was to extend this work to S-T GAMs.

6.6 Spatio-Temporal Generalized Additive Model (S-T GAM)

An S-T GAM model was fitted to the PM_{10} data set, with a Gamma response and a log link to accommodate nonlinear structure in the mean function. The same classes of covariates used in the above S-T LM and S-T GLM models were again considered for the development of an ST-GAM, where the response was PM_{10} concentration within the study area. To combine interacting covariates with space and time, a tensor-product structure was implemented in which the basis functions smoothing the individual covariates are combined product-wise. The product from the marginals in this experiment was achieved by using function "te" from R's "mgcv" package (Wood 2019) using `te(lon,lat,t).` In this study a thin-plate spline basis over space ("tp") and a cubic regression spline over time ("cr") were fitted. Sensitivity analysis in the context of S-T modeling was performed by evaluating the AIC as specific aspects of the model are varied. The term selection and number of knots (k) were established based on the steps described in Chapter 5.

6.6.1 S-T GAM Experiments and Results

Prediction from the constructed S-T GAM was performed on both test sets (random and block). Prediction plots for the block and random sets are provided in appendices G_A and G_B respectively. The PM_{10} prediction error for May 2015 on (A) block and (B) random test sets are shown in Figure 6.23 and Figure 6.24, respectively.

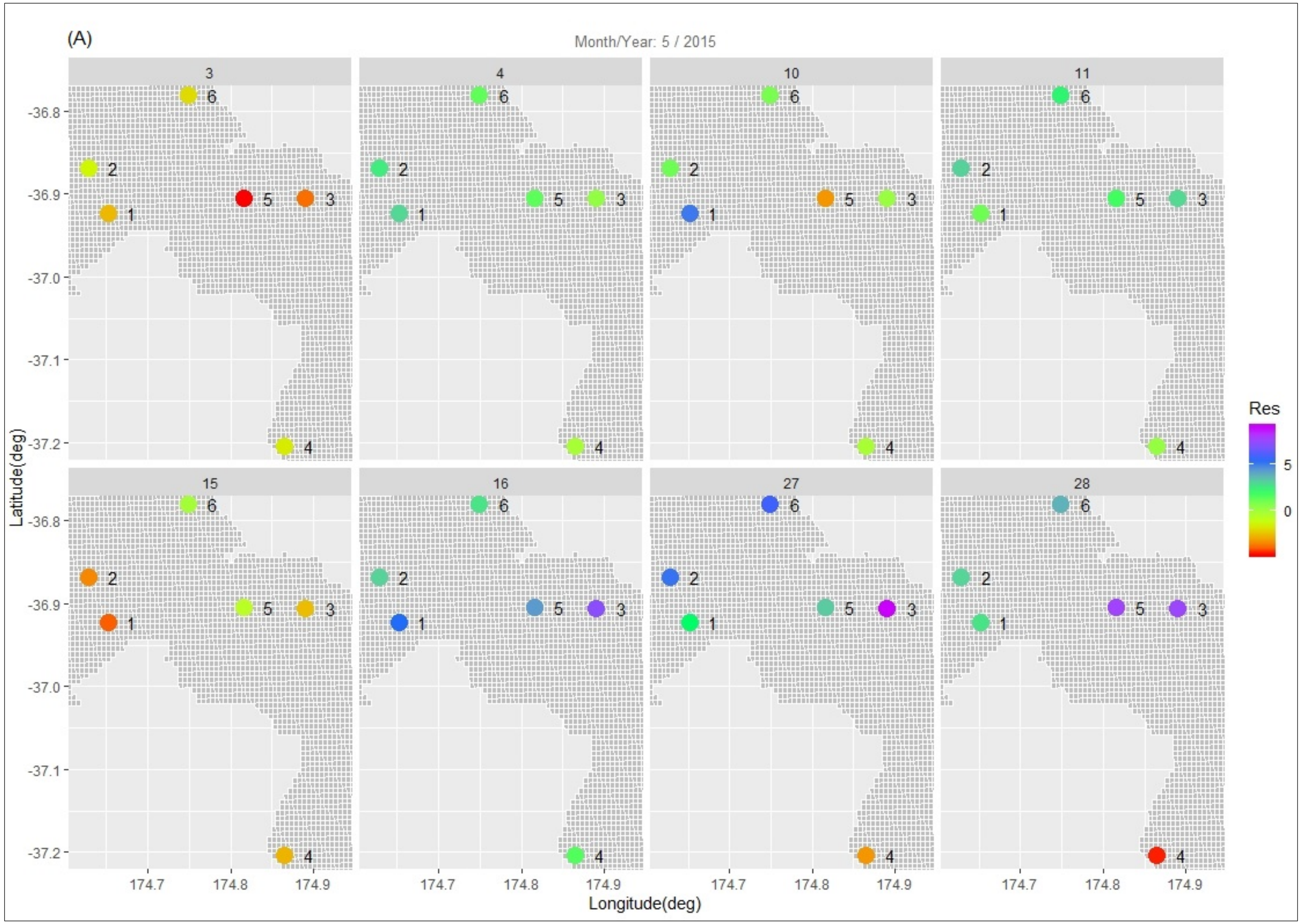


Figure 6.23: Residuals of predicted PM_{10} from fitted S-T GAM on block test set.

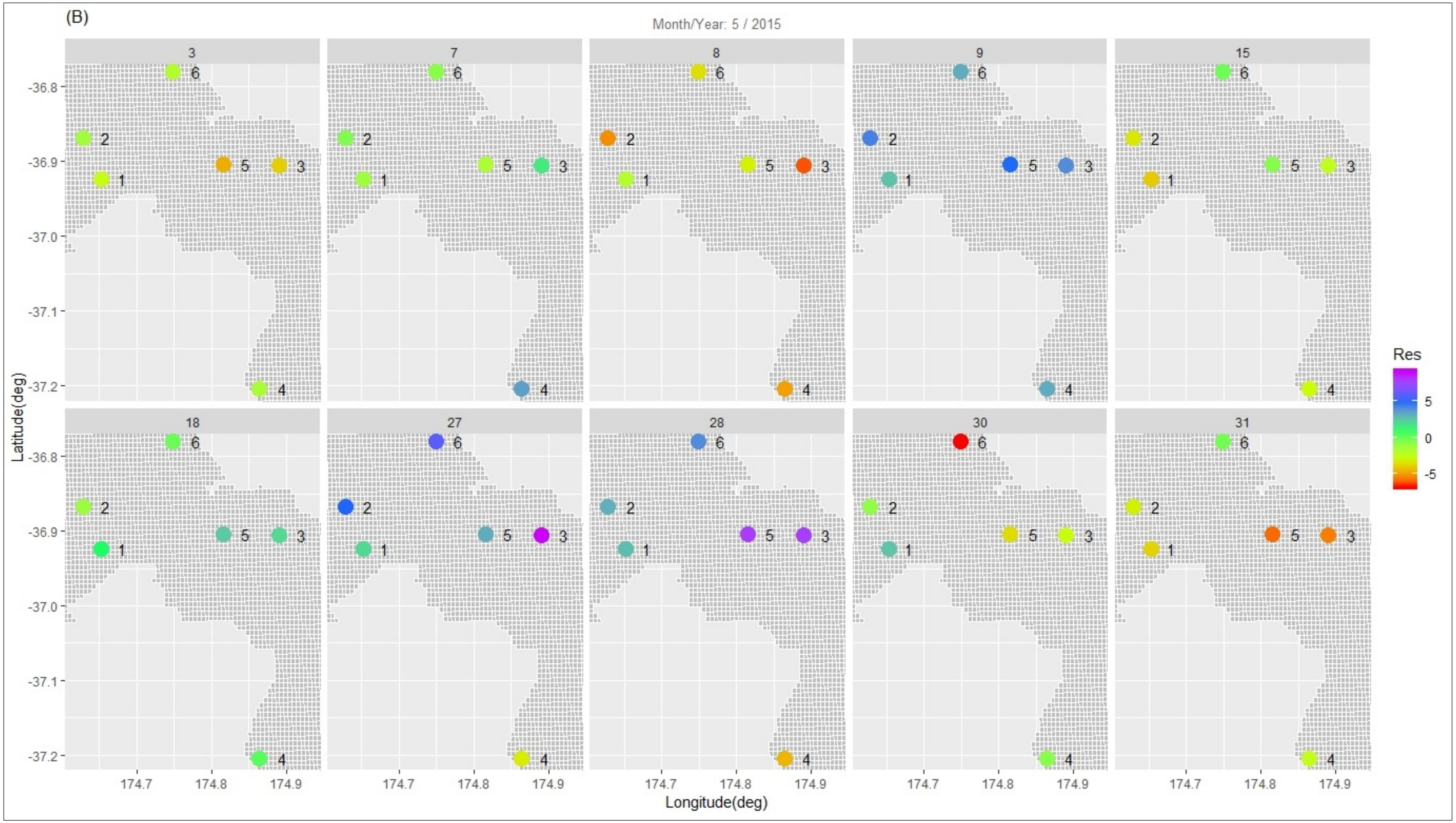


Figure 6.24: Residuals of predicted PM_{10} from fitted S-T GAM on random test set.

Spatial correlation of the residuals of the fitted S-T GAM on both test sets was evaluated using the I_M test. The results show the *p*-value is statistically significant at 5% level of significance. Therefore, the null hypothesis of no spatial correlation between residuals is rejected. This conclusion is consistent with the visualizations presented in Figure 6.23 and Figure 6.24. The *p*-value from the S-T I_M on residuals of predicted PM_{10} on both block and random test sets was low except for two months in the block set (Mar.2016 *p*-value 0.80 and Aug. 2011 *p*-value 0.15) and one month in the random set (Nov.2014 *p*-value:0.65). The calculated *p*-value on the block and random sets indicated that the H_0 of no spatial correlation between the residuals should be rejected. Temporal plots of residuals were used to investigate the temporal correlation between the residuals. The temporal plots showed high spatial and low temporal between the residuals of most stations in both sets. Temporal residuals of fitted S-T GAM for May 2015 is provided in Figure 6.25.

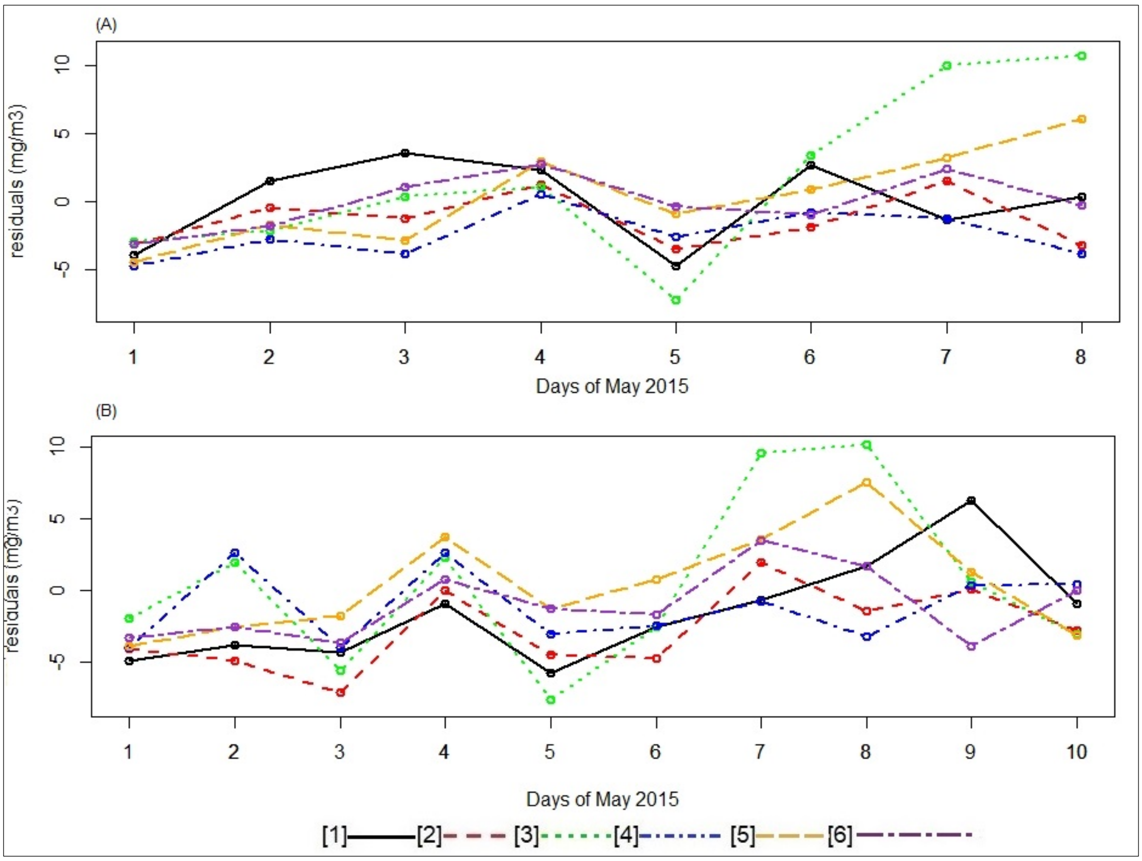


Figure 6.25: Temporal residuals of fitted S-T GAM on (A) block and (B) test sets. In the legend, [1]–[6] correspond to Glen Eden; Henderson; Pakuranga; Patumahoe; Penrose and Takapuna stations, respectively.

The plot presented in Figure 6.25 shows no temporal correlation between the residuals but sufficient spatial correlation (except for Pakuranga) as suggested by the I_M test and the spatial residual plots in Figure 6.23 and Figure 6.24 where Pakuranga's residual color was in high contrast with other stations on the 7th and 8th time points. DW test was carried on residuals of the fitted S-T GAM on both sets, and

Bonferroni value was calculated. The low ratio of Bonferroni value provided evidence that the temporal autocorrelation in the residuals is not significant.

Table 6.10: DW test on residuals of fitted S-T GAM model on block and random test sets, May 2015.

	Bock set		**Random set**	
	statistics	*p*-value	statistics	*p*-value
Glen Eden	2.37	0.71	1.11	0.06
Henderson	2.67	0.85	1.99	0.50
Pakuranga	0.87	0.05	1.64	0.27
Patumahoe	1.99	0.50	3.14	0.90
Penrose	0.89	0.05	1.19	0.08
Takapuna	1.45	0.20	1.85	0.40
Bonferroni Value	0		0	
H_a = *"true autocorrelation is greater than 0"*				

From Table 6.10 high *p*-value on both block and random sets can be noted therefore the H_0 of no temporal autocorrelation between the residuals at 5% for all stations are accepted.

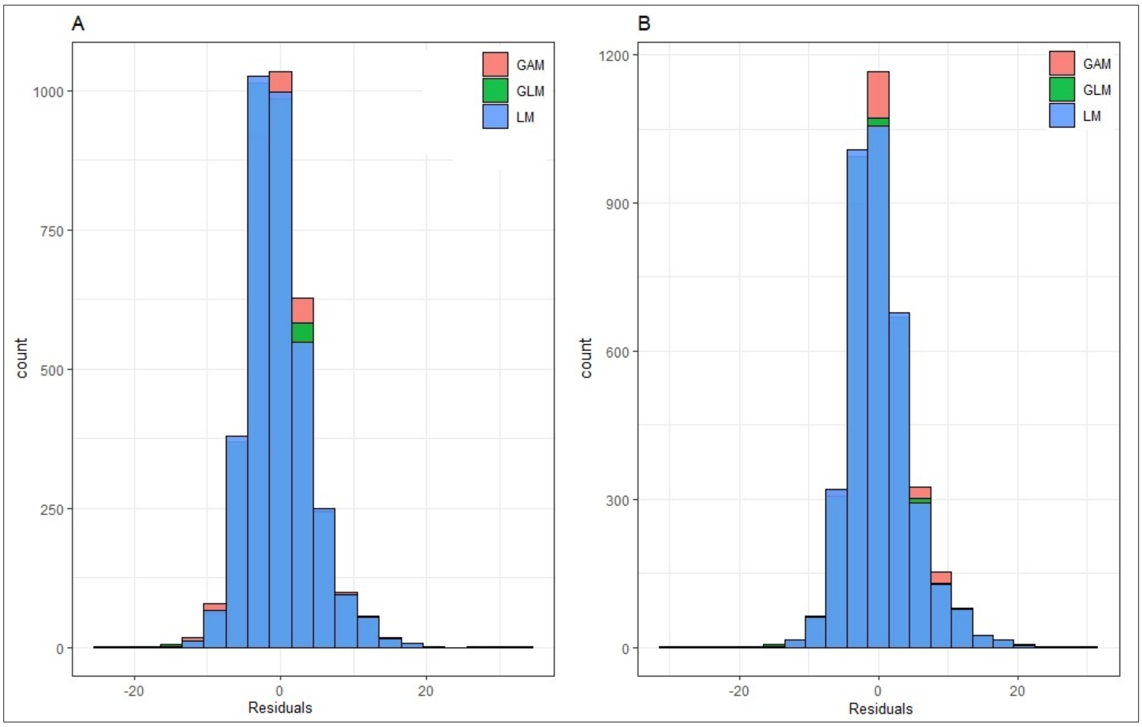


Figure 6.26: Histograms of residuals at test sets for the fitted S-T LM (blue), S-T GLM (green) and S-T GAM (red) models when the data are missing in a block (A) or at random (B).

Histograms of the S-T prediction errors using validation data for both block and missing at random sets were plotted to visually inspect the distributions of these errors from the S-T LM, S-T GLM, and S-T GAM models (Figure 6.26). The histograms reveal that the errors from the S-T LM model have a marginally larger spread. This is a first disadvantage of S-T LM when predicting the process across time points for which we have no data especially when missing at random. The variance of the residuals based on the S-T GAM model is slightly lower than the other two specially for missing at block set.

Scatter plots showing the degree of correlation between the predictions and the observations are presented in Figure 6.27. The errors distribution between the models for data points missing at random do not vary as much as they do for the test set where data are missing in multiple blocks, emphasizing the effect of missing time points on the quality of the predictions.

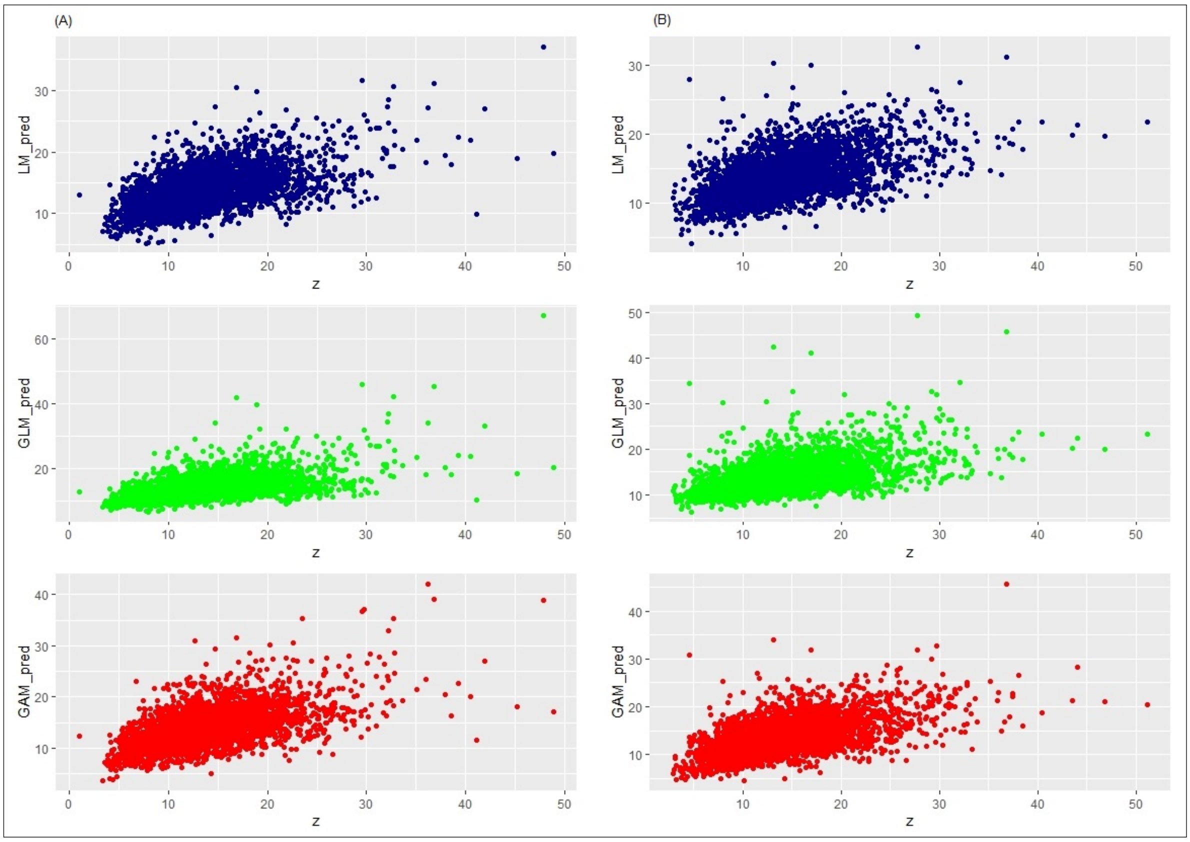


Figure 6.27: Scatter plots of the observed and predicted PM_{10} for the S-T LM (top/blue), S-T GLM (middle/green) and S-T GAM (bottom/red) models fitting block set (A) and random set (B).

Cross-validation diagnostics for each model were calculated and are summarized in Table 6.11. The S-T GAM model outperforms the S-T GLM and S-T LM models on most of the diagnostics for both block and random sets. The lower value of the MSEP for block sets is an indicator of sensitivity of missing data patterns on the performance of the models. For all models, it was noted that the SCV and CRPS need to be treated with care in a spatial or S-T setting, due to the correlation in errors that are not taken into account when computing these measures (C. K. Wikle et al., 2019b).

Table 6.11: Cross-validation metrics for the models fitted to the Auckland PM_{10} data set where data has been left out for two entire time intervals (resulting in two 'missing blocks') and where data has been removed at random. The bias (better when closer to zero), the predictive cross-validation measure (PCV, better when lower), the standardized cross-validation measure (SCV, better when closer to 1), and the continuous ranked probability score (CRPS, better when lower) are presented as performance metrics.

	Bock set					**Random set**				
	Bias	PCV	SCV	CRPS	MSEP	Bias	PCV	SCV	CRPS	MSEP
S-T LM	0.27	19.15	**0.96**	-0.16	19.15	-0.26	21.28	**1.01**	0.16	21.28
S-T GLM	0.28	19.91	1.65	**-0.17**	19.91	**-0.25**	22.10	1.71	**0.15**	22.10
S-T GAM	**0.17**	**18.92**	2.13	-0.11	**18.92**	-0.36	**20.43**	1.79	0.22	**20.43**

The AIC and BIC for all three models were computed using the number of estimated parameters and the size of the training set used to fit the models. Both the AIC and BIC presented in Table 6.12 are lower for the S-T GAM model than for the S-T GLM and S-T LM models.

Table 6.12: AIC and BIC for three models.

	Block set		**Random set**	
	AIC	BIC	AIC	BIC
S-T LM	56322	56437	54582	54697
S-T GLM	54926	55041	53303	53417
S-T GAM	**50534**	**51821**	**51147**	**51697**

Because the difference between the criterion for the models is large it is safe to conclude that **the S-T GAM model is a more accurate representation of the PM_{10} data and is preferable to the S-T GLM and S-T LM model for modeling and predicting the Auckland PM_{10} data set**.

6.7 Conclusion

The main aim of this Chapter was to report on the exploration of two objectives of spatio-temporal statistical modeling – interpolation of the PM_{10} concentrations in the study area using the S-T data and performing parameter inference. The uncertainty in the predictions and parameter estimates were also quantified. Two potential modeling solutions that initially consider the S-T error process were investigated. This enables the benefits and weaknesses of standard commonly used modeling approaches in modeling and predicting Auckland PM_{10} concentration to be investigated.

Section 6.3.2 presented deterministic prediction in order to obtain PM_{10} predictions at spatio-temporal locations using Auckland PM_{10} S-T data set. The IDW model precision across years was shown to vary between 9.64 µg m^{-3} and 12.87 µg m^{-3}, though this was slightly better in 2015 at 9.64 µg m^{-3} on average.

There are a very limited number of spatio-temporal models for prediction of PM_{10} in the current literature with which to compare the results of the models developed in this research. Yanosky et al. (2008) used IDW to predict monthly PM_{10} concentrations for 13 states in the northeast of the US. They mention in their report that a model with monthly spatial terms performed better than one with seasonal terms. Although they reported good model performance for their monthly model, both in rural and urban locations, they did not provide details of the IDW parameter values nor the feature/variable selection approach used. Because of the lack of model details, it was not possible to meaningfully compare the results reported here for Auckland with those of Yanosky et al.

In Section 6.3.3 a statistical regression model was used to obtain predictions for Auckland PM_{10} concentration, assuming that all the S-T dependence can be considered by using the trend terms. Such a model explicitly accounted for model error. The obtained model-based prediction-error variance provided useful insights into the model's performance. The assumption was that these trend terms can capture the PM_{10} concentration's large-scale S-T variability leaving behind the smaller-scale variability that can be statistically modeled using S-T covariances. The results of these experiments however, showed that the S-T linear regression model with trend terms was not capable of explaining the S-T variability of PM_{10} concentration over the study area. To accommodate the nonlinear structure in the mean function, S-T GAM model were further developed so that transformation of the mean response has an additive form. The difference between the AIC/BIC criterion for the S-T GAM model was notable compared to other two (S-T LM and S-T GLM) models concluding that S-T GAM model is a better representation of the data and is preferable to the S-T GLM model for modeling and prediction of Auckland PM_{10} data.

Chapter 7 Descriptive Spatio-Temporal Modeling

The work reported in this Chapter discusses descriptive approaches to S-T modeling from a high-level perspective and the study undertaken to develop descriptive (marginal) models by characterizing the S-T dependence structure through S-T covariances. This study in turn leads to models that are analogous to the ubiquitous geostatistical models that are employed in section 7.4.

Geostatistics distinguishes itself within the general field of statistics by having at its foundation, the theoretical principle of random spatial processes. Such random spatial processes are commonly present in environmental phenomena. In this Chapter separable and non-separable model are described and investigated. The best fitted model is then used to perform S-T Universal kriging as a means of modeling PM_{10} concentrations.

7.1 Introduction

Descriptive and dynamic approaches are able to address S-T statistical model problems by capturing statistical dependencies in S-T phenomena in two different ways although both have a common underlying probability model (C. K. Wikle et al., 2019b). Chapter 5 on time series analysis showed that both stationary AR(1) process models performed better ($0.62 <$ GAMM $R^2 < 0.67$) if not comparable ($0.32 <$ GAM $R^2 < 0.47$) compared to those reported for modeling PM_{10} in the Northeastern and Midwestern U.S. (GAMM $R^2 = 0.58$ and (GAM $R^2 = 0.49$) (Yanosky et al., 2014). From the dynamic point of view, the model stated that the value of PM_{10} at time *t* is equal to a transition factor times the value at the *t-1*, combined with an independent error. Descriptively, similar probability structures can be achieved by identifying the correlation between two values at any two given time points to be an exponentially decreasing function of the lag between the two time points (C. K. Wikle et al., 2019b). This descriptive approach can be more fit for purpose when full knowledge of the dynamics of the system are not available (C. K. Wikle et al., 2019b). As part of the descriptive approach attempts are made to characterize the S-T process in terms of its mean and its covariance function. This approach has been recognized historically in spatial statistics and is the basis of the Kriging (Cressie, 1990) methods (C. K. Wikle et al., 2019b). Using the mean and covariance functions a good fit to the data can be achieved and therefore the S-T variability can be appropriately described. In this research, in an attempt to provide comprehensive coverage, three commonly used S-T variogram models including the metric model (Myers, 2004), sum-metric model (Myers, 2004) and the product-sum model (Myers , 2004) are used.

Exploratory analysis of the marginal spatial and temporal variogram of PM_{10} is conducted for modeling the S-T empirical variograms. This analysis fulfills the second objective of the descriptive approach which is to discover the best variogram model for explaining PM_{10}. The best model as defined by Guo et al. (2015) is one that is flexible and efficient enough to measure the characteristics of S-T correlation structure of atmospheric PM_{10} and to provide the best prediction accuracy in creating the interpolation surfaces using S-T kriging. S-T geostatistics using relevant S-T variogram is adopted as a model for estimating PM_{10} concentration at daily and monthly scale. The methods and experiments detailed in this section focus on determining the statistically "optimal" weights using kriging which is a geostatistical linear combination method (Krige, 1951). Kriging uses statistical dependency properties, such as covariances between observed locations, to determine the weights while still respecting the measurement uncertainty (Smith et al., 2007).

7.2 Geostatistical Methods

The major challenge in S-T interpolation is identifying the S-T dependence structure (Li et al., 2016). Geostatistical methods include and model the spatial correlation between variables resulting in an unbiased estimation with a lowest and known variance (Oliver & Webster, 2015). In a purely spatial interpolation approach, time can be included by performing series of spatial interpolation snapshots and a one-dimensional slice of the space-time variogram that corresponds to zero-time lag is employed. Similar to a spatial variogram or covariance, the S-T dependency can be quantified by estimating and modeling S-T variogram functions that are conditionally negative definite (Donald E. Myers, 2004). Verifying such a condition is not straightforward in cases such as the Auckland PM_{10} dataset used in this research as the number of time points is larger than the number of spatial data points. Also, given that this study's focus within the descriptive approach is on the first two moments, means, variances, and covariances of $Y(\cdot;\cdot)$, it is assumed that the underlying process is Gaussian (C. K. Wikle et al., 2019c).

Time is implicitly treated as another dimension in simple S-T kriging, and covariance functions define co-variability among any two space-time locations in the domain of interest. In S-T kriging, the weighted residuals between the observations and their marginal means are taken by the conditional mean and the results are added back into the marginal mean related to the prediction location. These weights are functions of the covariances and the measurement error variance (C. K. Wikle et al., 2019c). To implement optimal prediction using geostatistical methods such as kriging, the joint S-T dependence structure of a S-T process needs to be described. Empirical S-T co-variograms are typically used to measure this joint S-T dependency. Unlike the spatial covariance estimates presented in Chapter 6, co-variability in S-T data is described as a function of specific lags in time and in space where the time lag is a scalar value and the lag in space is a vector.

The experimental S-T semivariogram $\gamma_{st}(\mathbf{h}_s, h_t)$ is the primary tool for verifying the applicability of S-T geostatistics. Assuming a spatial dependency for first moment and lag differences in space and time for the second moment the empirical semivariogram can be defined as:

$$\boldsymbol{\gamma}_{\mathbf{st}}(\mathbf{h_s}, \mathbf{h_t}) = \frac{\mathbf{1}}{\mathbf{2N(h_s,h_t)}} \sum_{\mathbf{i=1}}^{\mathbf{N(h_s,h_t)}} \left[\mathbf{z(s,t)_i} - \mathbf{z((s,t)_i)} + \left((\mathbf{h_s}, \mathbf{h_t})\right)\right]^{\mathbf{2}} \qquad \text{Eq. 7.1}$$

where

$\mathbf{N}(\mathrm{h_s}, \mathbf{h_t})$ is the number of any two locations parted by the vector **h**= $(\mathrm{h_s}, \mathbf{h_t})$,

$\mathrm{h_s}$ is the spatial lag,

$\mathbf{h_t}$ is the temporal lag and

$z(s,t)_i$is the value of the variable at the S-T location $(s,t)_i$.

Spatio-temporal kriging predictors require knowledge of the S-T covariances among the hidden random process evaluated at pairs of space and time locations. The covariance function must be non-negative-definite to guarantee non-negative kriging variances. The experimental S-T covariance C^st (h) for sh$_\mathrm{s}$ and $\mathbf{h_t}$ is defined as :

$$\boldsymbol{C}\hat{}_{\boldsymbol{st}}(\mathbf{h}_{\boldsymbol{s}}, \boldsymbol{h}_{\boldsymbol{t}}) = \frac{\mathbf{1}}{\mathbf{2}\boldsymbol{N}(\mathbf{h}_{\boldsymbol{s}},\boldsymbol{h}_{\boldsymbol{t}})} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}(\mathbf{h}_{\boldsymbol{s}},\boldsymbol{h}_{\boldsymbol{t}})} \boldsymbol{z}_{\boldsymbol{i}}(\mathbf{s},\boldsymbol{t})_{\boldsymbol{i}}.\boldsymbol{z}((\mathbf{s},\boldsymbol{t})_{\boldsymbol{i}}) + \left((\mathbf{h}_{\boldsymbol{s}}, \boldsymbol{h}_{\boldsymbol{t}})\right) - \hat{\boldsymbol{m}}_{-\boldsymbol{h}}.\hat{\boldsymbol{m}}_{+\boldsymbol{h}} \qquad \text{Eq. 7.2}$$

where

m^**-h** is the mean of the tail values and

m^**+h** is the mean of the head values.

A valid semivariogram may present a discontinuity in zero called *nugget* that quantifies the discontinuity among the observation and the smooth underlying process that may be due to the measurement error or to dynamics on a smaller scale than the one of our measurement grids. The distances between pairs of observed data at which the variogram is calculated are called lags. Another related property of the semivariogram is the sill value where the semivariogram model attains at the range. The relative nugget, ratio of the nugget to the total sill, tends to increase as the data scatters (Goovaerts, 1997). The semivariogram range quantifies the range of influence of the Gaussian process. The two elements of the process are uncorrelated when the distances are greater than the range.

Fitting a variogram model in the semivariogram modeling process always involves uncertainty related to the choice of the semivariogram model parameters. A model must generate variogram value for all separation distance. Next step is to select the best fitted function that captures the overall features of the experimental semivariogram (Αινσλιε Μ. Δενηαμ, 2012). Figure 7.1 shows variogram parameters and semivariogram modeling steps. The term semivariogram and variogram are often used interchangeably in literature. By definition, γ(*h*) is the semivariogram and the variogram is 2γ(*h*).

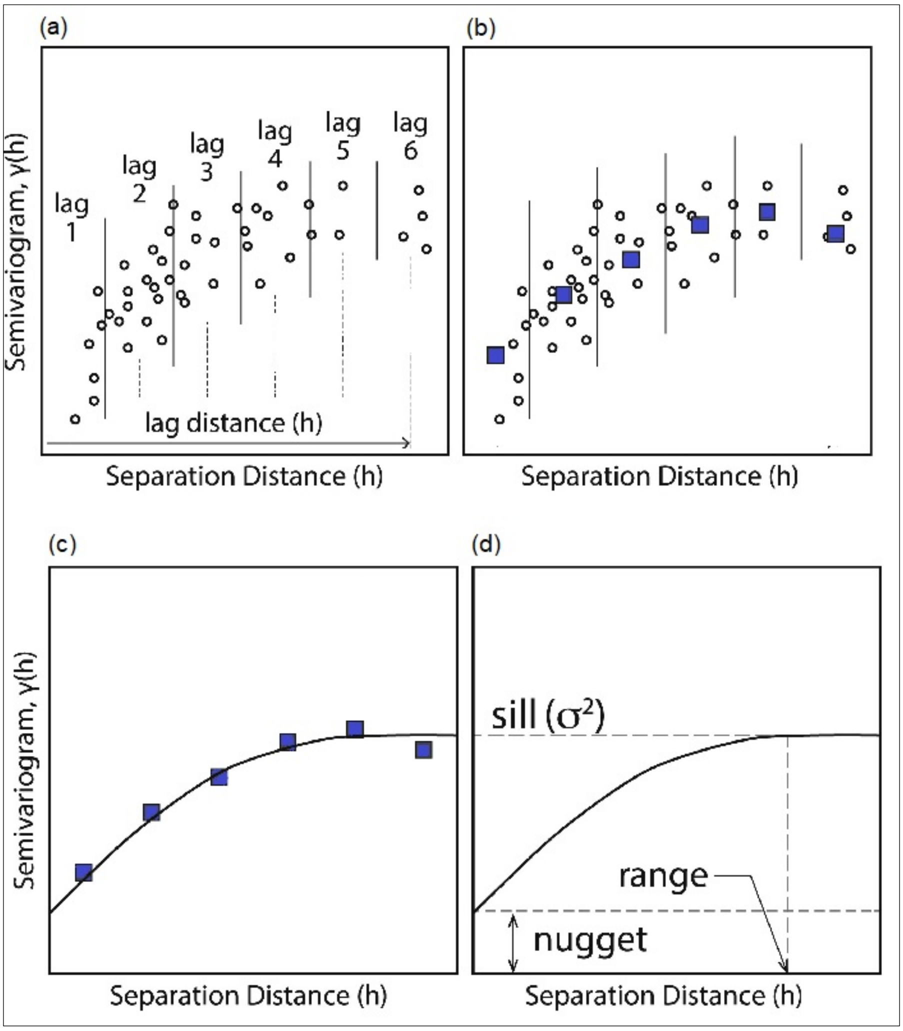


Figure 7.1: Semivariogram modeling steps: (a) data shown as circles are plotted as function of separation distance between each point. The separation distance is partitioned into lag bins. (b) calculate the empirical semivariogram for each bin and plot it as a function of h. (c) Model semivariogram fitted to empirical semivariogram data. (d) Model parameters for spatial structure of the data. Adopted from (Hanke et al., 2018) with permission.

In an isotropic phenomenon spatial continuity pattern is independent from the spatial direction, otherwise it is anisotropic. Temporal anisotropy is not applicable as there is only one temporal dimension. In geometric anisotropy, shape and sill of directional semivariograms do not change, but the value of the spatial range varies in different spatial directions. In a zonal anisotropic semivariogram the sill value depends on the spatial direction as well as the distance.

Isotropic case:

$$\boldsymbol{\omega(\theta) = \sum_{l=1}^{L} \omega_l[\gamma\hat{}(h_l) - \gamma(h_l, \theta)]^2} \qquad \text{Eq. 7.3}$$

where

θ is a parameter vector usually containing the range and sill for a given semivariogram model

$\gamma\hat{}(h_l)$ is sample semivariogram at lag distance $h_l = |\mathbf{h}_l|$,

$\gamma(h_l;\ \theta)$ is model semivariogram at lag distance h_l,

ω_l = weight for *l-th* squared error value $[\gamma\hat{}\ (h_l) - \gamma(h_l;\ \theta)]^2$ at lag distance h_l

Anisotropic case:

$$\boldsymbol{\omega}(\boldsymbol{\theta}) = \sum_{\mathbf{k=1}}^{\mathbf{K}} \sum_{\mathbf{l=1}}^{\mathbf{L}} \boldsymbol{\omega}_{\mathbf{l}}^{\mathbf{k}} \left[\boldsymbol{\gamma}\hat{}(\mathbf{h}_{\mathbf{l}}^{\mathbf{k}}) - \boldsymbol{\gamma}(\mathbf{h}_{\mathbf{l}}^{\mathbf{k}}, \boldsymbol{\theta})\right]^2 \qquad \text{Eq. 7.4}$$

where

$\gamma\hat{}(\mathbf{h}_l^k)$ is sample semivariogram at lag h_l for *k-th* direction,

$\gamma(\mathbf{h}_l^k, \boldsymbol{\theta})$ is model semivariogram at $\mathbf{h}_l^k$ for *k-th* direction

ω_l^k is weight for *l-th* squared error value $\left[\gamma\hat{}(\mathbf{h}_l^k) - \gamma(\mathbf{h}_l^k, \boldsymbol{\theta})\right]^2$ at lag distance $\mathbf{h}_l^k$ along *k-th* direction.

A general approach has been proposed to distinguish between isotropy and anisotropy by fitting a variogram model along four spatial directions (Chappell & Agnew, 2012; Eriksson & Siska, 2000). In this approach, the spatial correlation structure is isotropic if the regularity parameter changes in all variograms are all the same; otherwise, it is anisotropic. However, the challenge in this thesis's study area is the low number of data locations. In data sets where the total number of pairs is determined by the number of data locations, the directional sample variogram splits the pairs into four sets. Hence, having small number of pairs per distance class, as in the case of this research, makes it difficult to use the plotted sample directional variograms to fit models. To overcome this problem spatial trend analysis was performed using the ESRI Spatial Analyst by Esri. ArcGIS®.

7.3 Spatio-Temporal Variograms and Covariance Models

7.3.1 Separable Spatio-Temporal Variograms

Estimation of the mean and spatial structure function are the two key sequential steps involved in geostatistical interpolation in a purely spatial situation (D. E. Myers, 2004). Space-time interpolation involves similar steps but the estimation the structure function is more complex. Spatial-temporal models can be constructed by treating the space and time separately, or by including time as an additional dimension with a different unit of measurement (A. M. Denham, 2012). In a separable S-T covariance function the S-T process is a joint process of two individual, a purely spatial and a purely temporal process (De Cesare et al., 1997).

Separable classes of S-T covariance functions offer a convenient way to meet the space-time covariance function validity criteria. The valid separable class is given by (D. E. Myers, 2004):

$$\mathbf{c}(\mathbf{h};\boldsymbol{\tau}) = \mathbf{c}^{(\mathbf{s})}(\mathbf{h}).\mathbf{c}^{(\mathbf{t})}(\boldsymbol{\tau}) \qquad \text{Eq. 7.5}$$

where

$c^{(s)}(h)$ denotes a spatial covariance function, and

$c^{(t)}(\tau)$ denotes a temporal covariance function.

The structure of separable models suggests that the spatial structure is unchanged during the time and the temporal structure fixed at all locations (Cressie & Huang, 1999). The separability property also implies that the two are separate processes acting independently from each other, which is an unrealistic assumption that is very seldom true in the case of real-world processes (Bruno et al., 2009a). The joint S-T correlation function is obtained by using the marginal spatial and temporal correlation functions. Thus, instead of identifying realisations of two separate processes, only the joint process is observed (A. M. Denham, 2012). The separable S-T covariance models need to estimate a small number of parameters. This has led to use of separable S-T covariances where they are not physically reasonable (Bruno et al., 2009a; Cressie & Huang, 1999). Significant effort has been directed towards the development of non-separable S-T variograms and covariance models to model the space time interactions in environmental processes. A comparable list of valid spatial-temporal covariance functions was provided by De Iaco and Myers (2002). Various non-separable S-T models have been proposed by authors and researchers (De Iaco et al., 2002a; Gneiting, 2002; Mateu et al., 2008). A comparative review of many of these accepted and implemented models was performed in 2010 (De Iaco, 2010).

The product-sum model (De Iaco et al., 2002a) was found to give a marginally better fit than the Cressie–Huang model in a hydrological study estimating the runoff along a stream network topology (Skøien & Blöschl, 2005). Spatio-temporal analysis of daily Ozone data collected in Milan (Italy) showed, based on mean square error (MSE) values, that the product-sum model gave a better fit than Cressie–Huang's and Geniting's models (De Iaco, 2010). The S-T experimental semivariogram of the NO_2 emissions of soil from a tea plantation in central China was obtained using separable, product-sum, metric and sum-metric S-T semivariogram models. The sum-metric model was reported in this case to perform the best (Liu et al., 2016). The Sum-metric and product-sum models were used by

Menezes et al. (2017) to model the S-T concentration of NO_2 over Portugal's mainland. The models were compared using the ME and MSE metrics. The results for each of the models were very similar. The advantage of an extra parameter (anisotropy) in the sum-metric to deal with spatial and temporal distances as well as using particular variogram for space, time, and space–time resulted in the authors ultimately choosing the sum-metric model over the product-sum model. In the study by Ahmed et al. (2018) different covariance functions were utilized to model the S-T variogram of monthly averages of PM_{10}, NO_2 and sulfur dioxide (SO_2). They reported that in all cases the sum-metric model outperformed the separable, metric, and product-sum models.

Non-separable S-T models for nonstationary data have been proposed in recent years. However, only limited practical applications of these non-stationary covariance models are available in literature. Some researchers have implemented these models to model Ozone concentrations (Bruno et al., 2009b; Das & Bhattacharya, 2014; Fuentes et al., 2007). In another study, analysis of S-T changes in neurodegeneration related to Alzheimer's disease using a non-separable, nonstationary model was performed (Marco et al., 2015).

An overview of the mathematical descriptions of the most used separable and non-separable model is provided in the following subsection of this thesis. These models are employed in this study to explore the potential of commonly used S-T methods for daily PM_{10} measurements across Auckland. The choice of method was based on its widespread prior use and this choice was made to allow for comparison with other studies in literature, so this overview is by no means exhaustive. Finally, the results of applying each of these models to the case studies PM_{10} data are reported.

7.3.1.1 Sum Model

In the sum model (aka the zonal model), time is treated as another dimension resembling a 'zonal 'anisotropy. The sum model CST is computed according to Myers (2004) as follows:

$$\mathbf{C_{ST}}(\mathbf{h_s}, \mathbf{h_t}) = \mathbf{C_S}(\mathbf{h_s}) + \mathbf{C_T}(\mathbf{h_t}) \qquad \text{Eq. 7.6}$$

or

$$\boldsymbol{\gamma}_{\mathbf{ST}}(\mathbf{h_s}, \mathbf{h_t}) = \boldsymbol{\gamma}_{\mathbf{S}}(\mathbf{h_s}) + \boldsymbol{\gamma}_{\mathbf{T}}(\mathbf{h_t}) \qquad \text{Eq. 7.7}$$

where

$C_S(\mathbf{h}_s)$ is spatial covariances,

$C_T(h_t)$ is temporal covariances,

$\gamma_S(\mathbf{h}_s)$ is spatial variogram, and

$\gamma_T(h_t)$ is temporal variogram.

The issue with the sum model is that the obtained spatial-temporal covariance/variogram may not comply with strict definiteness and strict conditional negative definiteness even if the individual spatial and temporal covariances and variograms meet the criteria (Myers & Journel, 1999).

The sum model is unbounded when either a component in right side of the Equation 7.24 is unbounded (Myers 2004) and the spatial component can be combined with a geometric anisotropy as it is a separable model. According to Myers (2004) this will produce uncertainty in the results. However, some authors have suggested that using such a model is reasonable as the coefficient matrix may be invertible for some data locations. One such example is that of the ozone concentration model reported by Buxton and Pate (1994).

7.3.1.2 Metric Model

In this research the metric space-time covariance model and variogram are generated according to Myers (2004) in which $R^d \times T$ is a d+1 dimensional space. The model is formulated as follows:

$$\mathbf{C_{ST}(h_s, h_t) = C(a_1|h_s| + a_2|h_t|)} \qquad \text{Eq. 7.8}$$

and

$$\boldsymbol{\gamma}_{\mathbf{ST}}\mathbf{(h_s, h_t)} = \boldsymbol{\gamma}\mathbf{(a_1|h_s| + a_2|h_t|)} \qquad \text{Eq. 7.9}$$

where

C(ST) is a strictly positive definite function on R^{d+1},

γ(ST) is strictly conditionally negative definite on R^{d+1}, and

the coefficients a_1, $a_2 \in R^d$ facilitate the comparison between space and time.

7.3.1.3 Sum-Metric Model

Myer's Sum model can be combined with a metric model (D. E. Myers, 2004) to provide the sum-metric model γ_{ST} as follows:

$$\boldsymbol{\gamma}_{\mathbf{ST}}\mathbf{(h_s, h_t)} = \boldsymbol{\gamma}_{\mathbf{S}}\mathbf{(h_s)} + \boldsymbol{\gamma}_{\mathbf{T}}\mathbf{(h_t)} + \Upsilon\mathbf{(a_1|h_s| + a_2|h_t|)} \qquad \text{Eq. 7.10}$$

As with the sum-model it is possible to replace $a_1|\mathbf{h}_s| + a_2|\mathrm{h}_t|$ as a distance function on $R^d \times T$, with $a_1|\mathbf{h}_s|^2 + a_2|\mathrm{h}_t|^2$ such that

$$\boldsymbol{\gamma}_{\mathbf{ST}}\mathbf{(h_s, h_t)} = \boldsymbol{\gamma}_{\mathbf{S}}\mathbf{(h_s)} + \boldsymbol{\gamma}_{\mathbf{T}}\mathbf{(h_t)} + \ \Upsilon\mathbf{(a_1|h_s|^2 + a_2|h_t|^2)} \qquad \text{Eq. 7.11}$$

where

$\gamma_S(\mathbf{h}_s)$ and $\gamma_T(h_t)$ denote the space and the time variograms, respectively.

The respective marginals are given as:

$$\boldsymbol{\gamma}_{\mathbf{ST}}\mathbf{(h_s, 0)} = \boldsymbol{\gamma}_{\mathbf{S}}\mathbf{(h_s)} + \boldsymbol{\gamma}\mathbf{a_1|h_s|} \qquad \text{Eq. 7.12}$$

and

$$\boldsymbol{\gamma}_{\mathbf{ST}}\mathbf{(0, h_t)} = \boldsymbol{\gamma}_{\mathbf{T}}\mathbf{(h_t)} + \boldsymbol{\gamma}\mathbf{a_2|h_t|} \qquad \text{Eq. 7.13}$$

In a bounded model, the sill parameter in both space and time marginal will have a similar type of sill but the range of the parameter varies. Therefore, it is feasible to choose a sum-metric model with distinct marginals. As the strict conditional negative definiteness is guaranteed by γ, the remaining components can be semi-definite (D. E. Myers, 2004).

7.3.1.4 Product Model

In a product model the strictly positive definite covariance function is calculated according to Myers (2004) as follows:

$$\mathbf{C_{ST}(h_s, h_t) = C_S(h_s) \times C_T(h_t)} \qquad \text{Eq. 7.14}$$

The variogram and marginal are given by:

$$\boldsymbol{\gamma_{ST}(h_s, h_t) = C_T(0)\gamma_S(h_s) + C_T(0)\gamma_T(h_t) - \gamma_S(h_s) \times \gamma_T(h_t)} \qquad \text{Eq. 7.15}$$

$$\boldsymbol{\gamma_{ST}(h_s, h_t) = \gamma_{ST}(h_s, 0) + \gamma_{ST}(0_s, h_t) - [1/CT(0)CS(0)]\gamma_{ST}(h_s, 0) \times \gamma_{ST}(0_s h_t)} \qquad \text{Eq. 7.16}$$

7.3.2 Non-separable Models

7.3.2.1 Product-Sum Model

The product-sum model (De Iaco et al., 2001) is constructed by combining both “sum” and the “product” of the spatial and temporal covariance functions and can overcome the limitations related to separate models. A strictly positive definite $C_{ST}(\mathbf{h}_s, h_t)$ the on $R^d \times T$ is given by:

$$\mathbf{C_{ST}(h_s, h_t) = K_1 C_S(h_s) \times C_T(h_t) + K_2 C_S(h_s) + K_3 C_T(h_t)} \qquad \text{Eq. 7.17}$$

where

C_T and C_S are strictly positive definite temporal and spatial covariance models, respectively.

The variogram is given by:

$$\gamma_{ST}(\mathbf{h}_s, h_t) = [K_1 C_T(0) + K_2]\gamma_S(\mathbf{h}_s) + [K_1 C_s(0) + K_2]\gamma_T(h_t) - K_1\gamma_S(\mathbf{h}_s) \times \gamma_T(h_t)$$

Eq. 7.18

or

$$\boldsymbol{\gamma_{ST}(h_s, h_t) = \gamma_{ST}(h_s, 0) + \gamma_{ST}(0_s, h_t) - K\gamma_{ST}(h_s, 0) \times \gamma_{ST}(0_s, h_t)} \qquad \text{Eq. 7.19}$$

The sufficient condition for K is given by (De Iaco et al., 2001):

$$\mathbf{0 < K \leq 1/max\,(sill\,\gamma_{ST}(h_s, 0), sill\,\gamma_{ST}(0_s, h_t)} \quad \text{Eq. 7.20}$$

Unlike the sum, metric, and sum-metric models the assumption of second order stationarity is needed in such models (Myers 2004). This model does not need the use of a space–time metric and is more accommodating than the non-separable covariance models for estimating and modeling S-T correlation structures (De Iaco et al., 2002b).

7.4 Spatio-Temporal Interpolation Methods

Once a semivariogram or covariance model of S-T dependence is identified, estimation of the attribute value at the unsampled location (**h**,t) can be applied. In Kriging, the unknown value is estimated using weighted linear combinations of neighboring data, or a subset of the residuals, that are subject to the mean function definition. Selection of neighboring data is based on the spatial and temporal distance from the estimation data point (**h**,t) and weighted by including the closeness of each sample data to the location of prediction. Linear kriging methods are a Simple Kriging (SK) with known and constant mean, Ordinary Kriging (OK) where mean is not known but is constant and Universal Kriging (UK) where the mean is an unknown linear combination of random functions (Li et al., 2015). All kriging methods are modifications of the basic linear regression estimator Z*(**h**,t) Denham (2012):

$$\mathbf{Z^*(h,t) - m(h,t) = \textstyle\sum_{\alpha=1}^{n(h,t)} \lambda_\alpha(h,t)[Z(h,t)_\alpha - m(h,t)_\alpha]} \quad \text{Eq. 7.21}$$

where

$Z(\mathbf{h},t)_\alpha$ denotes realization of the $(\mathbf{h},t)_\alpha$,

$\lambda_\alpha(\mathbf{h},t)$ denotes the weight assigned to the $Z(\mathbf{h},t)_\alpha$,

$m(\mathbf{h},t)$ denotes the expected value of the random variable $Z(\mathbf{h},t)$, and

$m(\mathbf{h},t)_\alpha$ denotes the expected value of the random variable $Z(\mathbf{h},t)_\alpha$.

The number of neighboring data used in the estimation and the weights allocated to each data point could vary in different directions. Kriging is an unbiased method, as the errors mean is zero, aiming at reducing the estimation or variance of the errors. The variance is given by

$$\mathbf{\delta_E^2(u,t) = Var[Z^*(u,t) - Z(u,t)]} \quad \text{Eq. 7.22}$$

which is reduced under the constraint that

$$\mathbf{E[Z^*(u,t) - Z(u,t)] = 0} \quad \text{Eq. 7.23}$$

SK does not adjust to local trends; instead, it relies on a constant, global mean. The assumption of second-order stationarity in SK allows the random functions to be defined as residuals with a zero mean

(Mpanza, 2015). Prior to estimation, the mean is deducted from the observations. The prediction surface is adjusted using the residual values. The weights in SK are defined so that the error variance is reduced under the constraint of the unbiasedness of the estimator (Ainslie M. Denham, 2012).

OK assumes that random variables are stationary where the mean is not known (Armstrong, 1998). Local mean is used for approximating the mean at each weighting the search neighborhood. The unknown mean is estimated alongside the residual component. The kriging weights are forced to sum to 1 resulting in filtering the unknown local mean from the estimator (Ainslie M. Denham, 2012):

$$\mathbf{Z}^*_{\mathbf{OK}}(\mathbf{h},\mathbf{t}) = \sum_{\alpha=1}^{\mathbf{n}(\mathbf{h},\mathbf{t})} \lambda_\alpha(\mathbf{h},\mathbf{t})[\mathbf{Z}(\mathbf{h},\mathbf{t})_\alpha - \mathbf{m}(\mathbf{h},\mathbf{t})_\alpha] + \mathbf{m}(\mathbf{h},\mathbf{t}) \qquad \text{Eq. 7.24}$$

$$= \sum_{\alpha=1}^{n(\mathbf{h},t)} \lambda_\alpha(\mathbf{h},t)Z(\mathbf{h},t)_\alpha + \left[1 - \sum_{\alpha=1}^{n(\mathbf{h},t)} \lambda_\alpha(\mathbf{h},t)\right] m(\mathbf{h},t) \qquad \text{Eq. 7.25}$$

$$Z^*_{OK}(\mathbf{h},t) = \sum_{\alpha=1}^{n(\mathbf{h},t)} \lambda_\alpha^{OK}(\mathbf{h},t)Z(\mathbf{h},t)_\alpha \qquad \text{Eq. 7.26}$$

Like SK, the weights are determined so that the error variance is minimized. (Ainslie M. Denham, 2012).

7.5 Experiments and Results

The empirical variogram surface was calculated and utilized as input for the fitting of the different models. Figure 7.2 shows the empirical variogram as perspective wireframe with spatial bin set to 10 km apart.

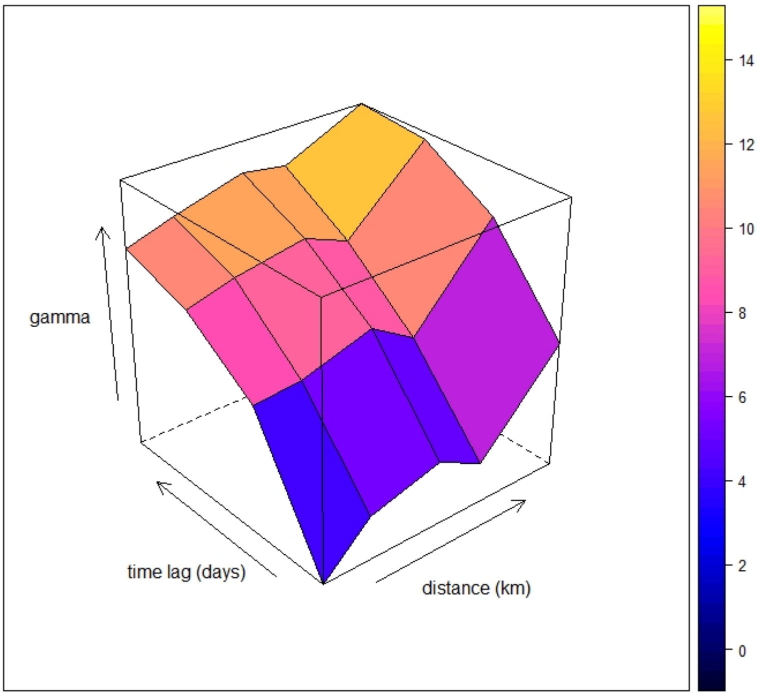


Figure 7.2: Empirical spatio-Temporal variogram, April 2012.

In addition to the selection of the S-T variogram, each component of this model was selected from two commonly applied one-dimensional variograms, namely Exponential (Exp) and Spherical (Sph) variograms. A selection of time specific fitted models (residuals compared to the sample variogram surface) are shown in Table 7.1. The same procedure was performed for each month of the year to find the best model for each variogram family. The result of the best fitting S-T model for the April 2012 time point shown in Figure 7.2 is highlighted in bold in Table 7.2.

Table 7.1: Weighted MSE (fit.method = 8) for different selections of the one-dimensional variogram components and different S-T variogram families.

	Year 2012- April				
	Joint	Exp+Exp	Exp+Sph	Sph+Exp	Sph+Sph
Metric	-	0.40	NA	NA	0.46
Product-sum (K=20)	-	0.83	1.15	15.56	15.36
Sum-metric	Exp	**0.034**	0.035	0.1074	0.1075
	Sph	0.04	0.041	15.56	0.1
Separable	-	1.21	1.26	15.79	15.78

Fitting routines these variogram models were implemented in R using the gstat library. To meet the criteria for some of the parameters, such as non-negative nuggets and positive ranges, the L-BFGS-B optimisation routine was used to enforce limits on the search space. In this study, the applied fitting routines were based on the (weighted) MSE of the model and sample variogram. Given the relatively small neighborhood size of the study area we needed to ensure that the model is fitted to the differences over the actual space and time that is used in the interpolation. Therefore, the spatial and temporal distances were reduced, and a cutoff measure was introduced. This adjustment also lowers the possibility of overfitting the variogram model. The S-T anisotropy was estimated in advance and fixed at 30 km/day. Different weighting schemes generate different model parameters and consequently result in different interpolation values. The best fitting S-T variogram of each family for April 2012 is shown in Figure 7.25.

Table 7.2: Best fitting S-T model for April 2012.

Metric model (weighted MSE: 0.40)					
	partial sill	model	range	nugget	anisotropy
Join	13.5711	(Exp, Exp)	43.62386		36.78 km/day
Product-sum model (weighted MSE: 0.72)					
	partial sill	model	range	nugget	k
Space	0.01	Exp	40.87	0.001	16.57
Time	21.47	Exp	6.47	0.95	
Sum-metric model (weighted MSE: 0.03)					
	partial sill	model	range	nugget	anisotropy
Space	5.77	Exp	56.62	0.10	
Time	16.55	Exp	17.39	0.03	
Join	8.46	Exp	68.32	0.02	68.43 km/day
Separable model (weighted MSE: 1.22)					
	partial sill	model	range	nugget	sp.-temp. sill
Space	0.994	Exp	2326.756	0.01	211.30
Time	0.99	Exp	63.78	0.009	

The month-specific Exponential model parameters were obtained in two stages: with and without the covariant as secondary information. The model parameters were estimated for each month of data at a time, to consider the possible device drift and seasonal variability in the regression coefficients and semi-variance parameters. The same approach was taken by Zoet, Osei, Hoek & Stein (2020) for S-T modeling of NO_2 concentration in city of Eindhoven, the Netherlands. Our results showed changes in the range of nugget value that indicate the degree of non-spatial variability of the observations. It can be concluded that adding the covariant to the model might improve the model by reducing the non-measurement error in the observed data. Table 7.3 shows the parameters after fitting the best model (sum-metric) with and without covariant for April 2012.

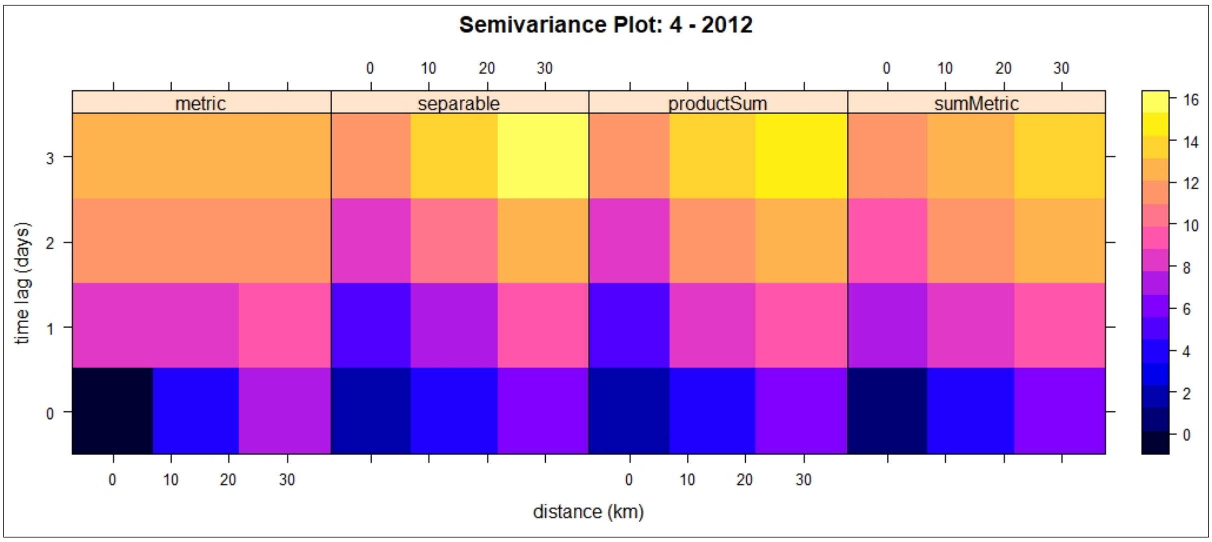


Figure 7.3: Best fitting S-T variogram of each family (April 2012).

The surface plot of predicted standard error is presented in Figure 7.4 showing a decrease in prediction standard errors while increasing the information in the model. Data for 15 April 12 (randomly selected) was also deliberately omitted from the original data set to investigate how predictions on a day with missing value are affected. It can be observed from the plots that prediction standard errors are significantly larger for 15 April 2012.

Table 7.3: Fitted Exponential Model parameters with and without covariant.

Sum-metric model without covariant (weighted MSE: **0.04**)					
	partial sill	model	range	nugget	anisotropy
Space	3.49	Exp	25.76	0.50	
Time	16.46	Exp	16.53	0.02	
Join	8.51	Exp	66.64	0.021	72.35 km/day
Sum-metric model with covariant (weighted MSE: **0.034**)					
	partial sill	model	range	nugget	anisotropy
Space	1.04	Exp	23.45	1.09	
Time	20.71	Exp	17.52	0.00	
Join	7.98	Exp	58.18	0.00	68.43 km/day

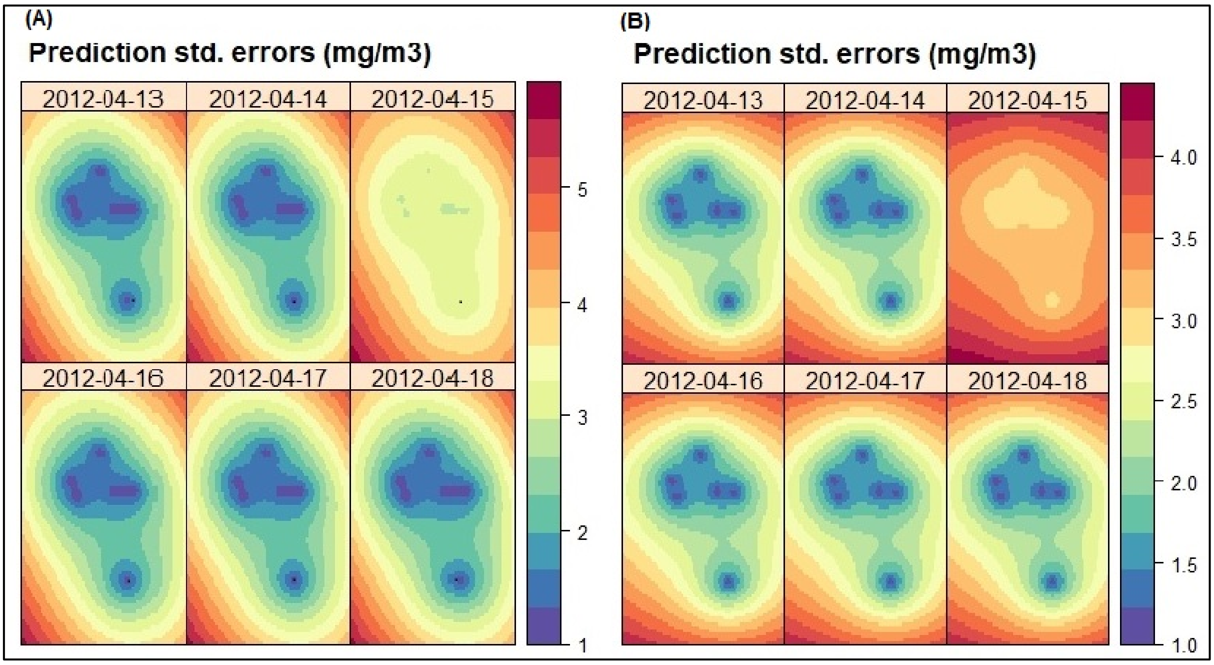


Figure 7.4: Spatio-temporal kriging prediction standard errors of PM_{10} with covariates (A) and without covariates (B) within the Auckland study area for six days in April 2012[2].

The month-year specific variogram models that obtained the smallest RMSE (Table 7.2) were then used to produce a gridded prediction using Universal S-T kriging where the latitude was used as a covariate. The interpolation domain consists of spatial locations between 174.42° and 175° west, and 37.3° and -36.7° north extended to 80 km apart. For the temporal grid, six equally spaced days were considered. Figure 7.5 shows the interpolated grid of prediction and prediction standard errors for six continuous days alongside the sampling locations for April 2012. The maps visualise prediction standard errors associated to the prediction location's proximity to an observation. It is also notable that the interpolated surface for the omitted day is much smoother than the observed days.

[2] Data for 15 April 2012 was excluded from the original data set (see discussion in page 229).

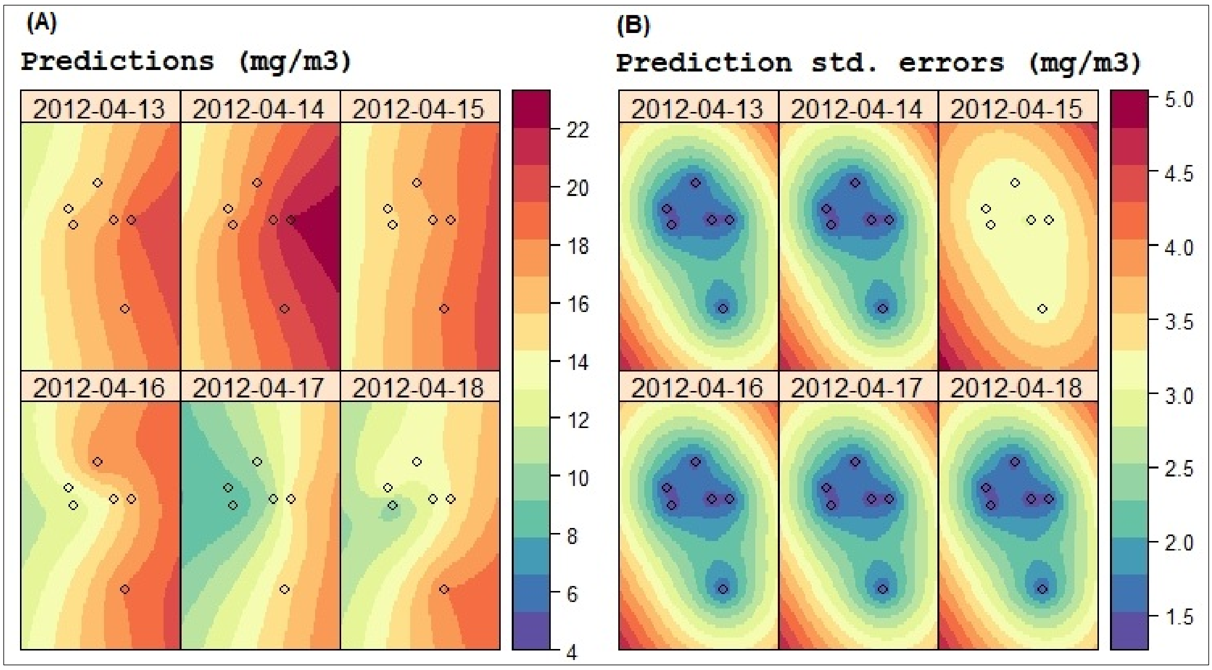


Figure 7.5: Spatio-temporal kriging prediction (A) and prediction standard errors (B) of PM_{10} over the Auckland study area for six days in April 2012.[3]

The predictive performance of the kriging procedure within the study area was performed using KrigeST function of gstat package. Similar to approaches taken in literature (Gräler et al., 2016; D. Hu et al., 2017; Van Zoest et al., 2020) for measuring the accuracy of predicted PM_{10}, using the KrigeST function and LOOCV , the accuracy of predicted PM_{10} at all observed space-time locations were measured. Actual and predicted values were then assessed to measure the model performance by calculating the errors of the interpolated values. The diagnostic measures employed were the RMSE, ME and MAPE of the residuals to the prediction errors. The value of MAPE should be 0 as the residuals from cross-validation should be equal to prediction errors at each point that was held out. Results of cross-validation on 2012 showed RMSE 0.692, average standard error 1.60, and the MSPE of the 10.42, than was higher than the ideal 0, meaning that the predictions are rather less variable than the true value; this is to be expected, as kriging is a smoothing estimator. The year-specific cross-validation metrics are presented in Table 7.4.

[3] Note: Data for 15 April 2012 was excluded from the original data set (see discussion in page 229).

Table 7.4: Year-specific LOOCV cross-validation results**.**

	RMSE	Average Standard Error	MSPE
2011	0.72	1.40	10.07
2012	0.69	1.60	10.44
2013	0.77	1.15	11.84
2014	0.91	1.45	12.02
2015	0.33	1.24	7.185
2016	1.22	1.47	10.14

Although spatio-temporal kriging modeling is generally accepted as an effective tool for modeling air pollutant concentrations (Van Zoest et al., 2020), the cross-validation results in Table 7.4 showed that spatio-temporal kriging only performed marginally better when compared to IDW for predicting Auckland PM_{10} concentration. The prediction quality of spatio-temporal kriging is known to improve if sufficiently strong correlated locations are added with the temporal dimension (Gräler et al., 2016). However, for Auckland given the monitoring networks limitations, with few stations at a low spatial density, including more stations is not currently possible.

7.6 Conclusion

In this Chapter we sought to describe the space time interactions of Auckland's PM_{10} through a marginal model developed by spatio-temporal covariances. The descriptive approach was chosen as the dynamic systems that drive the spatio-temporal PM_{10} concentration in the Auckland study area was not clearly understood or indeed even known.

In this Chapter, the dependent random process was identified in terms of first-order and second-order moments in terms of covariances of its marginal distribution. In the S-T kriging process for modeling and predicting Auckland PM_{10} concentration, it was assumed that the true process can be described in terms of S-T fixed effects combined with a S-T dependent random process. The S-T kriging in Section 7.4 aimed to specify the dependence structure in the PM_{10} random process aiming to achieve the second goal of S-T modeling as discussed in Section 7.1, that is, S-T prediction. The causal structure that causes the dependence in a random process was not of concern. Four models for spatio temporal variograms were examined in this Chapter. These stationarity assumptions were taken in determining the experimental marginal spatio-temporal variograms. The spatio-temporal variogram for the sum-metric was found to be the best fitted model compared to metric, product-sum and separable models based on the weighted MSE metric. The spatial and temporal dependencies were not only modelled independently, but also their joint dependencies. In our case of a pure nugget spatial variogram, these joint dependencies were stronger than the purely spatial dependencies.

The main challenge in characterising the spatio-temporal dependence structure using the marginal covariance model is its ability to model real-world spatio-temporal interactions according to the rules that govern the spatio-temporal variability. These rules make the underlying process a dynamical system. Chapter 8 uses machine learning and ensemble methods to account for these dynamics and therefore is attempted to provide more realistic and generalisable models.

Chapter 8 Machine Learning and Model Building

The overall goal of this Chapter is to describe work towards the development of a model, using data from routine monitoring networks, from which inferences can be drawn about site specific PM source characteristics and dispersion mechanisms. In that sense, it is important to utilise simple modeling techniques that depend only on data from routine and cost-effective monitors. Specifically, the non-linear statistical approaches of artificial neural networks (ANN), Long short-term memory (LSTM) and Random Forest (RF) were chosen and used for further evaluation. This Chapter is divided into two main sections: feature selection and forecasting.

Each of the subsections of this Chapter provides a theoretical background to these techniques with a summary of their current applications in air pollution modeling. These modeling techniques were then used to model the concentrations of PM_{10} pollutants collected from the monitoring sites in Auckland. The results are presented and discussed as a basis of the analyses presented in the subsequent sections.

8.1 Introduction

An approach in modeling of atmospheric pollutant concentration is to employ deterministic models that are based on the governing dynamic and chemical transformation procedures. Deterministic models are limited by their requirement for such detailed knowledge. In Chapter 5, purely statistical models were investigated with the aim of providing reasonable predictions using Auckland's routinely available data. However, these models were restricted by their incapability to provide insight into dispersion mechanisms and hence did not perform well. The challenge therefore remains. How do we create air quality models to effectively obtain the variability in observed concentrations using the available meteorological and temporal data?

The traditional methods presented in Chapters 5, 6 and 7 often require hand-crafted features, and expert knowledge of the field. As elaborated in Chapter 5, autoregressive models involve either strict or weak stationarity in the data, which does not hold in real world time series. This means that before we can apply traditional methods, we need to transform our data with techniques such as detrending algorithms, which introduce their own set of problems. It was concluded that the ARIMA class of models only perform well on linear processes. ANNs do not suffer from these problems since they are non-linear in nature and data driven. ANNs learn to model processes based on example input and output values, and performance gets better when more data is used in the training of the models (though one has to be careful of the risk of overfitting). This shifts the required expert knowledge, that was required with techniques such as ARIMA presented in Chapter 5, from the working field of the data to knowledge of the algorithm. This inherently makes it cheaper and easier to make meaningful predictions from time series data.

In this Chapter the use machine learning techniques, namely ANN Multilayer Perceptron (MLP), LSTM, and RF (an ensemble method) to forecast PM_{10} concentrations in Auckland study area is explored through experiments. These models can incorporate as many parameters as predictor variables. However, the complexity and multivariate behavior of environmental models require additional techniques to involve the appropriate variables within the modeling structure. Very few studies have given attention to understanding the meteorological variables that are the most important predictor variables for determining PM_{10} concentrations and identifying the influential time scales on the emission patterns; whether daily, weekly, or monthly, for example. Therefore, the site-specific most influential meteorological variables in determining PM_{10} concentrations and the important time scales which influence the emission patterns (daily or monthly) are identified using two common input optimization techniques, namely forward selection, and backward elimination. Inputs from these two methods are then used to build the models and to investigate to what extent these models can learn the time variation of concentrations through training without the need for comprehensive air pollution data. The quality of MLP output can be enhanced by de-correlating the input variables prior to the MLP training process stage (Yu et al., 2004). Intelligent hybrid systems are generally used for the finding the optimised ANN parameters and input selection. In this Chapter, a hybrid system is used through combinations of ANNs with the data mining technique, Principal Component Analysis (PCA), for eliminating correlation in the sample data before they are being presented to an MLP. Performance of these hybrid PCA-MLP time series modeling technique and to understand the complex time series of PM_{10} concentrations in a different site located in Auckland were analyzed.

Sections 8.2 to 8.4 describe the datamining and machine learning methods utilized in this Chapter. The similarities and differences of machine learning modeling approaches are discussed. Section 8.5 presents the experiment methodologies followed by section 8.6 presenting the results of the machine learning technique to understanding site-specific air pollution dispersion mechanisms.

8.2 Data Mining Techniques for Input Parameter Selection

8.2.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a multivariate statistical analysis method built on feature extraction aiming to find a new set of uncorrelated variables to explain the maximum variance. The PCA enables extraction of the main patterns in the matrix in terms of a complementary set of score and loading plots. In analysis of environmental data calculation of the principal components (PCs) is mostly performed using the solution of an eigenvalue problem through the matrix of covariances of anomalies of the dataset (Hannachi et al., 2007). The first PC accounts for the main proportion of variability, whilst the subsequent PCs explain the remaining variability which have not been explained by their predecessors (Sun & Sun, 2017). For every eigenvalue, there is an eigenvector that is not zero. Factor loadings provide an indication of the degree of correlation between a PC and a variable where sum of the squares of coefficients of correlation between a variable and all the components is equal to one (Abdi & Williams, 2010). Deciding on the number of retained principal components is a challenging factor as it may lead to information loss or overfitting. The two commonly used criteria in the literature are the Kaiser criteria and the Percentage of Accumulated Variance (PVA) criteria (Lau et al., 2009; Pearce, Beringer, Nicholls, Hyndman, & Tapper, 2011). In the case of the Kaiser criteria only the PCs with eigenvalues larger than one are retained (Jolliffe, 2011). A less restricted threshold was suggested by Lau et al. (2009) by retaining the PCs whose eigenvalues are equal to or greater than 0.7. The PVA_n criteria retains the PCs with accumulated percentages of variance exceeding the value n.

Forward Selection (Greedy Search): Forward selection, or greedy searching, is an incremental linear search approach that chooses input variables one by one (Olden et al., 2004). It starts by building single input networks and choosing the variable that maximizes the performance of model based on the selected optimality criteria (the lowest MSE in this case). The network iteratively trains by adding the remaining input to the previous input sets. The procedure continues until introducing another input variable does not enhance model performance.

Backward Elimination: In contrast with forward selection, backward elimination begins by training a network with all input variables and successfully eliminates inputs one at a time (Olden et al., 2004). The process is stopped when the removal of an input variable does not improve the model performance.

K-mean Clustering of PM_{10} Concentrations

K-means clustering is one of the most commonly used unsupervised machine learning algorithms and aims to split a dataset into k distinct groupings (Steinbach et al., 2000). Due to uneven distribution and lack of extremes in PM_{10} concentration, k-means clustering was used to ensure samples are evenly selected during partitioning data into train, test, and validation. The cluster rankings were used for partitioning the dataset for train, test, and evaluation. A k-means cluster analysis of the concentration is performed by randomly selecting k data points from the space of PM_{10} data that are being clustered into groups. The selected data points are taken as initial centroids. Each data point is then allocated to the cluster with the closest centroid. The position of the centroid is recalculated by finding the mean of the cluster once all points are allocated to a cluster. The last two step are repeated until the centroid no longer moves. K-means clustering require a pre-specified number of clusters, which in this Chapter's case study was set to the number of seasonal variations for each station identified in our primally data exploration presented Chapter 4. The clusters of PM_{10} concentration for each station are presented using a colour index plot (Appendix D).

8.3 Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANNs) learning capability, generalization, and extracting important features from inputs makes them an efficient modeling approach (Liu, 2001). Neural networks are comprised of one input layer with neuron(s), at least one hidden layer and one output layer that provides the output result(s). The connections between the neurons are weighted. Neural networks are trained by adjusting these connecting weightings so that the best relationship between the input and the output signal is established. The output of each neuron is produced from weighted inputs that are summed and passed through the transfer functions. Log-sigmoid is a commonly used transfer functions in Multilayer Perceptron (MLP) networks with backpropagation. The generated output by log-sigmoid function is within the range of zero to one (Dorofki et al., 2012). Learning in a neural network can be stated as search for weights for the network which can optimally model the data based upon certain criterion. This is usually referred to as a function of the training data that represents a gain to be maximized or loss to be minimized. Learning procedures in neural networks can be divided into three broad classes: Firstly, Supervised Learning in which the desired outputs of the input vectors are known throughout the learning procedure. Secondly, Unsupervised Learning in which the desired outputs of the input vectors are unknown and regularities within the input vectors are captured during the learning process. Thirdly, Reinforcement Learning which only requires a single scale evaluation of the generated output and if considered "good" then the network is "rewarded"; otherwise, the network is "penalized". Thus, reinforcement learning is sometimes referred to as reward-penalty or learning with critic. Depending on the neural network architecture and training method, ANNs can be divided into many sub categories (Haykin, 1999).

The architecture of an ANN defines the connection pattern and arrangement of neurons in relation to each other. Feedforward neural network (no feedback link) and feedback or Recurrent Neural Network (RNN) are the types of ANN architecture. One-layer perceptron, MLP, and Radial Basis Function (RBF) are types of ANNs using feedforward neural networks. The main networks using multiple-layer feedforward architectures are MLP and RBF with the generalized delta rule and the competitive/delta rule learning algorithms, respectively (Silva et al., 2017). The case studies in this Chapter address the MLP architecture with a hyper tangent sigmoid function. The hyper tangent sigmoid function is known to be a popular choice in multivariate functions approximation (Anastassiou, 2011; Reyes et al., 2013; Sramka et al., 2019) and PM_{10} modeling (Saufie et al., 2015; Valencia et al., 2006) as its learning rate is faster than sigmoid activation function in terms of number of training iterations.

The ANN modeling technique is known to be an effective nonlinear statistical modeling approach in air quality studies when compared with other approaches. ANN modeling was first used for the prediction of air pollutants in 1993. This was an ANN model for predicting SO_2 concentrations in an industrial area of Solvenia (Boznar et al., 1993).

An MLP neural network was used by Abdebrahim et al. (2016) to forecast the daily averaged concentration of the PM_{10} in Algiers (Algeria) using various meteorological parameters collected over four contiguous years. The authors compared the model's overall performance using different numbers of neurons and hidden layers. A recent study reviewed air pollution models and reported that ANN performed better than other statistical techniques in forecasting outdoor air pollutants (Cabanerosa et al., 2019). The authors selected 139 peer-reviewed articles (from January 2001 to February 2019) with 28% focusing on PM_{10} modeling and prediction. Few recent studies have investigated the relative contribution of meteorological parameters to the observed levels of air pollution concentrations (Singh et al. 2012; Yan Chan and Jian 2013), though some studies have used basic meteorological parameters and pollutants as predictor variables, without considering or justifying the parameter choices, to model the concentration of several pollutants in the model (Singh et al. 2012). The approach in general to inclusion of parameters in pollutant models appears to be opportunistic – if we have it use it – rather than considered.

A number of ANN models were used to analyze and predict PM_{10} concentration up to three days ahead using variety of pollutant measurements such as $PM_{2.5}$ and CO concentrations, O_3, NO, NO_2, SO_2, benzene as well as meteorological factors from Pescara, Italy (Biancofiore et al., 2017). The recurrent ANN (RC ANN) was reported to have performed the best when compared to a feed forward ANN and a MLR model. The authors noted that the inclusion of CO as a parameter in PM_{10} forecasting significantly improved the results in their models. This is not entirely surprising as typically CO is highly correlated with PM. Time-lagged models have also been found to give reliable predictions (Elangasinghe et al., 2014) but have limited practical use when missing data is presented in a dataset.

Perez (2001) used an ANN to predict hourly mean SO_2 concentrations eight hours ahead in Chile using hourly average temperature, RH, and wind speed. They reported an average error of 30%. In similar work, Chelani et al. (2002) used ANN to predict SO_2 concentrations at three sites in Delhi using wind speed, a wind direction index, RH, and temperature. Another study used a large number of predictor variables including date, maximum and average temperatures, pressure, humidity, wind, cloud coverage and daily precipitation as input to an ANN model. The aim was to forecast daily average total suspended particles for SO_2, PM_{10} and NO_2 (Jiang et al., 2004). The correlation between predicted and observed concentrations was found to be 0.7.

A PCA based ANN model was used by Karatzas and Kaltsatos (2007) to model hourly ozone concentration in Thessaloniki, Greece. The reported Index of Agreement (IA) for the two locations investigated in this study was reported 83 and 83.7 %.

The diverse studies noted in this section have used various input parameters to improve the forecasting accuracy. While the accuracy of ANN forecasts was found to be higher than traditional statistical forecasts in the literature reviewed here, ANN has its own shortcomings subject to their input variables. ANN models can show tendency to perform poorly and mislead due to noise in all the parameters from adding an excessive number of input parameters (M. A. Elangasinghe, 2014). These shortcomings of ANNs facilitated the development of the RNN models. As noted earlier, RNN has been shown by one 2017 study to outperform other forms of ANN when modeling $PM_{2.5}$ (Biancofiore et al., 2017).

8.4 Recurrent Neural Network (RNN): Long Short-Term Memory (LSTM) Models

Design of a Recurrent Neural Network (RNN) is similar to feedforward neural networks, where neurons' outputs are used as response inputs for other neurons. The recurrent component modifies these networks for active information processing and adaptive control. During back propagation, RNNs suffer from vanishing gradient-values which are used to update a neural network's weight. This may lead to increasing the learning time, or worse, result in the RNN not working (Akbari et al., 2014). Figure 8.1 illustrates the back propagation in a recursive module in a standard RNN with a single $tanh$ layer.

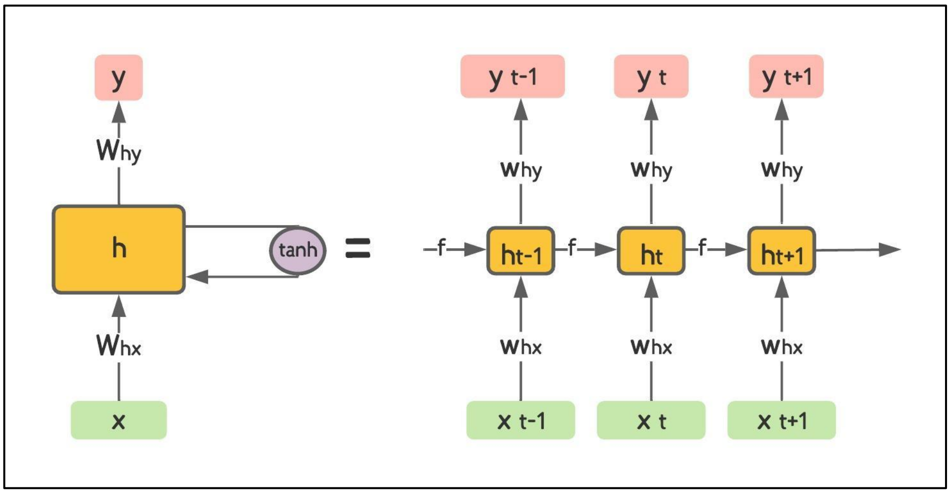


Figure 8.1: RNN with back propagation unfolded in time Adopted from (Colah, 2015) with permission.

The recursive RNN formulas are as follows (Pascanu et al., 2014):

$$h_t = tanh(W_h h_{t-1} + W_x x_t) \qquad \text{Eq 8.2}$$

$$y_t = W_h h_t \qquad \text{Eq 8.3}$$

where x_t , h_t, y_t, and W_h are input vector, hidden layer, output vector and weighted matrix, respectively.

Long Short-Term Memory (LSTM) networks (Hochreiter & Schmidhuber, 1997) are a modified version of RNN, that resolve the vanishing gradient problem by removing/adding information in a single cell through 'gates'. There are three gates namely input, forget, and output gates within an LSTM. LSTM are best suited for processing and predicting time series with time lags of unknown duration as well as deep neural network architectures composed of several LSTM layers and other types of layers (Rosato et al., 2019). Like standard RNN, LSTM trains the model by using back-propagation however, the repeating module in LSTM has different structure from RNN. Figure 8.2 shows an LSTM. The vector from each node's output is carried via the solid line to the inputs of others.

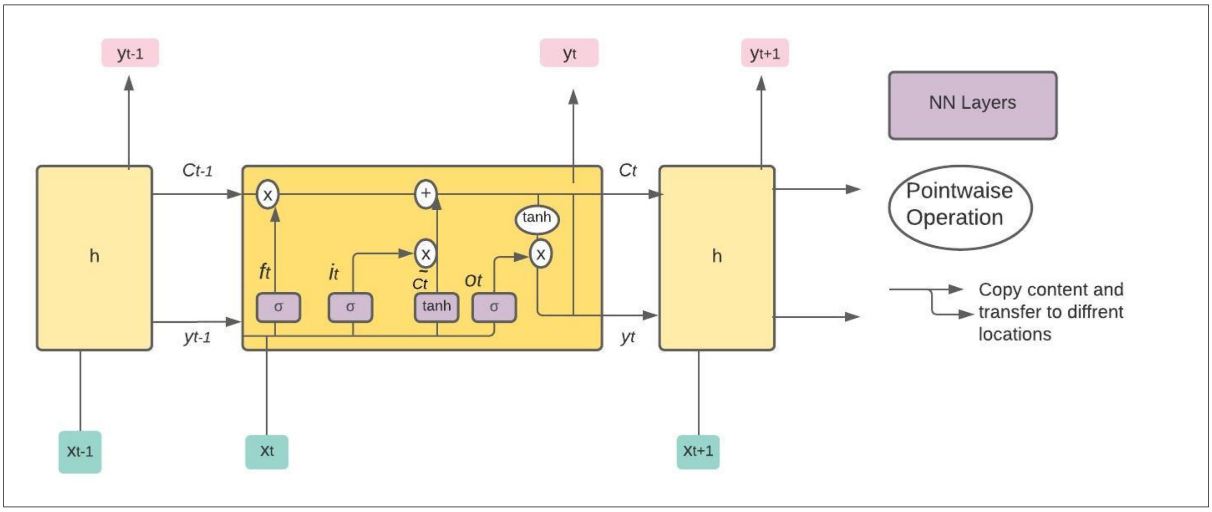


Figure 8.2: LSTM structure with four neural network layers. Adopted from (Colah, 2015) with permission.

The hidden states in an LSTM architecture is calculated as below (Graves et al., 2013):

$$\sigma = \frac{1}{1+e^{-1}} \qquad \text{Eq 8.4}$$

$$i_t = \sigma\left(W_f[y_{t-1}, x_t] + b_i\right) \qquad \text{Eq 8.5}$$

$$f_t = \sigma\left(W_f[y_{t-1}, x_t] + b_f\right) \qquad \text{Eq 8.6}$$

$$\tilde{c}_t = tanh\left(W_f[y_{t-1}, x_t] + b_c\right) \qquad \text{Eq 8.7}$$

$$c_t = f_t * c_{t-1} + i_t * \tilde{c}_t \qquad \text{Eq 8.8}$$

Where σ denotes the logistic sigmoid function. The input, forget, and output gates are presented by $i, f,$ and o, respectively. Information from cell to gate vectors are transformed through weight matrix (W). The new candidate value ($\tilde{c}_t$) is created by hyperbolic tangent function (*tanh)* layer. To calculate the output, a sigmoid layer is run to decide the part of the new cell state (c_t) to be used as output. In final step, *tanh* is applied to the cell state and multiply it by the sigmoid layer's output to calculate the output of a hidden state (y_t) (Graves et al., 2013):

$$o_t = \sigma(W_o[y_{t-1}, x_t] + b_0) \quad \text{Eq 8.9}$$

$$y_t = o_t * tanh(c_t) \quad \text{Eq 8.10}$$

Stacked LSTM (Graves et al., 2013), is an LSTM model with multiple LSTM layers where the output of the each neuron's hidden state in the first LSTM layer is used as an input to the next LSTM layer. In such structure, a sequence output is provided by the prior LSTM that is one output per input time step, instead of one output time step for all input time steps. The stacked LSTMs can significantly increase the network's generalization capability (Sagheer & Kotb, 2019).

Recently, the use of deep learning in urban air pollutant studies has become prevalent in interdisciplinary research (Ong et al., 2016; Qin et al., 2019). An LSTM model was developed by Kim et al. (2019) for daily prediction of $PM_{10/2.5}$ concentration. The model used 11 input parameters including observations for $PM_{10/2.5}$, various meteorological parameters as well as concentrations of several pollutants including SO_2, O3, NO_2 and CO. Measurements (January 2014 to April 2016) that were collected from seven monitoring sites located in the major cities of South Korea were used as a model training set. The model was optimized by providing a validation set and tested using a bench dataset used for evaluating models in South Korea. According to the authors, the model's structure was limited to three to five hidden layers with 100 hidden nodes. The authors reported an IOA, between 62% and 79%, and relatively high prediction errors and biases which were attributed to the high PM_{10} events which caused notable spikes in the concentration. The authors suggested that adding more informative input parameters would improve the model performance. A study by Wu et al. (2020) used multivariant LSTM model to predict PM_{10} concentrations using 12 months of PM_{10}, AOD, rainfall, evaporation capacity, relative humidity, sunshine intensity, wind velocity, SO_2, CO, and O_3 values collected during 2017 for Wuhan, China. The predicted value by LSTM for the maximum and minimum values were below average. The proposed multivariant LSTM algorithm was not optimized and that could be the cause of the reported underprediction (below average) of maximum and minimum values. In another recent study, PM_{10} concentrations in the Upper Hunter valley, Australia was forecasted using LSTM for October 2019 using the PM_{10} (only) values (from 30 September 2012 to 30 September 2019) as the network input (Delgado et al., 2020). To detect PM_{10} they used 60-day lags with 130 iterations. The graphical results in their paper showed that the model detected a 'quite close to actual behavior' trend of PM_{10} for the next month. No numerical measure of accuracy or closeness was reported.

LSTM has also been applied by Xayasouk et al. (2020) to forecast PM_{10} concentrations ten days ahead. The input data was collected from 25 monitoring stations in Seoul, South Korea reporting $PM_{2.5}$, meteorological measurements (rainfall, wind speed and direction, temperature, and relative humidity), and sky condition. The network was optimised by obtaining the lowest RMSE value while using different batch sizes and keeping the learning rate and epoch constant at 0.01 and 100, respectively. Communication with the first author of the paper clarified that the average of hourly data from 25 stations were used as input data – this information was missing from the paper.

LSTM have also been used for shorter term prediction (12, 24, 48, and 120 hours) of PM_{10} measurements (Becerra-Rico et al., 2020). The input datasets contained 12 months (for 2016) of hourly SO_2, CO, NO_2, ozone, PM_{10} and $PM_{2.5}$ measurements obtained from few monitoring stations located in Mexico City. The model was trained using 3000 epochs based on the asymptote given around 3000 epochs and 4000 batch sizes. Their results showed the lowest RMSE was obtained for 12 hours ahead prediction. The second lowest RMSE belonged to the 120 hours forecast while the highest RMSE was for the 24-hour forecast. The authors concluded that the inconsistency between the prediction results needed to be investigated further.

8.5 Random Forest (RF)

Random Forest (RF) is an ensemble learning method that trains several decision trees and constructs an ensemble of them. The RF is random as bootstrap samples of the data are used to build each tree, and the nodes are divided based on the best subsets of randomly selected features. The average of each tree's output is used to get a final RF ensemble prediction once the training is finalized (L. Breiman, 2001). The ensemble method enables the time of training to be consistently tuned. Variable importance measures provided by RF determines the prediction power of each variable, thus producing more interpretable results than neural networks (X. Hu et al., 2017).

RF with ensemble decision tree ML method was used by Grange et al. (2018) to perform daily PM_{10} trend analysis using meteorological variables (wind speed, wind direction, and atmospheric temperature), daily boundary layer heights, synoptic weather pattern, air mass cluster and seasonal terms collected from 31 stations across Switzerland. The R^2 value for their predictive RF models ranged from 54% to 71%. The application of the RF model for daily 24h averaged $PM_{2.5}$ concentration using meteorological measurements (temperature, dew point, visibility, pressure, potential evaporation, downward longwave radiation flux, downward shortwave radiation flux, RH, and wind vectors), AOD, land use and the GEOS-Chem model (GEOS-Chem, 2017) has been evaluated by Hu et al., (2017). Their results achieved an overall CV R^2 value of 0.80 using a convolutional layer. AOD, land use and meteorological variables were used as predictor variables.

A RF was able to explain 78% of daily PM_{10} variation using AOD, land use information, weather data, and MODIS active fire data in China (Chen et al., 2018). An RF model was reported to have captured PM_{10} variability in Italy, with a CV R^2 of 0.75 (Stafoggia et al., 2019). The study used a rich set of predictors namely meteorological data, AOD, Planetary boundary layer (PBL), monthly estimates of MODIS Normalized Difference Vegetation Index (NDVI) at 1-km^2 spatial, desert dust advection days, emission data (SO_2, NO_2, CO and NH_3), road density, elevation, satellite-based nighttime imagery, land based data, land cover, imperviousness surface areas, population, geo-climatic zones, and administrative areas (Regions, Provinces, Municipalities). The results showed temperature, PBL, wind components, AOD and Julian day as well as elevation, spatial coordinates, and administrative regions were the most important predictors. A RF model was reported that described around 64% of PM_{10} variability (Sweden, during 2005–2016) in test set using meteorological, AOD and 16 spatial (including residential population, road lengths, % green space, % residential space, and % forest space) predictors (Stafoggia et al., 2020). To the author's knowledge RF has not been used for modelling Auckland PM_{10} data to this date. In this Chapter, Auckland PM_{10} concentration is modeled using RF method and its performance is evaluated and compared to those which used a rather rich set of input variables.

8.6 Experimental Methods

PCA: To perform site specific PCA, all predictive variables were first included in the analysis. KMO (Kaiser, 1970) and Bartlett's test of sphericity (Tobias & Carlson, 1969) were used to test whether the data were fit for PCA. Extraction analysis was performed based on correlation matrix to extract based on eigenvalues greater than one. To find the number of PCs the Kaiser rule and PVA_{70}, as suggested by Lau et al. (2009), as well as PCA scree plots are used in this study as an indication to select the number of PCs to keep for each site.

MLP: A multilayer perceptron ANN topology was selected to build a functional relationship between PM_{10} concentration and other inputs as baseline. The month and DOW variables were introduced to the model in their respective discrete numbers (months 1-12 and weekday 1-7) as temporal inputs. Model predictions are site-specific and, once trained to a particular site, can confidently be used to predict concentrations only at that site. The time series of data is first randomized to ensure that a normally distributed sample of data covering all seasons is drawn for training. The data set was then divided into a training set (70% of the data), a validation set (15%) and a test set (15%) using the cluster numbers as partitioning value. The best-found multilayer perceptron model consisted of one hidden layer network with a Levenburg Marquardt back propagation algorithm, hyperbolic tangent transfer function in the hidden layer and bias transfer function in the output layer. To test if the ANN model was capturing any non-linearity that was not picked up by MLP model, a site-specific ANN-PCA hybrid model was developed using the PCs as inputs. The ANN model performance was then compared against the results of ANN-PCA hybrid model.

LSTM: The LSTM experiments were programmed in Python's SciPy environment using the Keras package (Falbel et al., 2020). Prior to fitting the LSTM model to the Auckland PM_{10} dataset, the time series were transformed into a supervised learning problem. Data was organised into input and output patterns so that the lagged observation is used as an input to forecast the observation at the current time step. All observations were then scaled between -1 and 1 to meet the requirements of the models' default hyperbolic tangent activation function. These transforms are then inverted on prediction values to rescale them into the original scale prior to calculating an error score to evaluate the model. To avoid bias with knowledge from the test set the scaling coefficients (RMSE) were calculated on the training set and were used to scale the test dataset. In this study, Auckland PM_{10} concentration is analyesd using an n=1 time-step of multiple (PM_{10} and meteorological) features. Hyper tangent activation function is used to activate output layer because it has a steady state at 0 (Witten et al., 2017); also sigmoid function suffers from vanishing gradient problem due to continuous multiplication of gradients (Witten et al., 2017).

8.6.1 Model Configuration and Tuning of the LSTM Parameters

LSTM was set to LSTM 'stateful' to give it control when the LSTM layer state is cleared. A batch size holding a fixed-sized number of train dataset was used to define the number of processing patterns prior to updating the network's weights.

The structure of the $LSTM_{multi}$ was determined from the iterative training multiple repeats using walk-forward validation in a loop of 30 repeats. A cost function was used to measure model performance based on training samples and the related prediction outputs. To meet the purpose of minimizing the regression cost, the RMSE was used as a cost function. RMSE is commonly chosen in literature as it retrains the model to have small errors at each point (Lin & Huang, 2020) , (Yadav et al., 2020). The RMSE score for each iteration was calculated and their distribution was analysed to find the optimal configuration. Adaptive Moment Estimation (ADAM) was applied to train the neural networks (Kingma & Ba, 2015). The ADAM approach combines the advantages of Adaptive Gradient Algorithm (Liu et al., 2020), and the Root Mean Square Propagation (e.g. how quickly it is changing). To obtain a fair trade off between precision and generalization the LSTM model parameters were tuned in following three stages. Given its obvious advantages, in this research ADAM method is therefore adopted.

Number of Epochs- The initial $LSTM_{multi}$ model was constructed based on two batch sizes (number of training examples in one forward/backward pass of an RNN before a weight update) and one single neuron. Thirty runs were completed for epoch values of 50, 100, 120, and 180 and RMSE summary statistics were used to find the best epoch configuration ($Epoch_{Optim}$). The metrics in summary statistics includes the mean (the average expected performance of a configuration), standard deviation (SD) (the variance), and min and max RMSE scores (the range of possible best and worst-case examples).

Batch size- Once the $\text{Epoch}_{\text{Optim}}$ is selected, the impact of different number of batch sizes in the network is investigated while keeping the number of epochs constant at $\text{Epoch}_{\text{Optim}}$. The batch size must be a factor of the test and training sets to meet the Keras requirement (Falbel et al., 2020). The average test RMSE performance was utilized to find the optimal batch size number ($\text{Batch}_{\text{Optim}}$).

Number of Neurons- The impact of differing the number of neurons in the network was investigated while keeping the $\text{Epoch}_{\text{Optim}}$ and $\text{Batch}_{\text{Optim}}$ constant. The number of neurons influences the learning capability of the network. Usually, increasing the number of neurons increase the learning of structure from the problem at the cost of the rise in training time and potentially the training data overfitting. The effect of rising the number of neurons, in the range of 1 to 5, was empirically evaluated. The average test RMSE performance was used to find the optimal number of neurons whilst other network configurations are kept unchanged.

Random Forest: According to Breiman (2001) the use of bagging appears to improve accuracy when used in tandem with random features and provides continuing estimates of the correlation and generalization error of the combined ensemble of trees out-of-bag (OOB). In the case studies presented in 8.7.5, two different approaches were used to assess the performance of the site-specific RF models through cross-validation. Firstly, predictions and observations from the OOB data of the RF were compared. Bagging is used in conjunction with random feature selection at each node. The “in-bag” samples held two thirds of the observations to train the model. Each new training set is drawn, with replacement, from the original training set. The RF implementation uses CART to grow the trees. Because RF is an ensemble method trees are grown on the new training set without pruning. The remaining third OOB is used for model validation. Second approach was taken by using random 80% of the data for training set. The test set of 20% was then put down each forest to get a test set error for both. Random forest was run using 100 iterations and the size of randomly selected attributes of the group was set to the first integer less than $\log_2 m+1$, where m is the number of predictors. The Random Forest classifier of Weka which implements Breiman’s RF algorithm (L Breiman, 2001) was used to build the RF model.

8.7 Results and Discussion

8.7.1 Input selection Results and Discusion

The descriptive analysis provided in Chapter 3 formed the basis for using input parameters in the MLP model. To make good model predictions, the emission patterns included in the model by including a numeric value for ‘month of the year’ (1 [January] -12 [December]) and DOW (1 [Sunday]to 7[Saturday]) as inputs, along with other meteorological input parameters, including wind speed, wind direction, temperature, relative humidity, rainfall, and solar radiation. Table 8.1 provides the results of site-specific model input feature selection using forward selection and backward elimination methods. Different inputs were selected when forward selection and backward elimination were applied to Henderson, Pakuranga, and Takapuna and Patumahoe. In the case of Patumahoe backward elimination selected all the predictor variables as important in making the best PM_{10} prediction

Table 8.1: Site-specific forward and backward optimization technique.

Site	**Optimization Technique**	
	Forward	Backward
Glen Eden	Lag_1, Temp, Rain, DOW	Lag_1, Lag_2, Temp, WD, Solar, DOW
Henderson	Lag_1, RH, Temp, Month, WS, Rain	Month, DOW, Solar, WD, WS, Lag_2, Rain, Lag_1, Temp
Pakuranga	Lag_1, WD	DOW, Lag_2, Rain, WD, Lag_1, RH
Patumahoe	Lag_1, WD, RH, Temp, Rain, WS, Month	Month, DOW, Solar, WD, WS, Rain, Lag_1, RH, Temp
Penrose	Lag_1, RH, Temp, WS, Month, Rain, WD, Solar	Month, Solar, WD, WS, Rain, Lag_1, Temp, RH
Takapuna	Lag_1, WD, WS, RH, Temp, month, Solar, Lag_2	Month, Solar, WD, Lag_2, WS, Lag_1, Temp

8.7.2 PCA Method Reuslts and Discussion

The initial site-specific PCA gave a KMO value of ≥ 0.5 at the $p < 0.0001$ significance level for all sites, suggesting the sample is adequate for PCA (Kaiser, 1970) The initial PCA for Glen Eden (PCA_{GE}) showed that there are four distinct constructs explaining 64.75% of the total variance for eigenvalues of greater than one (Table 8.2). The same conclusion was drawn using the scree plot as at fourth PC the plot starts to taper off gradually (Figure 8.3).

Table 8.2: Eigenvalues results, Glen Eden

	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	**2.391**	23.907	**23.907**
2	**1.601**	16.013	**39.921**
3	**1.374**	13.740	**53.661**
4	**1.109**	11.085	**64.746**
5	0.999	9.9920	74.738
6	0.775	7.7500	82.489
7	0.570	5.7030	88.192
8	0.536	5.3560	93.548
9	0.434	4.3410	97.889
10	0.211	2.1110	100.000

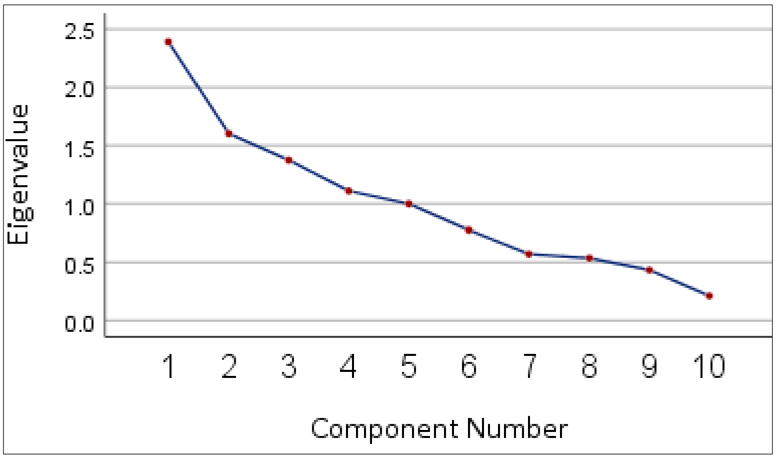


Figure 8.3: Scree Plot of PCs Eigenvalue, Glen Eden

The four constructs identified by PCA_{GE} are presented in rotated component matrix sorted by size in Table 8.3. For the first component, explaining ~ 24% of the variance, three items RH, solar radiation and rain have high communalities (> 0.7). Any items with communalities less than 0.3 are considered too low to load on the factor (Osborne, 2008). For component 2 only lag_1 and lag_2 have a high loading with temperature having a weak loading just exceeding the communality cut-off. Because this component is hypothetically related the previous two day's PM_{10} lag, temperature is not considered to be part of the construct and a communality value cut-off of 0.5 is employed. The third component, explaining ~ 14% of the variance, has only wind speed and direction loading on the factor. While component four which represents seasonality showing strong loading between month and temperature and their effect on PM_{10} concentration. It is notable that DOW does not load on any of all four components.

Table 8.3: Rotated Component Matrix, Glen Eden

	Component			
Item	**1**	**2**	**3**	**4**
RH	**.830**	.112	-.199	.003
Solar	**-.791**	-.295	.099	-.167
Rain	**.717**	-.126	.297	-.119
Lag_1	.054	**.852**	.018	.065
Lag_2	.076	**.839**	.107	.017
Dow	-.020	-.056	.008	.052
WS	-.025	-.109	**.867**	.011
WD	-.021	.209	**.705**	.125
Month	-.029	-.134	.214	**.859**
Temp	-.195	-.413	.126	**-.735**

In next step, the DOW item, with low factor loadings in all components, was removed and the experiment is repeated to investigate its impact on the component structures within the data. It is worth noting that KMO value improved from 0.5 to 0.64 ($p < 0.0001$). Table 8.4 shows that the number of distinct components remains the same and the total variance explained is increased to 71.92%.

Table 8.4: Rotation Sums of Squared Loadings

Component	Total	% of Variance	Cumulative %
1	1.881	20.895	20.895
2	1.775	19.727	40.623
3	1.463	16.259	56.882
4	1.354	15.042	71.924

The rotated component matrix (Table 8.5) showed almost the same loading pattern for each component. The same steps were taken for each site and the final site-specific PCs were generated (Table 8.5).

Table 8.5: Site-specific rotated components using Varimax rotation with Kaiser Normalization.

Site	**Rotated Components**				
	1	**2**	**3**	**4**	**5**
Glen Eden	0.83RH-0.79Solar+0.72Rain	0.85Lag$_1$+0.84Lag$_2$	0.87WS+0.70WD	0.86Month-0.74Temp	-
Henderson	-0.85Solar+0.84RH+0.6Rain	0.86Lag$_1$-+86Lag$_2$	-0.75Temp+0.76Month+0.51WD	0.85WS	0.98DOW
Pakuranga	0.92Rain+0.67RH-0.53Solar	0.86Temp+0.73Solar	0.89Lag$_1$+0.88Lag$_2$	0.86WS+0.74WD	-
Patumahoe	0.88Lag$_1$+0.83Lag$_2$	0.82Solar-0.82RH	0.97WS	-	-
Penrose	-0.84RH+0.83Solar-0.66Rain	0.84Lag$_1$+0.84Lag$_2$	-0.78Temp+0.74Month+0.5WD	0.53Rain+0.84WS	-
Takapuna	0.87RH-0.85Solar-0.50Rain	-0.81Temp+0.74Month+0.57WD	0.84Lag$_1$+0.83Lag$_2$	0.82WS+0.56Rain	-

The site-specific PCs are presented in Table 8.5, showing the variables with the highest loading (cut off > 0.5) in each component identified by PCA. Henderson and Patumahoe had the highest (5) and lowest (3) number of PCs. number of PCs (5 PCs) It is notable that relative humidity, solar radiation, and rain were selected by first PCs in all sites except for Patumahoe with rural background where lag_1 and Lag_2 were selected in its first PC. Both lag_1 and lag_2 were selected in second PCs for Glen Eden, Henderson, and Penrose. The wind components were included in the 3rd PC for Glen Eden, Henderson, Patumahoe, and Penrose. The temporal variable, month, was only included in Henderson, Penrose, and Takapuna. DOW was selected in Henderson's fifth PC only.

8.7.3 MLP Results and Discussion

The ANN model gave the best predictions for PM_{10} at the Pakuranga site for backward and PCA input selection methods. The PM_{10} concentrations at this site are mainly affected by emissions from traffic as well as nightly home heating during the winter. It is notable that, given the presence of mixed effects from these two sources, the relationship between PM_{10} concentrations and meteorology is not detected by the ANN model at the Henderson site which is also a similar urban environment to Pakuranga site. However, a similar relationship was for both Henderson and Takapuna: these two sites also showed similar variation in PM_{10} concentration (Chapter 3). Overall, the results of the time series predictions of PM_{10} for Auckland using MLP showed a moderate level of accuracy for all the site's models. These findings agree with Elanasinghe's MLP model (2014) to predict hourly PM_{10} concentration which reportedly gave a RMSE of 8.2 µg/m^3 and an R^2 of 0.66.

Table 8.6: Performance of MLP model (forward selected inputs)

	Train		**Validation**		**Test**		
	RMSE	R^2	RMSE	R^2	RMSE	R^2	IOA
Glen Eden	4.50	0.61	4.82	0.46	4.45	0.62	0.63
Henderson	3.58	0.57	3.88	0.54	3.97	0.59	0.59
Pakuranga	3.85	0.58	3.94	0.57	4.09	0.61	0.65
Patumahoe	4.91	0.57	5.67	0.52	5.06	**0.63**	**0.66**
Penrose	4.72	0.53	4.77	0.50	4.50	**0.53**	**0.54**
Takapuna	4.30	0.54	3.99	0.53	4.29	0.56	0.58

Table 8.7: Performance of MLP model (backward selected inputs)

	Train		**Validation**		**Test**	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
Glen Eden	4.36	0.64	4.38	0.62	4.72	**0.52**
Henderson	3.57	0.57	3.92	0.58	3.59	0.58
Pakuranga	3.90	0.60	3.99	0.61	3.92	**0.64**
Patumahoe	4.88	0.59	5.20	0.50	5.37	0.63
Penrose	4.71	0.53	4.70	0.51	4.51	0.53
Takapuna	4.30	0.53	4.05	0.57	4.07	0.60

Table 8.8: Performance of MLP model (PCA inputs)

	Train		**Validation**		**Test**	
	RMSE	R^2	RMSE	R^2	RMSE	R^2
Glen Eden	4.21	0.66	4.36	0.59	4.90	0.60
Henderson	3.81	0.53	3.83	0.49	4.08	0.55
Pakuranga	4.96	0.59	5.21	0.43	4.82	**0.65**
Patumahoe	3.76	0.58	3.89	0.59	4.54	0.62
Penrose	4.69	0.51	4.72	0.48	4.53	0.52
Takapuna	4.28	0.53	4.28	0.47	4.23	0.56

The modest prediction results shown in Table 8.6 to Table 8.8 achieved for PM_{10} could be because of lack of data related to additional gaseous co-pollutants to represent the contributions of home heating and traffic, as well as the effects of removal mechanism such as rainfall washout, re-suspension and depositing on surface. These factors are also suggestive of dispersion complexity that cannot be explained solely by meteorological and temporal parameters making modeling of PM_{10} a challenging task.

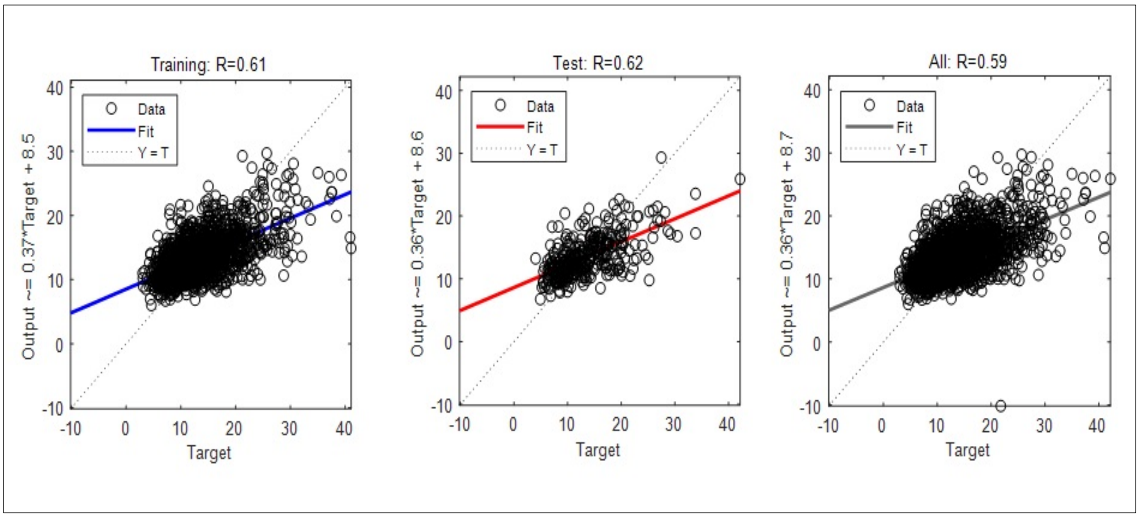


Figure 8.4: Daily average observed versus predicted PM_{10} concentrations using MLP with forward input selection (Glen Eden).

The impact of the three input selection methods on performance of MLP models are slightly different on different sites with MLP-PCA performing best on Pakuranga. Only insignificant differences in performance statistics of the models were found between forward selection and backward elimination techniques for Takapuna site with R^2= 0.56 and RMSE=4.29 µg/m^3; and R^2=0.60 and RMSE= 4.07 µg/m^3, respectively. The MLP$_{fw}$ and MLP$_{bck}$ performed better in Glen Eden with 0.23 µg/m^3 decrease in RMSE and 0.10% increase in R^2. It is notable that DOW was selected in the backward elimination method for Pakuranga. This led to the expectation that the model would serve well for such site because it is influenced by local traffic sources like the case raised by Carslaw et al.'s model (2006). However, there was only a marginal improvement in the model using the MLP$_{bck}$ features (a 0.17 decrease in RMSE µg/m^3 and a 0.02 increase in R^2) when compared to MLP$_{fw}$. This lack of notable improvement when including DOW in the model, could be due to introducing noise as the number of inputs increased. Based on these results and considering the impact of size of dataset on LSTM models to be implemented in this Chapter, it was decided to choose inputs selected by the forward selection method.

8.7.4 LSTM Results and Discussion

Number of Epochs: Table 8.9 shows the RMSE summary statistics from each population of results for the Pakuranga Site. The mean RMSE scores suggest that epoch configured to 100 is better than the other alternatives. The best possible performance might accomplish by using epochs of 180, but this improvement is at the expense of poor performance on average (the mean RMSE almost doubles).

Table 8.9: Summary statistics for different four epochs.

	RMSE			
	Number of Epochs			
	50	**100**	**120**	**180**
Count	30	30	30	30
Mean	5.89	5.83	5.86	10.20
SD	0.53	0.19	0.50	24.35
Min	5.22	5.25	5.21	5.17
25%	5.83	5.83	5.83	5.83
50%	5.83	5.83	5.83	5.83
75%	5.83	5.83	5.83	5.83
Max	8.62	6.58	8.35	139.11

Batch Size: The results of mean performance (Table 8.10) suggest a lower RMSE is achieved with a batch size of one. Depending on the site-specific results, this might improve further by increasing training of the epochs. For Pakuranga station (Table 8.9), a batch size of one is an ideal result, this size affording a low mean error (8.94) with low variability (0.54) indicating that the tuned network is reproducible.

Table 8.10: Summary statistics for each of the three batch size configurations.

	RMSE		
	Batch Size		
	1	**2**	**4**
Count	30	30	30
Mean	8.94	11.22	12.46
SD	.54	1.462	1.04
Min	7.23	9.67	11.58
25%	7.27	9.76	11.82
50%	10.02	11.39	11.90
75%	10.09	12.64	12.92
Max	10.09	12.64	14.09

Number of Neurons: From the mean performance alone, see Table 8.11, the experimental results suggest using three neurons has the best performance with lowest variance over 100 epochs and batch size of one.

Table 8.11: Summary statistics for each of the four neurons.

	RMSE		
	Number of Neurons		
	1	**3**	**4**
Count	30	30	30
Mean	1.24	0.79	1.49
SD	-0.71	-0.89	-0.79
Min	0.93	0.45	1.32
25%	0.95	0.45	1.36
50%	1.28	1.00	1.38
75%	1.53	1.02	1.58
Max	1.53	1.02	1.82

The same diagnostic steps were taken for the remaining five sites to objectively compare the impact of different number of epochs, batch size and neurons on the tuning of the site-specific LSTM networks.

Table 8.12: Statistical analysis with modelled and observed PM_{10}

	MAPE	RMSE	R^2
Glen Eden	35.28	5.21	0.45
Henderson	22.37	3.77	0.59
Pakuranga	30.84	5.30	0.56
Patumahoe	29.42	3.90	0.56
Penrose	24.78	4.56	0.52
Takapuna	24.03	5.06	0.48

In final step, a many-to-one $LSTM_{multi}$ model is developed where input time-steps have multiple (m) features, based on the input selection method presented in section 8.7.1, with $n = 1$ time-steps. The framed dataset in $LSTM_{multi}$ contains $n \times m + m$ columns where $n * m$ columns are taken as input for the observation of all features across the previous n days. The input data is then reshaped to a three-dimensional structure (samples, timesteps, and features) to reflect the actual timesteps and features. The PM_{10} data is taken as output at the following day.

The results from these experiments are tabulated in Table 8.12. The RMSE between the LSTM predictions and observations ranged from 3.77 to 5.21 $\mu g\, m^{-3}$. Among the six sites, the PM_{10} predictions at Henderson showed the lowest error (3.77 $\mu g/m^3$). The highest RMSE belongs to Glen Eden site (5.21 $\mu g/m^3$).

8.7.5 RF Results and Discussion

Site specific RF models for prediction of PM_{10} concentration were grown using the same explanatory variables as for LSTM. Figure 8.5 shows the site-specific explanatory variable importance for PM_{10} concentrations. Generally, lag_1 had greatest importance for prediction of PM_{10} concentrations. DOW contributed to the models' predictive ability despite being, in general, the least important of the variables in the RF models. Amongst the meteorological variables, wind direction was the most important variable for Penrose and Patumahoe. This observation agrees with the Explanatory analysis performed as part of this research and is reported in Chapter 4 of this thesis. The performance of an RF model, unlike ANNs and LTSMs, is not negatively affected by the inclusion of attributes that have lower predictive power (Grange et al., 2018) and therefore these variables were not removed from the RF models developed as part of this research.

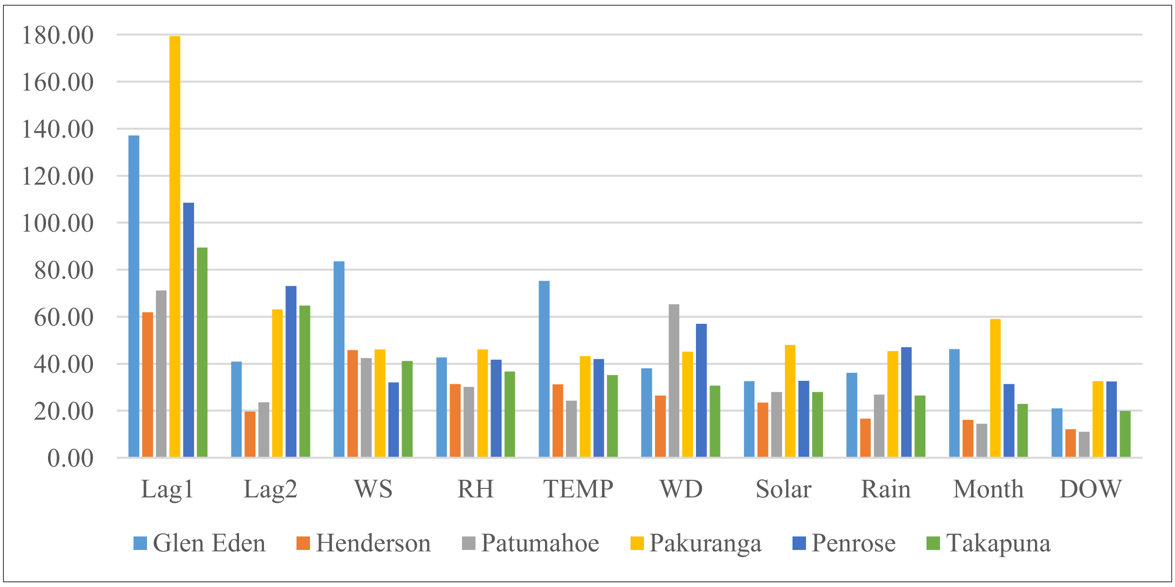


Figure 8.5: Importance of parameters based on average impurity decrease and number of nodes using that parameter. Y axis represents the importance.

The comparisons of both validation approaches are summarized in Table 8.13 in terms of R^2, RMSE, as well as IOA, obtained from linear regression between observations and CV predictions. The R-squared values for OOB and test set ranged from 0.56 to 0.71 and 0.52 to 0.71, respectively.

The results show that for most sites in Auckland, between 56%-65% of PM_{10} concentration were explained using simple meteorological and temporal variables. Unlike when using random left out samples as train-test sets where bias is presented at unknown extend, the OOB estimates are unbiased as it runs past the point where the test set error converges (L. Breiman, 2001). These experimental results, shown in Table 8.13, also provide empirical evidence, over and above that of Breiman, that the OOB estimate is as accurate as using a test set if not better concluding that using the OOB error estimate removes the need a test set.

Table 8.13: RF Statistical Metrics on OOB and Test Set.

	OOB					**Test Set**				
	R^2	RMSE	RAE	RRSE	IOA	R^2	RMSE	RAE	RRSE	IOA
Glen Eden	**0.71**	3.03	0.70	0.70	0.79	**0.71**	3.00	0.71	0.71	0.79
Henderson	0.63	3.47	0.76	0.78	0.69	0.62	2.77	0.78	0.79	0.70
Patumahoe	0.66	3.69	0.75	2.85	0.72	0.66	3.82	0.75	0.76	0.73
Pakuranga	**0.56**	5.05	0.83	0.83	0.69	**0.52**	4.92	0.88	0.86	0.69
Penrose	0.64	4.18	0.77	0.77	0.72	0.66	4.38	0.76	0.76	0.72
Takapuna	0.63	3.92	0.79	0.78	0.73	0.64	3.88	0.78	0.77	0.73

Performance of the site-specific MLP models varied with respect to their generalization capability, which in turn was dependent upon the data representation. The MLP models explained 53 to 63 % of PM_{10} concentration variability in selected sites. In comparison to related work (see section 8.5) the proposed model performed relatively poorly. We believe that this poor performance may be due to nature of Auckland PM_{10} concentration distribution as well as a limitation in the variety of predictors available for Auckland. The most relatable MLP model was developed by Elangasinghe (2014) to predict hourly PM_{10} concentration in a traffic/residential background site (Mt Wellington Highway Interchange on SH 1) in Auckland for nearly 3 months during autumn and winter (April to July) in 2010. Elangasinghe's model gave an $R^2 = 0.66$. In comparison similar sites with a traffic/industrial and traffic/residential background, in this research, at Penrose and Pakuranga gave a prediction performance of $R^2 = 0.53$ and $R^2 = 0.61$ respectively.

An empirical approach was taken in for tuning site-specific LSTM models to provide diagnostic examination of different experiment designs over time, along with objective analysis of test RMSE. The optimised LSTM model showed minimum RMSE values (3.77) for Henderson and highest RMSE value (5.21) for Takapuna site. Based on the R^2 value ranging between (0.45-0.59), performance of the LSTM for Auckland PM_{10} time series problems were not as satisfactory as expected when considering LSTM results provided in the reviewed literature (Section 8.4). The Auckland LSTM models could not characterize the complex features of the sequential data efficiently as they learnt long-interval time series PM_{10} data with high non-linearity. A deep multi-layer LSTM model with varied numbers of neurons in each layer may overcome this problem and this may be worth exploring in future work. In contrast, the site-specific RF models developed and presented in this study proved to be the most efficient and effective – giving improved the model fit when compared with MLP models. Despite the limited number of key predictors used when compared to those used in other similar studies (Section 8.5) the RF models performed relatively well (R^2: 0.63-0.71) in the main urban areas as well as rural Patumahoe (R^2: 0.66) which is characterized by a very different background when compared to Auckland's urban areas.

For comparison of the accuracies of these models the IOA (Table 8.14) was used to provide an overall measure and determine which method is applicable for daily operational PM_{10} predictions. Based on IOA, RF outperformed both MLP and LSTM model. The LSTM model had the lowest IOA in the range of 0.51 to 0.62.

Table 8.14: Index of Agreement for all methods.

	Index of Agreement		
	MLP	LSTM	RF
Glen Eden	0.63	0.51	0.79
Henderson	0.59	0.57	0.69
Patumahoe	0.65	0.59	0.72
Pakuranga	0.66	0.62	0.69
Penrose	0.54	0.53	0.72
Takapuna	0.58	0.51	0.73

8.8 Conclusions

This Chapter reports on the experiments that investigated methods based on **routinely available and parsimonious meteorological and temporal parameters**, to build PM_{10} forecasting models in two ways. Firstly, the application of two input selection algorithms was evaluated. It was found that the best method differed and that a case-based selection of which approach is suitable is required when determining whether to use features selected by forward selection or backward elimination. Secondly, a PCA dimension reduction approach was evaluated to determine the most influential meteorological and temporal variables with regards to the underlying structures in the data. The components that contributed the most towards explaining the variance in the data were used as inputs to the models and helped to examine the primary relationships among variables. In this study, predictive models of one day ahead PM_{10} concentration using three different machine learning approaches were developed and their site-specific performance evaluated.

In recent years, large numbers of studies have started to utilize machine-learning methods to predict PM_{10} daily concentrations over large geographical domains. A comparison of previous studies with the one proposed here is not easy due to the unique and (un)known microclimate characteristics of Auckland area (Chapter 3) and the limitations of in the data sources which are used as predictors. The outcomes of prediction using the RF models developed in this study suggest that RF can help to understand the extent of the variability in concentrations that can be explained by the governing meteorological variables. The RF method based on the findings of this research is the most promising of all the approaches investigated. It is likely that the main limitations of the Auckland case study area such as the little variability in observed PM_{10}

concentrations, and the lack of key predictors including traffic and population data resulted in the RF ensemble method being a more effective approach. RF’s can, unlike LSTM and MLP, handle missing data, automatically balance datasets, is good at training using a small number of samples. This means RF is a model worth considering in cases like predicting Auckland’s PM_{10} where air quality management is carried out using few air quality measurement sites, and where the data is simply not available to have the detailed emission inventories that are essential for the implementation of complex physically based dispersion models.

Given the characteristics of the Isthmus city, use of emission rates might increase model performance by capturing the complexities that have not been depicted in the present model. In this study, injecting the bagging and random features as the only kind of randomness made RF models accurate classifiers and regressors compared to the MLP and LSTM models. As suggested by Breiman (2002) other types of injected randomness such as use of random Boolean combinations of features may improve the results.

It is important to note that RF is ***not*** good at detecting rare items or events such as the Auckland’s Sky Tower fire in 2019 and the Australian bush fires which both resulted in high spikes in PM concentrations in Auckland. Using data stream mining methods such as change detection along with RFs would be one avenue worth exploring though the current state of work in this area means that change detection approaches are still not effective in detecting rare and extreme events.

Chapter 9 CONCLUSIONS AND SIGNIFICANT OUTCOMES

In the 2020's there is almost undisputed scientific consensus that air quality degradation is a significant environmental and health hazard and, thus, a great deal of investigation is being conducted in this field. New Zealand is perceived as a green and clean country when compared to the rest of the world. However, it is important to recognise that population growth and urbanisation are major contributors to rising levels of air pollutants. These drivers are being seen in the Auckland region with infill housing and high-density housing along with a rapidly increasing population. While the COVID19 pandemic of 2020-21 has slowed this rapid growth, the growth continues, and it is expected that a rapid and sudden growth will occur once international travel and immigration is opened up again.

A general view of air quality at the global level, the state of New Zealand ambient air (from the start of monitoring until the present) and for the Auckland region in particular was provided in Chapter 2. Apart from the prevailing weather conditions, the main source of PM_{10} concentration in New Zealand is clearly related to anthropogenic activities (e.g., traffic flows, domestic heating patterns).

From the literature review on recent PM_{10} modeling approaches it was concluded that there is a general lack, apart from a few short-term single-site case studies (of at most a year in length), of PM_{10} studies in Auckland. There are certainly, as of the time of writing this thesis, no other longer-term studies located in either the white or grey literature. The first running and still-current monitoring site in Auckland's airshed, namely Penrose, has been operating since 1994. Over the last ten years, the number of PM_{10} monitoring stations in Auckland's network has been reduced. Consequently, the network collecting PM, weather and other gaseous pollutants that contribute to PM_{10} concentration is sparse. This has the consequence that any models developed for Auckland are site-specific. While the reduction of monitoring sites might help to reduce, at least in the short-term, operational costs, it also results in limitations with respect to models of and estimations of PM_{10} concentration. In some cases, temporary dense small area monitoring networks have been implemented in Auckland by research institutes, such as Universities and NIWA, as part of government funded short term projects. Such sites are only live for the term of the project which may be from a few months to three years. Consequently, these initiatives do not contribute significantly to the monitoring efforts or to the development of models for understanding and forecasting Auckland's PM. Because of the challenges in maintaining and growing these monitoring networks this research has focused on parsimonious models which use readily available data rather than specialized research monitoring stations.

What follows is a discussion of the key findings, contributions and ideas for future research listed by research question.

9.1 Research Questions: Data Exploration, Quality, and Imputation

Q1.1. What is the quality of the meteorological and PM data available for the Auckland airshed?

The quality of both PM_{10} and meteorological data was addressed in Chapter 3. The analysis of the data found that on the whole the quality of the data available was relatively poor. Rainfall in particular was missing completely for one station (Henderson, onsite rainfall was not collected) and missing most readings for two another stations (Takapuna, 85 continuous days with missing data points and Penrose with missing 72, 30 and 22 continuous days). In order to ensure that the data used in the models reported in this thesis was as complete and as accurate and reliable as possible WHO guidelines for PM_{10} were adopted and methods for obtaining or imputing the missing data were explored.

To ensure compliance with WHO regulations, 1-hour data was checked against the WHO criteria to determine whether they could be used to estimate average PM_{10} concentrations. Daily average concentration was calculated for a site only if the hourly concentration data in one day was available for at least 50% of the day (i.e.,12 hours in 24 hours). Any days where there was less than 50% of the data were excluded. Even if the WHO 50% threshold is met, calculating daily average from midnight to midnight can cause overestimation of daily average PM_{10} concentration when only the evening data are available. To address this issue, data with such missing patterns were identified and excluded from the averaging procedure. The site with highest missing days was Penrose (1.4%). There was notable unexplained exceedance of PM_{10} at Patumahoe and 27 days of continuous missing PM_{10} data in Henderson. In total, 65 days of peak PM_{10} events were observed.

The meteorological data for all of the urban sites was obtained from the Auckland City Council and was provided in 1-hour resolution. For the rural Patumahoe station, daily averaged data was collected from another nearby weather station owned by NIWA, New Zealand through their "*clifo*" database. Quality control of hourly meteorological data was carried out by eliminating missing data and obviously unreasonable data. Missing rainfall values were calculated using the National Oceanic and Atmospheric Administration (NOAA)'s Hydro-Estimator data which is derived from satellite images.

Q1.2. What is the spatial/temporal relationship between Auckland's PM_{10} site measurements and meteorological over the entire region in different time scales (daily, monthly, and yearly)?

Daily mean temperature showed a negative exponential relationship with the daily average PM_{10} measurements. Seasonal and yearly analysis showed high PM_{10} concentrations also appear to correlate with low temperature days, that can be attributed to the use of wood burners during cold seasons. Rainfall appeared to have some site-specific predictive ability as shown by the low pairwise Pearson's correlation coefficients. Solar radiation had limited predictive ability except in the case of the Glen Eden site. The relationships between the wind speed and wind direction on PM_{10} concentrations were site-specific. Highest PM_{10} concentrations were observed in Henderson when the wind is from the westerly and easterly directions. This aligns with marine aerosols driven in from the Tasman Sea and Pacific Ocean. Analysis of the seasonal variation of PM_{10} by wind speed, showed highest PM_{10} concentrations occur at the two extremes of wind speed in Henderson. Peak PM_{10} are highest on cold calm winter days under inversion conditions or with a light southerly wind, particularly for anticyclones synoptic conditions. The high concentration of PM_{10} was observed in Patumahoe during warmer months that can be attributed to dusts and soil sources originated from agricultural and land use activities in the area.

Q1.3. Can missing rainfall data be imputed from satellite-based sources for the Auckland airshed? and How accurate are these satellite rainfall measurements?

Since there was not a priori basis for excluding rainfall variables from a model, satellite rainfall measurements were collected from the nearest proximity to each site through the procedure noted on the National Oceanic and Atmospheric Administration (NOAA) website. So, rainfall data can be imputed for Auckland's airshed using this method, which was originally implemented for use largely in the Northern Hemisphere. While computing these values this thesis researcher noted inconsistencies in the results and communicated with the NOAA researcher who worked on the original method (mathematical equations) – it was found that a value had not been negated for the Southern Hemisphere. This oversight was corrected, and the rainfall values appeared to be within reasonable and known ranges for the Auckland region. As a test of the accuracy of the rainfall data ground truth rainfall gauge data was compared with the rainfall in millimeters predicted by the NOAA Hydro-Estimator method there was only on average 60% agreement but given this was the only source of data available it was employed in the models with this known caveat.

So, in answer to this question satellite imagery can be used to impute rainfall – other researchers have also used this approach – however, more work needs to be done to improve this approach and its accuracy for the Southern Hemisphere. This is a thesis in itself and was considered to be outside of the scope of this thesis.

Q1.4. How is PM_{10} concentration influenced by Aerosol Optical Thickness (AOT)?

As part of the research undertaken during this thesis work code was written to obtain and extract AOD data ranging from 0.47- 2.13 microns along with Cloud fraction from the land aerosol cloud mask (MODIS satellite imagery). However, exploration of the data extracted revealed very poor-quality data due to missing values and also low values for the "quality" flag parameter included in the data. This appears to be an issue unique to NZ as the AOT measures obtained from MODIS is used widely for studies in the Northern Hemisphere. It was not clear why the data is so poor for Auckland but in any case, it was decided not to use it this data in this thesis's research. As a result, this question remains a question for future researchers to investigate with a specific focus first on the estimation of AOT from satellite imagery and later once this is reliable for NZ and Auckland in particular the role of AOT in PM_{10} prediction can be explored.

Q1.5. Do seasonality trends exist in Auckland's PM_{10}? And if these trends exist what is the nature of seasonal patterns in Auckland's PM_{10}? Which seasonality detection method should be used to account for any observed seasonal trends in Auckland's PM_{10}?

The time series analysis presented in Chapter 3 was intended to form the basis for further modeling approaches conducted in remaining Chapters and (non)stationarity of the time series were examined through KPSS and ADF tests. Unlike ADF test, we failed to reject the null hypothesis of stationarity around a deterministic trend in KPSS test. After discussing this result with one of the package's authors (Miranda & Yee, 2018), it was concluded that the contradictory results observed could be due to the presence of multiple seasonality in the time series that was not observed by ADF. To answer the question in regard to detecting the nature of seasonality and its pattern seasonal decomposition were carried on in Chapter 4. The seasonal component of Auckland PM_{10} time series captured the expected variation in the mean of the daily PM_{10} concentrations during the cold season by exhibiting a peak in mid-winter using STL decomposition method. Auckland time series with higher frequency of daily observation of PM_{10} observations, exhibited complicated seasonal patterns that could not picked up by classical decomposition methods such as STL. The two distinct weekly and yearly seasonal patterns were identified using the TBATS model for each site.

So, in conclusion seasonality pattern do exist in Auckland's PM_{10} and these patterns are a result of **multiple complex seasonalities including daily, weekly, and annual trends**. Of the seasonality detection methods explored it appears that **TBATS** is the best method for detecting seasonal trends in Auckland PM_{10} timeseries.

9.2 Research Questions: Prediction Models

Q2.1. In the absence of comprehensive emission inventories or information on potential sources of emission affecting a particular site, to what degree can the daily concentration of PM_{10} in Auckland airshed be explained by site-specific predictors variables?

To handle multiple seasonality in the estimation of next day PM_{10} concentration, harmonic regression using the Fourier terms (seasonal periods of 7 and 365 days based on the generated ACF and PACF plots), and two relatively new state-space modeling frameworks namely harmonic regression with Fourier terms and TBATs were employed in Chapter 4. The TBATS model differed from dynamic harmonic regression in that the seasonality was changed gradually over time, while seasonal patterns were forced by the harmonic regression terms to replicate periodically without changing. **Site specific TBATS models with a lesser number of parameters and non-integer seasonality were found to be a more suitable technique for forecasting of Auckland's PM_{10} with minimal errors, when compared to harmonic regression models, when there is the presence of complex seasonality.**

In order to explore these findings further the literature in the area was examined. It is clear from the literature that, at this point in time, there is no one single model that is globally suitable for PM_{10} prediction/forecasting and that different models performed differently in different regions. These differences appear to be related to data quality, choice of variables and in particular one of the most influencing factors appears to be seasonality and temporal cycles in the data. Perhaps in the end the best that we can hope for is a set of heuristics for selecting a model based on recent or historic trends in a region and one possible avenue for future research should be examining machine learning approaches for model learning so that the models are learnt and change with the changing input data. In this thesis, in Chapter 8 a Random Forest ensemble machine learning approach is evaluated and showed promising results – this suggests that further exploration of real-time streaming data and a machine learning approach in which models are learnt on the fly might be in the end the best approach to forecasting PM_{10}. This is of course with the proviso that data of sufficient quality is available. Advances in modeling need to be considered hand in hand with improvements in monitoring – especially in the case of Auckland.

In this research, the non-parametric Mann-Kendall tests indicated no significant changing trend in PM_{10} for Henderson, Patumahoe and Penrose sites. The Box Cox test also accepted the null hypothesis of no monotonic trends in Glen Eden, Henderson, Patumahoe and Penrose sites. The Modified Mann-Kendall Test for serially correlated data was used to account for the serial correlation present in the daily PM_{10} concentration values. The reported *p-value* after variance correction (*value^{b}*) concluded that there was a significant downward trend for Takapuna and a negative but not significant trend for Glen Eden. The trends for the remaining sites were positive but not significant. This illustrates the very localised nature of PM concentration trends which cannot be explained fully by land usage (rural vs, urban).

The relationship between the simple and routinely available meteorological factors and PM_{10} concentrations is relatively underexplored in literature and the significance of this relationship to assign site-specific source contributions is not well understood. In Chapter 5 and 6 the influence of meteorological and temporal factors on the distribution of the daily average of PM_{10} concentration is modelled using GAMs and GAMMs. PM_{10} concentrations for each of the six stations were estimated separately using various models, with all available meteorological variables included. Although prior research examined collinearity solely based on correlation coefficients, in this research concurvity of each term with the whole the GAM model was calculated with assumption of independent and identically normally distributed errors for inference purposes. For all input variables, the estimated values were lower than 1, with a maximum value of 0.5 found between relative humidity and solar radiation in Takapuna that are likely due to the seasonal trends of these two terms. Therefore, it was assumed that collinearity do not exists and that a GAM method can therefore be applied. The partial effect results showed a clear difference between the effects of anthropogenic sources (crudely adapted using temporal variables) and atmospheric conditions. **To the author's knowledge, this study is the first effort to use GAMs and GAMMs to model PM_{10} concentration in the Auckland airshed, where multiple sources contribute to the observed PM_{10} concentrations.**

Chapter 8 further investigated how to extract the key information required from routinely available meteorological and temporal parameters, to build PM_{10} forecasting models in two manners. Application of input selection algorithms to identify the optimal set of model inputs showed different results, therefore a suitability study was carried out using a case-specific analysis. Different inputs were selected when forward selection and backward elimination were applied to Henderson, Pakuranga, and Takapuna and Patumahoe. In the case of Patumahoe backward elimination selected all the predictor variables as important in making the best PM_{10} prediction. Penrose selected the same features in both forward selection and backward elimination (Lag_1, relative humidity, temperature, wind speed, month, rain, wind direction, and solar radiation). The results of PCA showed the first component in all urban sites included relative humidity, solar radiation, and rainfall. The first component for Patumahoe included lag_1 and lag_2. It was noted that day of week was only selected in fifth component of Henderson.

At an early point in this research elevation was considered as a predictor variable. The range in elevations within the study regions is relatively small. It was found that elevation did not contribute to the models. However, it may be that in topologies with a greater range of elevations that elevation does become important.

Q2.2. Can computationally simple semi-empirical methods such as GAMs and GAMMs be used to model Auckland's daily PM_{10} concentrations reliably?

As presented in Chapter 5, GAM model showed a moderate strength of the linear relationship between the observed and predicted values (R^2) ranging from 0.55 to 0.69 for Pakuranga and Patumahoe stations, respectively. For low concentrations of PM_{10}, the values fall within the FAC2 region, whereas for higher concentrations of PM_{10} scattering was more evident. The GAM model in this study performed in a similar manner ($0.20<R^2<0.44$) to a GAM model ($R^2 = 0.49$) for estimating annual average PM_{10} concentration in the U.S. using several geographic information systems–derived covariates. They used number of meteorological, geographical and traffic volume variables. In the research for this doctoral thesis traffic volume was not readily/openly available and therefore not included. In Auckland traffic volume is monitored by Waka Kotahi (The NZ Transport Agency) and is only available as summarized data and only key sites such as the Water View tunnel are monitored and often for only short periods of time.

GAMM modeling was applied to the Auckland case with the assumption that the errors can exhibit a particular form of dependence and the smooth functions can allow much more flexible functional forms than simple linear terms can. These conclusions were confirmed based on the GAMM results, where autocorrelation was also modeled. The result of cross validation of the fitted GAMMs on each site's testing dataset showed positive values of Fractional Bias (FB). However, the calculated values of FB are too small and close to zero indicating that there is no systematic tendency to under prediction. The predictive ability of GAMM models for Auckland's PM_{10} was comparable ($0.36<R^2<0.48$) than those reported for modeling PM_{10} in the Northeastern and Midwestern U.S. ($R^2=0.58$).

While simple semi-empirical methods can be used to model Auckland's PM_{10} concentrations and on the whole these models perform better for Auckland's PM_{10} prediction than for other countries and regions (as reported in the literature), the performance is not sufficient to be able to forecast with sufficient confidence. For example, it would not be recommended on this basis of this thesis's work that semi-empirical models are used to report future concentrations in order to high light periods of risk for citizens with respiratory conditions.

Q2.3. How can we, from a high-level perspective, use descriptive (marginal) models to characterize spatio-temporal dependence structures for PM_{10} modeling?

Due to a lack of high spatial resolution of meteorological data, IDW was used to interpolate the PM_{10} concentrations between the air quality monitoring stations. This interpolation was conducted without considering potential influencing factors such as weather/meteorological conditions. Instead, only spatial (latitude and longitude) and temporal variables (days) were used as input variables. Three other spatio-temporal statistical models (regression with trend, ST_GLM and ST-GAM) were also considered in order to be able to account for the meteorological factors as well as spatial coordinates in their data driven models. Using the S-T approach and considering these covariates in form of basis functions, these **S-T models were found to provide a better representation of the current situation in the study area with S-T GAM outperforming S-T GLM, regression with trend, and IDW.**

In Chapter 7, descriptive approaches were used to obtain statistical dependencies in spatio-temporal distribution of PM_{10}. Different S-T variograms were fitted to the data and the best model was used in Kriging to model and predict PM_{10} concentration in study area. **The advantage of the S-T kriging over deterministic methods (Chapter 6) was its ability to provide a better prediction precision by describing the spatio-temporal dependence structure through the inclusion of spatio-temporal covariances**.

Further work to improve the performance of S-T kriging for modeling of Auckland PM_{10} data set should be undertaken in future work and include careful consideration of the data preprocessing steps such as: employing additional covariates in linear modeling step and use of a temporal AR-model in preprocessing step followed by spatio-temporal residual kriging.

Q2.4. How reliable and efficient are machine learning methods for predicting next day PM_{10} concentration?

In Chapter 8, predictive models of one day ahead PM_{10} concentration using three different machine learning approaches were developed and their site-specific performance evaluated. In the last ten years, there has been a proliferation of studies using machine-learning methods to predict PM_{10} daily concentrations over large geographical domains. It is difficult to compare the performance of the methods used in previous studies with the one proposed here due to unique and well-known microclimate characteristics of Auckland region and differences in predictor data sources. Given the main limitations of the research presented in this thesis such as the low variability in observed PM_{10} concentrations and the lack of key predictors including traffic and population data, **the predictive outcomes of the RF models developed in this thesis suggest that RF can help to explain the extent variability in PM_{10} concentrations that can be explained by the governing meteorological variables.**

Q2.5. How accurate are non-linear statistical and ensemble approaches in predicting next day PM_{10} concentration?

In Chapter 8, injecting the bagging and random features as the only kind of randomness made RF models accurate classifiers and regressors compared to the MLP and LSTM models. As suggested by Breiman (2002) other types of injected randomness such as use of random Boolean combinations of features may improve the results. Using data stream mining methods such as change detection along with RFs would be one avenue worth exploring though the current state of work in this area has reported that change detection approaches are still not effective in detecting rare and extreme events. **The random forest methodology was determined to be the best option for modeling/predicting PM_{10} of the methods explored in this research.**

Q2.6. To what extent does data quality become a significant factor in determining the performance of PM_{10} concentration models?

Given the characteristics of the isthmus as a predominant geographical feature of Auckland city the use of emission rates might improve model performance by capturing the complexities that have not been depicted in the models built in this thesis's research. While statistical approaches are restricted in their ability to explain the physical parameters controlling the pollutant dispersion, development of models to successfully simulate atmospheric processes require an understanding of the fine detail of dispersion processes, and to develop source receptor relationships. The practical application of such techniques is however limited due to their extensive data requirements.

It should be noted that the traffic characteristics and other anthropogenic source were not directly included as predictors for modeling PM_{10}. Accordingly, further investigation using these factors over a longer period to be directly included in the models is suggested. Use of imputation/interpolation methods for replacing missing PM_{10} concentration and meteorological measurements is also suggested for future work. In this study it was presumed these missing data are missing completely at random, and therefore this will influence the trend estimates and test.

Similarly, another notable limitation of the work presented in this thesis is related to the of lack of rainfall data for Henderson (no availability), Takapuna and Penrose site (Chapter 3), rainfall data was imputed based on satellite imagery provided by NOAA/NESDIS. As a test of the accuracy of the rainfall data ground truth rainfall gauge data was compared with the rainfall in millimeters predicted by the NOAA Hydro-Estimator method there was only on average 60% agreement but given this was the only source of data available it was employed in the models with this known caveat. This could explain why Henderson rainfall was not selected by the GAM and GAMM smooth terms in Section 5.4. Also, Henderson rainfall was found to have the lowest importance level (17.1) compared to remaining sites' rainfall level (above 24) when used in RF model (Section 8.7.5). Imputing missing data almost always adds some uncertainty to any model developed.

9.3 Final Thoughts

Modeling and prediction of PM_{10} is not straight forward there are a number of key aspects that make this challenging. Those aspects that are in control of the stakeholders include:

- Sensor choice, maintenance, and calibration. Obviously, there is still a need for the development of more reliable and cost-effective sensors for both pollution and weather monitoring.
- Sensor network density – It is clear that due to the localized nature of PM that denser networks are required for the purposes of modeling and predicting PM. Other regions in New Zealand (for example Canterbury in the South Island) have more comprehensive monitoring in place.
- Consistency in same data collection across stations in the network.
- Lack of open data sources. Many data collections are proprietary with respect to data and access is difficult for researchers. Different data is held by different entities – sharing of this data would enable better imputation of missing data points. Only some of the data from organizations such as Auckland Council is shared. However, with the advent of their new data portal the data is becoming more open and readily available to researchers.
- Improvement of satellite imagery and methods for extracting features such as precipitation and AOT from these images for the Southern Hemisphere. Much of the work in this area has been conducted and fully evaluated only in the Northern Hemisphere. Improving computations in for example the Hydro-Estimator would improve and help: triangulate and verify ground truth readings impute data

Once these issues are addressed then while modeling will still be challenging the data quality will result in better performance of the models. This is critical when key policies and decisions are being made both at a local and national level based on these models.

Finally, machine learning approaches should be considered as methods for validating sensor data and as a means to correct questionable readings. It is clear from this research that ensemble machine learning methods – methods that learn models as new data is introduced – are the way forward.

References

Abdi, H., & Williams, L. J. (2010, 2010/07/01). Principal component analysis [https://doi.org/10.1002/wics.101]. *WIREs Computational Statistics, 2*(4), 433-459. https://doi.org/https://doi.org/10.1002/wics.101 AC. (2006). *The ambient air quality monitoring network in the Auckland Region* (296).

Adhikari, R., & Agrawal, R. K. (2013). Time Series Forecasting Using Stochastic Models In *An Introductory Study on Time Series Modeling and Forecasting*. LAP Lambert Academic Publishing.

Afzali, A., Rashid, M., Sabariah, B., & Ramli, M. (2014). PM10 Pollution: Its Prediction and Meteorological Influence in PasirGudang, Johor. *Earth and Environmental Science, 18*. https://doi.org/10.1088/1755-1315/18/1/012100

Aga, E., Samoli, E., Touloumi, G., Anderson, H. R., Cadum, E., Forsberg, B., Goodman, P., Goren, A., Kotesovec, F., Kriz, B., Macarol-Hiti, M., Medina, S., Paldy, A., Schindler, C., Sunyer, J., Tittanen, P., Wojtyniak, B., Zmirou, D., Schwartz, J., & Katsouyanni, K. (2003). Short-term effects of ambient particles on mortality in the elderly: results from 28 cities in the APHEA2 project. *European Respiratory Journal, 21*, 28-33. https://doi.org/DOI: 10.1183/09031936.03.00402803

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. Second International Symposium on Information Theory, Budapest.

Akbari, M., Asadi, P., Besharati-Givi, M. K., & G.Khodabandehlouie, G. (2014). Artificial neural network and optimization. In M. K. Besharati-Givi (Ed.), *Advances in Friction-Stir Welding and Processing* (Vol. 1). Woodhead Publishing. https://doi.org/https://doi.org/10.1533/9780857094551.543

Aldrin, M., & Haff, I. H. (2005, 2005/04/01/). Generalised additive modelling of air pollution, traffic volume and meteorology. *Atmospheric Environment, 39*(11), 2145-2155. https://doi.org/https://doi.org/10.1016/j.atmosenv.2004.12.020

Amaral, S. S., De Carvalho, J. A., Jr., Costa, M. A. M., & Pinheiro, C. (2015). An Overview of Particulate Matter Measurement Instruments. *Atmosphere, 6*, 1327-1345. https://doi.org/10.3390/atmos6091327

Anastassiou, G. A. (2011, 2011/02/01/). Multivariate hyperbolic tangent neural network approximation. *Computers & Mathematics with Applications, 61*(4), 809-821. https://doi.org/https://doi.org/10.1016/j.camwa.2010.12.029

Ancelet, T., & Davy, P. (2014). *Multi-elemental analysis of PM10 and apportionment of contributing sources-Tokoroa* (4093981, Issue.

Ancelet, T., Davy, P. K., Trompetter, W. J., & Markwitz, A. (2014, 2014/10/01/). Sources of particulate matter pollution in a small New Zealand city. *Atmospheric Pollution Research, 5*(4), 572-580. https://doi.org/https://doi.org/10.5094/APR.2014.066

Appelhans, T. (2010). *A climatology of particulate pollution in Christchurch* University of Canterbury].

Appelhans, T., Sturman, A., & Zawar-Reza, P. (2013). Synoptic and climatological controls of particulate matter pollution in a Southern Hemisphere coastal city. International Journal of Climatology,

Armstrong, M. (1998). *Basic Linear Geostatistics*. Springer-Verlag.

Barmpadimos, I., Hueglin, C., Keller, J., Henne, S., Prev´, A. S. H., & 179, P. (2011). Influence of meteorology on PM10 trends and variability in Switzerland from 1991 to 2008. *Atmos. Chem. Phys., 11*, 1813–1835.

Becerra-Rico, J., Aceves-Fernández, M. A., Esquivel-Escalante, K., & Pedraza-Ortega, J. C. (2020). Airborne particle pollution predictive model using Gated Recurrent Unit (GRU) deep neural networks. *Earth Science Informatics*. https://doi.org/https://doi.org/10.1007/s12145-020-00462-9

Belusic, A., Bulić, I. H., & Klaic, Z. B. (2015). Using a generalized additive model to quantify the influence of local meteorology on air quality in Zagreb. *Geofizika* DOI: 10.15233/gfz.12015.15232.15235.

Belušić, A., Herceg-Bulić, I., & Klaić, Z. B. (2015). Using a generalized additive model to quantify the influence of local meteorology on air quality in Zagreb. *GEOFIZIKA, 23*. https://doi.org/10.15233/gfz.2015.32.5

Bertaccini, P., Dukic, V., & Ignaccolo, R. (2012). Modeling the Short-TermEffect of Traffic andMeteorology on Air Pollution in Turin with Generalized AdditiveModels. *Advances in Meteorology*, doi:10.1155/2012/609328.

Biancofiore, F., Busilacchio, M., Verdecchia, M., Tomassetti, B., Aruffo, E., Bianco, S., Di Tommaso, S., Colangeli, C., Rosatelli, G., & Di Carlo, P. (2017, 2017/07/01/). Recursive neural network model for analysis and forecast of PM10 and PM2.5. *Atmospheric Pollution Research, 8*(4), 652-659. https://doi.org/https://doi.org/10.1016/j.apr.2016.12.014

BOPRC. (2014). *Offsets Guidance for the Rotorua Airshed*. B. o. P. R. Council.

Box, G. E. P., & Jenkins, G. M. (1990). *Time series Analysis: Forecasting and Control.* Francisco Holden-Day.

Boznar, M., Lesjak, M., & Mlakar, P. (1993, 1993/06/01/). A neural network-based method for short-term predictions of ambient SO2 concentrations in highly polluted industrial areas of complex terrain. *Atmospheric Environment. Part B. Urban Atmosphere, 27*(2), 221-230. https://doi.org/https://doi.org/10.1016/0957-1272(93)90007-S

Breiman, L. (2001, 2001/10/01). Random Forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001). RANDOM FORESTS. *Machine Learning, 45*, 5-32. https://doi.org/http://dx.doi.org/10.1023/A:1010933404324

Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica, 47*(5), 1287-1294. (The Econometric Society)

Brimblecombe, P. (1999). Air Pollution and Health History. In S. T. Holgate, J. M. Samet, & R. L. Maynard (Eds.), *Air Pollution and Health* (pp. 5-18). https://doi.org/https://doi.org/10.1016/B978-0-12-352335-8.X5074-1

Brockwell, J. P., & Davis, R. A. (2002). *Introduction to Time Series and Forecasting, Second Edition*. Springer Texts in Statistics.

Bruno, F., Guttorp, P., Sampson, P. D., & Cocchi, D. (2009a). A simple non-separable, non-stationary spatiotemporal model for ozone. *Environ Ecol Stat*, 515–529.

Bruno, F., Guttorp, P., Sampson, P. D., & Cocchi, D. (2009b, 2009/12/01). A simple non-separable, non-stationary spatiotemporal model for ozone. *Environmental and Ecological Statistics, 16*(4), 515-529. https://doi.org/10.1007/s10651-008-0094-8

Burkhardt, J., & Grantz, D. A. (2016). Plants and Atmospheric Aerosols. In *Progress in Botany* (Vol. 78). https://doi.org/10.1007/124_2016_12

Cabanerosa, S. M., Calautitb, J. K., & Hughesa, B. M. (2019). A review of artificial neural network models for ambient air pollution

prediction. *Environmental Modelling & Software, 119*, 285-304. https://doi.org/https://doi.org/10.1016/j.envsoft.2019.06.014

Carslaw, D. C., & Ropkins, K. (2012). openairdAn R package for air quality data analysis. *Environmental Modelling & Software, 52*, 52-61.

Chaloulakou, A., Kassomenos, P., Demokritou, N., & Koutrakis, P. (2003). Measurements of PM10 and PM2.5 Particle Concentrations in Athens, Greece. *Atmospheric Environment, 37*(5), 649-660.

Chappell, A., & Agnew, C. T. (2012). Geostatistical Analysis and Numerical Simulation of West African Sahel Rainfall. In A. J. Conacher (Ed.), *The GeoJournal Library: Land Degradation* (Vol. 58, pp. 19-35). https://doi.org/https://doi.org/10.1007/978-94-017-2033-5_2 (Springer, Dordrecht)

Chatea, D. M., & Praneshab, T. S. (2004). Field studies of scavenging of aerosols by rain events. *Journal of Aerosol Science, 35*(6). https://doi.org/https://doi.org/10.1016/j.jaerosci.2003.09.007

Chen, G., Wang, Y., Li, S., Cao, W., Ren, H., Knibbs, L. D., Abramson, M. J., & Guo, Y. (2018). Spatiotemporal patterns of PM 10 concentrations over China during 2005-2016: A satellite-based estimation using the random forests approach. *Environ Pollut, 242*, 605-613. https://doi.org/10.1016/j.envpol.2018.07.012

Chien, L. C. (2009). *Multi-city time series analyses of air pollution and mortality data using Generalized Geoadditive Mixed Models* University of North Carolina]. Chapel Hill, North Carolina.

Chlebowska-Styś, A., Sówka, I., D., K., & Pachurka, L. (2017). Analysis of concentrations trends and origins of PM10 in selected European cities *E3S Web of Conferences, 17*. https://doi.org/DOI: 10.1051/e3sconf/20171700013

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. J. (1990). STL: A Seasonal-trend Decomposition Procedure Based on Loess. . *Journal of Official Statistics, 6*(1), 3-33. https://doi.org/http://bit.ly/stl1990

Cleveland, W. S., & Devlin, S. J. (1988). Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association, 83*(403).

Clifford, S. (2013). *Spatio-temporal modelling of ultrafine particle number concentration* Queensland University of Technology].

CliFlo, N. (2015). *CliFlo* https://cliflo.niwa.co.nz/

Cook, R. D., & Weisberg, S. (1983). Diagnostics for heteroscedasticity in regression [Article]. *Biometrika, 70*(1), 1-10. https://doi.org/10.1093/biomet/70.1.1

Council, A. (2020). *Ceramco Park Function Centre.* https://bookings.aucklandcouncil.govt.nz/facilities/facility/ceramco-park-function-centre

Cressie, N., & Christopher, K. W. (2011). *Statistics for Spatio-Temporal Data.*

Cressie, N., & Huang, C. (1999). Classes of nonseparable, spatio-temporal stationary covariance function. *Journal of the American Statistical Association*, 1330-1340.

Czernecki, B., Półrolniczak, M., Kolendowicz, L., Marosz, M., Kendzierski, S., & Natalia Pilguj, N. (2016). Influence of the atmospheric conditions on PM10 concentrations in Poznań, Poland. *Journal of Atmospheric Chemistry, 74*, 115–139.

Dagum, E. B. (2010). Time Series Modeling and Decomposition. *STATISTICA, 4*, 433-457.

Das, M., & Bhattacharya, S. (2014). Nonstationary, Nonparametric, Nonseparable Bayesian Spatio-Temporal Modeling Using Kernel Convolution of Order Based Dependent Dirichlet Process. https://doi.org/arXiv:1405.4955

David, F., Pyle, J. A., Sutton, M. A., & Williams, M. L. (2020). Global Air Quality, past present and future: an introduction. *Phil. Trans. R. Soc. A., 378*(2183). https://doi.org/https://doi.org/10.1098/rsta.2019.0323

Davy, P. K., Trompetter, W. J., & Markwitz, A. (2011). *Source apportionment of airborne particles in the Auckland region: 2010 Analysis*.

Davy, P. K., Trompetter, W. J. T., Ancelet, T., & Markwitz, A. (2017). *Source Apportionment and Trend Analysis of Air Particulate Matter in the Auckland Region* (TR2017/001). A. Council.

De Cesare, L., Myers, D. E., & Posa, D. (1997). Spatial-temporal modeling of SO2 in Milan district. *10.1007/978-94-011-5726-1_34*, 1031-1042.

De Gooijer, J. G., & Hyndman, R. J. (2006). 25 Years of Time Series Forecasting. *International Journal of Forecasting*, 443–473.

De Iaco, S. (2010). Space–time correlation analysis: A comparative study. *Journal of Applied Statistics*, 1027-1041.

De Iaco, S., Myers, D. E., & Posa, D. (2001). Space-time analysis using a general product-sum model. *Statistics and Probability Letters*, 21-28.

De Iaco, S., Myers, D. E., & Posa, D. (2002a). Nonseparable space-time covariance models: some parametric families. *Mathematical Geology*, 23-42.

De Iaco, S., Myers, D. E., & Posa, D. (2002b). Space–time variograms and a functional form for total air pollution measurements. *Computational Statistics & Data Analysis*, 311-328.

De Iaco, S., & Posa, D. (2012). "Predicting spatio-temporal random field: Some computational aspects. *Comput. Geosci.*, 12–24.

De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011, 2011/12/01). Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American Statistical Association, 106*(496), 1513-1527. https://doi.org/10.1198/jasa.2011.tm09771

Delgado, A., Ricardo, R., Maque, A., & Chiara, C. (2020). Air Quality Prediction (PM2.5 and PM10) at the Upper Hunter Town - Muswellbrook using the Long-Short-Term Memory Method. *International Journal of Advanced Computer Science and Applications, 11*(4), 318-332. https://doi.org/10.14569/IJACSA.2020.0110443

Denham, A. M. (2012). *Geostatistical spatiotemporal modelling with application to the western king prawn of the Shark Bay managed prawn fishery*. http://ro.ecu.edu.au/theses/435

Denham, A. M. (2012). *Geostatistical spatiotemporal modelling with application to the western king prawn of the Shark Bay managed prawn fishery.* EDITH COWAN UNIVERSITY]. Faculty of Computing, Health and Scienc. http://ro.ecu.edu.au/theses/435

Dickey, D. A., & Fuller, W. A. (1976). Distribution of the Estimators for Autoregressive Time Series With a Unit Root. *Journal of the American Statistical Association, 74*(366), 427-431. (Taylor & Francis, Ltd. on behalf of the American Statistical Association)

Dokumentov, A., & Hyndman, R. J. (2015). STR: A Seasonal-Trend Decomposition Procedure Based on Regression. Article JEL classification: C10,C14,C22. http://www.buseco.monash.edu.au/depts/ebs/pubs/wpapers/

Dorofki, M., Elshafie, A. H., Jaafar, O., Karim, O. A., & Mastura, S. (2012). Comparison of Artificial Neural Network Transfer Functions Abilities to Simulate Extreme Runoff Data. International Conference on Environment, Energy and Biotechnology, Singapore.

Doychin, V., Gocheva-Ilieva, S., Ivanov, A., & Iliev, I. P. (2015). Studying the Effect of Meteorological Factors on the SO2 and PM10 Pollution. Seventh Conference of the Euro-American Consortium for Promoting the Application of Mathematics in Technical and Natural Sciences AMiTaNS '15, Albena, Bulgaria.

Durlauf, S. N., & Peter, C. B. P. (1988). Trends versus Random Walks in Time Series Analysis. *Econometrica, 56*(6), 1333-1354. https://doi.org/10.2307/1913101

EEA. (2012). *Particulate matter from natural sources and related reporting under the EU Air Quality Directive in 2008 and 2009* (EEA Technical report., Issue.

EEA. (2020). *Air quality in Europe -2020 report*.

Eilers, P. H. C., & Marx, B. D. (2010). Splines, knots, and penalties. *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(6), 637–653.

Elangasinghe, M. (2014). Applications of semi-empirical and statistical techniques in urban air pollution modelling.

Elangasinghe, M. A. (2014). *Applications of semi-empirical and statistical techniques in urban air pollution modelling* The University of Auckland].

Elangasinghe, M. A., Singhal, N., Dirks, K. N., & Salmond, J. A. (2014). Development of an ANN–based air pollution forecasting system with explicit knowledge through sensitivity analysis. *Atmospheric Pollution Research, 5*(4), 696-708. https://doi.org/https://doi.org/10.5094/APR.2014.079

Emery, C., Liu, Z., G. Russell, A., Odman, M. T., Yarwood, G., & Kumar, N. (2017). Recommendations on statistics and benchmarks to assess photochemical model performance. *Journal of the Air & Waste Management Association*, 582-598.

Emmanuel Hernández Mayoral, M. A. H. L., Edwin Román Hernández, Hugo Jorge Cortina Marrero, José Rafael Dorrego Portela and Victor Ivan Moreno Oliva. (2017). Fourier Analysis for Harmonic Signals in Electrical Power Systems. In *Fourier Transforms High-tech Application and Current Trends* (pp. 43-66). IntechOPen.

EPA. (2018). *Particulate Matter Emissions* (Report on the Environment, Issue.

Eriksson, M., & Siska, P. (2000). Understanding Anisotropy Computations. *Mathematical Geology, 32*(6), 683–700.

Fadnavis, S. (2020). Atmospheric Aerosols and Trace Gases. In R. Krishnan, J. Sanjay, C. Gnanaseelan, M. Mujumdar, A. Kulkarni, & S. Chakraborty (Eds.), *Assessment of Climate Change over the Indian Region*. Springer, Singapore. https://doi.org/https://doi.org/10.1007/978-981-15-4327-2_5

Falbel, D., Allaire, J., & Chollet, F. (2020). *R Interface to 'Keras'.* In (Version 2.3.0.0)

Fang, X., Li, R., Xu, Q., Bottai, M., Fang, F., & Cao, Y. (2016). A Two-Stage Method to Estimate the Contribution of Road Traffic to PM2.5 Concentrations in Beijing, China. *International Journal of Environ Resources and Public Health, 13*(124). https://doi.org/doi:10.3390/ijerph13010124

Fuentes, M., Chen, L., & Davis, J. M. (2007, Nov 5). A class of nonseparable and nonstationary spatial temporal covariance functions. *ENVIRONMETRICS, 19*(5), 487-507. https://doi.org/10.1002/env.891

Galatioto, F., Bell, M. C., & Hill, G. (2014, 2014/11/01). Understanding the characteristics of the microenvironments in urban street canyons through analysis of pollution measured using a novel pervasive sensor array. *Environmental Monitoring and Assessment, 186*(11), 7443-7460. https://doi.org/10.1007/s10661-014-3939-7

Gaynor, P. E., & Kirkpatrick, R. C. (1994). *Introduction to Time-Series Modeling and Forecasting in Business and Economics*. McGraw-Hill Education (ISE Editions).

GEOS-Chem. (2017). *GEOS-Chem*. Retrieved 15/09/2020 from http://acmg.seas.harvard.edu/geos/geos_overview.html

Gerasopoulos, E., Kouvarakis, G., Babasakalis, P., Vrekoussis, M., Putaud, J. P., & Mihalopoulos, N. (2006, 2006/08/01/). Origin and variability of particulate matter (PM10) mass concentrations over the Eastern Mediterranean. *Atmospheric Environment, 40*(25), 4679-4690. https://doi.org/https://doi.org/10.1016/j.atmosenv.2006.04.020

Giechaskiel, B., Maricq, M., Ntziachristos, L., Dardiotis, C., Wang, X., Axmann, H., Bergmann, A., & Schindler, W. (2014, 2014/01/01/). Review of motor vehicle particulate emissions sampling and measurement: From smoke and filter mass to particle number. *Journal of Aerosol Science, 67*, 48-86. https://doi.org/https://doi.org/10.1016/j.jaerosci.2013.09.003

Gimson, N., Chilton, R., & Xie, S. (2010). *Meteorological Datasets for the Auckland Region– User Guide.* A. R. C. T. Report.

Gneiting, T. (2002, 2002/06/01). Nonseparable, Stationary Covariance Functions for Space–Time Data. *Journal of the American Statistical Association, 97*(458), 590-600. https://doi.org/10.1198/016214502760047113

Gocheva-Ilieva, S., & Ivanov, A. (2019). Assaying SARIMA and generalised regularised regression for particulate matter PM10 modelling and forecasting. *International Journal of Environment and Pollution, 66*(1/2/3), 41-62. https://doi.org/10.1504/IJEP.2019.104520

Gocheva-Ilieva, S., Stoyanova, D., Doychin, V., & Ivanov, A. (2014). Time series analysis and forecasting for air pollution in small urban area: An SARIMA and factor analysis approach. *Stochastic Environmental Research and Risk Assessment, 28*(4), 1045-1060.

Goovaerts, P. (1997). Local Estimation: Accounting for a Single Attribute. In *Geostatistics for Natural Resources Evaluation*. Oxford University Press.

Gräler, B., Pebesma, E., & Heuvelink, G. (2016). Spatio-Temporal Interpolation using gstat *The R Journal, 8*(1), 204-218.

Grange, S. K., Carslaw, D. C., Lewis, A. C., Boleti, E., & Hueglin, C. (2018). Random forest meteorological normalisation models for Swiss PM10 trend analysis. *Atmospheric Chemistry and Physics, 1*. https://doi.org/https://doi.org/10.5194/acp-2017-1092

Graves, A., Mohamed, A. R., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. EEE International Conference on Acoustics, Speech and Signal Processing,

Green, P. J., & Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach.* (1st Edition ed.). Chapman and Hall Ltd. https://doi.org/https://doi.org/10.1201/b15710

Griffin, R. J. (2013). The Sources and Impacts of Tropospheric Particulate Matter. *Nature Education Knowledge, 4*(5).

Guarnaccia, C., Cerón Bretón, J. G., Quartieri, J., Tepedino, C., & Cerón Bretón, R. M. (2014). An Application of Time Series Analysis for Forecasting and Control of Carbon Monoxide Concentrations. *INTERNATIONAL JOURNAL OF MATHEMATICAL MODELS AND METHODS IN APPLIED SCIENCES, 8*.

H. Hamed, K., Rao, R., & Chen, H.-L. (2003). *Nonstationarities in Hydrologic and Environmental Time Series*. Springer.

Haenfling, C. (2020). *Particulate matter concentrations (PM10)*. M. U. Environmental Health Intelligence New Zealand.

Hamed, K. H., & Rao, R. (1998). A Modified Mann-Kendall Trend Test for Autocorrelated Data. *Journal of Hydrology 204*, 182-196.

Han, Y., Cao, J., An, Z., Chow, J. C., Watson, J. G., Jin, Z., Fung, K., & Liu, S. (2007, 2007/09/01/). Evaluation of the thermal/optical reflectance method for quantification of elemental carbon in sediments. *Chemosphere, 69*(4), 526-533. https://doi.org/https://doi.org/10.1016/j.chemosphere.2007.03.035

Hanke, J. R., Fischer, M. P., & Pollyea, R. M. (2018). Contents lists available atScienceDirectJournal of Structural Geologyjournal homepage:www.elsevier.com/locate/jsgDirectional semivariogram analysis to identify and rank controls on thespatial variability of fracture networks. *Journal of Structural Geology, 108*, 34-51.

Hannachi, A., Jolliffe, I. T., & Stephenson, D. B. (2007). Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology, 27*(9), 1119-1152. https://doi.org/https://doi.org/10.1002/joc.1499

Harrison, R. M. (2020). Airborne particulate matter. *Phil. Trans. R. Soc. A, 378*(2183). https://doi.org/https://doi.org/10.1098/rsta.2019.0319

Hart, J. E., Yanosky, J. D., Puett, R. C., Ryan, L., Dockery, D. W., Smith, T. J., Garshick, E., & Laden, F. (2009). Spatial Modeling of PM10 and NO2 in the Continental United States, 1985–2000. *Environ Health Perspect, 117*(11), 1690-1696. https://doi.org/10.1289/ehp.0900840

Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Scienecs, 1*(3), 297-318.

Haykin, S. O. (1999). *Neural Networks: A Comprehensive Foundation* (2nd Edition ed.). Prentice-Hall, Upper Saddle river, N.J, 842.

He, S., Mazumdar, S., & Arena, V. C. (2006). A comparative study of the use of GAM and GLM in air pollution research. *ENVIRONMETRICS, 17*, 81–93. https://doi.org/10.1002/env.751

Henebr, G. M. (1994). Spatial model error analysis using autocorrelation indices *Ecological Modelling, 82*, 75-91.

Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and l1 penalized regression: a review. *Statistics Surveys, 2*, 61-93.

Hochreiter, S., & Schmidhuber, J. (1997, Nov 15). Long short-term memory. *Neural Comput, 9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[Record #23 is using a reference type undefined in this output style.]

Hu, X., Belle, J. H., Meng, X., Wildani, A., Waller, L. A., Strickland, M. J., & Liu, Y. (2017, 2017/06/20). Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach. *Environmental Science & Technology, 51*(12), 6936-6944. https://doi.org/10.1021/acs.est.7b01210

Hu, Z. (2008). *Time Series Forecasting Model for Chinese Future Marketing Price of Copper and Aluminum* Georgia State University].

Hughes, D. (1993). *Pan's Travail: Environmental Problems of the Ancient Greeks and Romans (Ancient Society and History)* (1 ed.). The Johns Hopkins University Press.

Hydro-Estimator. (2014). *Global Hydro-Estimator - Algorithm Description*. https://www.ospo.noaa.gov/Products/atmosphere/ghe/algo.html

Hyndman, R. J. (2018). Forecasting complex seasonality. In *Forecasting: principles and practice,2nd edition*. OTexts.com/fpp2.

Ibrahim, M. Z., Zailan, R., Ismail, M., & Safiih Lo, M. (2009). Forecasting and Time Series Analysis of Air Pollutants in Several Area of Malaysia. *American Journal of Environmental Sciences, 5*(5), 625-632.

Janhäll, S. (2015, 2015/03/01/). Review on urban vegetation and particle air pollution – Deposition and dispersion. *Atmospheric Environment, 105*, 130-137. https://doi.org/https://doi.org/10.1016/j.atmosenv.2015.01.052

Jentsch, C., & Subba Rao, S. (2015). A test for second order stationarity of a multivariate time series. *Journal of Econometrics*, 124-161.

Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J., & Shao, D. (2004, 2004/12/01/). Progress in developing an ANN model for air pollution index forecast. *Atmospheric Environment, 38*(40), 7055-7064. https://doi.org/https://doi.org/10.1016/j.atmosenv.2003.10.066

Jolliffe, I. (2011). Principal Component Analysis. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1094-1096). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_455

Kaiser, H. F. (1970, 1970/12/01). A second generation little jiffy. *Psychometrika, 35*(4), 401-415. https://doi.org/10.1007/BF02291817

Kang, E. L., Liu, D., & Cressie, N. (2009). Statistical analysis of small-area data based on independence, spatial, non-hierarchical, and hierarchical models. *Computational Statistics & Data Analysis, 53*, 3016–3032.

Kim, H. S., Park, I., Song, C. H., Lee, K., Yun, J. W., Kim, H. K., Jeon, M., Jiwon, L., & Han, K. M. (2019). Development of a daily PM10 and PM2.5 prediction system using a deep long short-term memory neural network model. *Atmospheric Chemistry and Physics, 19*, 12935–12951. https://doi.org/https://doi.org/10.5194/acp-19-12935-2019

Kingma, D. P., & Ba, J. (2015). *Adam: A Method for Stochastic Optimization* http://arxiv.org/abs/1412.6980

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy, 52*(6), 119-139.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. How sure are we that economic time series have a unit root? *Journal of Econometrics, 54*, 159-178.

Kyriakidis , P. C., & Journel, A. G. (1999). Geostatistical space-time models: A review. *Math. Geol*, 651–684.

Lang, P. E., Carslawa, D. C., & Moller, S. J. (2019). A trend analysis approach for air quality network data. *Atmospheric Environment: X.* https://doi.org/100030

Laongsri, B. (2013). *Studies of the Properties of Particulate Matter in the UK Atmosphere* University of Birmingham, UK].

Lau, J., Hung, W. T., & Cheung, C. S. (2009, 2009/02/01/). Interpretation of air quality in relation to monitoring station's surroundings. *Atmospheric Environment, 43*(4), 769-777. https://doi.org/https://doi.org/10.1016/j.atmosenv.2008.11.008

Lee, E., Chan, C. K., & Paatero, P. (1999, 1999/08/01/). Application of positive matrix factorization in source apportionment of particulate pollutants in Hong Kong. *Atmospheric Environment, 33*(19), 3201-3212. https://doi.org/https://doi.org/10.1016/S1352-2310(99)00113-2

Legates, D., & Maccabe, G. J. (1999). Evaluating the Use Of "Goodness-of-Fit" Measures in Hydrologic and Hydroclimatic Model Validation. *Water Resources Research, 35*(1), 233-241.

Li, G., Sun, J., Jayasinghe, R., Pan, X., Zhou, M., Wang, X., Cai, Y., Sadler, R., & Shaw, G. (2012). Temperature Modifies the Effects of Particulate Matter on Non-Accidental Mortality: A Comparative Study of Beijing, China and Brisbane, Australia. *Public Health Research, 2*(2), 21-27. https://doi.org/10.5923/j.phr.20120202.04

Li, L. (2008). An Application of a Shape Function Based Spatiotemporal Interpolation Method to Ozone and Population-Based Environmental Exposure in the Contiguous U.S. *Journal of Environmental Informatics, 12*(2), 120-128.

Li, L., & Revesz, P. (2004). Interpolation methods for spatio-temporal geographic data. *Computers, Environment and Urban Systems, 28*(3201-227).

Li, L., Romary, T., & Caers, J. (2015, 2015/11/01/). Universal kriging with training images. *Spatial Statistics, 14*, 240-268. https://doi.org/https://doi.org/10.1016/j.spasta.2015.04.004

Li, L., Zhou, X., Kalo, M., & Piltner, R. (2016). Spatiotemporal Interpolation Methods for the Application of Estimating Population Exposure to Fine Particulate Matter in the Contiguous U.S. and a Real-Time Web Application. *Int J Environ Res Public Health.*, 13(18): 749.

Li, Q., Guo, Y., Song, J. Y., Song, Y., Ma, J., & Wang, H. J. (2018). Impact of long-term exposure to local PM10 on children's blood pressure: a Chinese national cross-sectional study. *Air Quality, Atmosphere & Health, 11*(6), 705–713. https://doi.org/10.1007/s11869-018-0577-1

Li, W., Cao, Y., Li, R., Ma, X., Chen, J., Wu, Z., & Xu, Q. (2018). The spatial variation in the effects of air pollution on cardiovascular mortality in Beijing, China. *Journal of Exposure Science and Environmental Epidemiology, 28*, 297–304.

Lin, S. L., & Huang, H. W. (2020, 2020/03/26). Improving Deep Learning for Forecasting Accuracy in Financial Data. *Discrete Dynamics in Nature and Society, 2020*, 5803407. https://doi.org/10.1155/2020/5803407

Lin, X. H., & Zhang, D. W. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 61*, 381-400.

Liu, G. P. (2001). Neural networks. In *Nonlinear Identification and Control: A neural network approach* (1 ed., pp. 210). Springer-Verlag London.

Liu, M., Mroueh, Y., Ross, J., Zhang, W., Cui, X., Das, P., & Yang, T. (2020). *Towards Better Understanding of Adaptive Gradient Algorithms in Generative Adversarial Nets* https://openreview.net/forum?id=SJxIm0VtwH

Liu, X. L., Fu, X. Q., Li, Y., Shen, J. L., Wang, Y., Zou, G. H., Wu, Y. Z., Ma, Q. M., Chen, D., Wang, C., Xiao, R. L., & Wu, J. S. (2016). Spatio-temporal variability in N2O emissions from a tea-planted soil in subtropical central China. *Geoscientific Model Development Discussions*, DOI: 10.5194/gmd-2015-5251.

Longley, I. (2020). *Pollution levels soar in Level 3, says NIWA*. Retrieved 21/10/2020 from https://niwa.co.nz/news/pollution-levels-soar-in-level-3-says-niwa

Mahapatra, P. S., Jain, S., Shrestha, S., Senapati, S., & Puppala, S. P. (2018). Ambient endotoxin in PM10 and association with inflammatory activity, air pollutants, and meteorology, in Chitwan, Nepal. *Science of The Total Environment, 618*, 1331-1342. https://doi.org/https://doi.org/10.1016/j.scitotenv.2017.09.249

Manisalidis, I., Stavropoulou, E., Stavropoulos, A., & Bezirtzoglou, E. (2020). Environmental and Health Impacts of Air Pollution: A Review. *Front Public Health*. https://doi.org/https://doi.org/10.3389/fpubh.2020.00014

Mao, M., Zhang, X., & Yan Yin, Y. (2018). Particulate Matter and Gaseous Pollutions in Three Metropolises along the Chinese Yangtze River: Situation and Implications. *International Journal of Environmental Research and Public Health 15*(1102). https://doi.org/10.3390/ijerph15061102

Marco, L., Ziegler, a., Alexander, D. C., & Ourselin, S. (2015). Modelling Non-Stationary and Non-Separable Spatio-Temporal Changes in Neurodegeneration via Gaussian Process Convolution. *Machine Learning Meets Medical Imaging*, 35-44.

Markham, A. C. (1994). Adam C. Markham. In A. C. Markham (Ed.), *A Brief History of Pollution* (1 ed.). Earthscan Pubs. https://doi.org/https://doi.org/10.4324/9780429344879

Marra, G., & Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis, 55*(7), 2372-2387. https://doi.org/10.1016/j.csda.2011.02.004

Masri, S., Kang, C.-M., & Koutrakis, P. (2015). Composition and Sources of Fine and Coarse Particles Collected during 2002–2010 in Boston, MA. *Air Waste Manag Assoc*, 287-297.

Mateu, J., Porcu, E., & Gregori, P. (2008, 2008/05/01). Recent advances to model anisotropic space–time data. *Statistical Methods and Applications, 17*(2), 209-223. https://doi.org/10.1007/s10260-007-0056-6

McCulloch, C. M., Searle, S. R., & Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models* (2nd Edition ed.).

McMurry, P., Shepherd, M., & Vickery, J. (2004). *Particulate Matter Science for Policy Makers: A NARSTO Assessment* (Vol. ISBN: 0 52 184287 5). Cambridge University Press.

McVey, A., Pernak, R., Hegarty, J., & Alvarado, M. (2018). *El Paso Ozone and PM2.5 Background and Totals Trend Analysis*.

Memon, M. (2018). *Forecasting of Delhi Air Pollution With The Help of Performance Evaluation of Advanced Time Series Models*. N. C. o. Ireland.

MFE, M. f. t. E. (2002). *Ambient Air Quality Guidelines* (Air Quality Report No 32 Issue. M. f. t. Environment.

MFE, M. f. t. E. (2009). *Environmental report card: Air quality (particulate matter – PM10)*.

MFE, M. f. t. E. (2011). *2011 Users' Guide to the revised National Environmental Standards for Air Quality*.

MFE, M. f. t. E. (2012). *Indicator update: Air quality (particulate matter – PM10)*.

MfE, M. f. t. E., & Stats NZ, S. (2018). *New Zealand's Environmental Reporting Series: Our air 2018.* (ME 1384). M. f. t. E. a. S. NZ. https://www.mfe.govt.nz/sites/default/files/media/Air/

Miranda, A. I., Marchi, E., Ferretti, M., & Millán, M. M. (2008). Chapter 9 Forest Fires and Air Quality Issues in Southern Europe. In A. Bytnerowicz, M. J. Arbaugh, A. R. Riebau, & C. Andersen (Eds.), *Developments in Environmental Science* (Vol. 8, pp. 209-231). Elsevier. https://doi.org/https://doi.org/10.1016/S1474-8177(08)00009-0

Miranda, V., & Yee, T. (2018). *Package 'VGAMextra'*. In *Additions and Extensions of the 'VGAM' Package* (Version 0.0-1)

Moliné, J., Bernard, M.-L., Komorowski, J.-C., Euphrasie-Clotilde, L., Brute, F.-N., & Roussas, A. (2014). Particle size distribution and PM10 of volcanic ashes in Guadeloupe during the major eruption of Soufrière Hills in February 2010. *European geosciences union general assembly*, 16357.

Moran, P. A. P. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*(1/2), 17-23.

Mpanza, M. (2015). A Comparison of Oridinary Kriging and Simple Kriging On a PGR Resource in the Esatern Limb of the Bushveled Complex. *Master Thesis*.

Murase, H., Nagashima, H., Yonezaki, S., Matsukura, R., & Kitakado, T. (2009). Application of a generalized additive model (GAM) to reveal relationships between environmental factors and distributions of pelagic fish and krill: a case study in Sendai Bay, Japan. *CES Journal of Marine Science, 66*(6), 1417–1424.

Myers, D. E. (2004). Estimating and Modeling Space-Time Variograms. Proceedings of the joint meeting of The 6th International Symposium On Spatial Accuracy Assessment In Natural Resources and Environmental Sciences and The 15th Annual Conference of The International Environmetrics Society, Portland, Maine, USA.

Myers, D. E. (2004). stimating and modeling space-time variograms. McRoberts R (ed) Proceedings of the joint meeting of TIES-2004 and ACCURACY-2004,

Myers, D. E., & Journel, A. (1999). Variograms with Zonal Anisotropies and Non-Invertible Kriging Systems. *Mathematical Geology*, 779-785.

Naveen, V., & Anu, N. (2017). Time Series Analysis to Forecast Air Quality Indices in Thiruvananthapuram District, Kerala, India. *Int. Journal of Engineering Research and Application 7*(6), 66-84.

NOAA. (2014). *National Oceanic and Atmospheric Administration*. https://www.noaa.gov/

Okkaoğlua, Y., Akdi, Y., & DemirberkÜnlü, K. (2020). Daily PM10, periodicity and harmonic regression model: The case of London. *Atmospheric Environment, 238*. https://doi.org/https://doi.org/10.1016/j.atmosenv.2020.117755

Olden, J. D., Joy, M. K., & Death, R. G. (2004, 2004/11/01/). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling, 178*(3), 389-397. https://doi.org/https://doi.org/10.1016/j.ecolmodel.2004.03.013

Oliver, M. A., & Webster, R. (2015). *Basic Steps in Geostatistics: The Variogram and Kriging*. SpringerBriefs in Agriculture.

Onyutha, C. (2016). Statistical Uncertainty in Hydrometeorological Trend Analyses. *Advances in Meteorology*, http://dx.doi.org/10.1155/2016/8701617.

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics, 35*, 526-528.

Pascanu, R., Gulcehre, C., Cho, K., & Bengio, Y. (2014). *How to Construct Deep Recurrent Neural Networks*

Patel, H., Talbot, N., Salmond, J., Dirks, K., Xie, S., & Davy, P. (2020). Implications for air quality management of changes in air quality during lockdown in Auckland (New Zealand) in response to the 2020 SARS-CoV-2 epidemic. *Science of The Total Environment, 746*. https://doi.org/https://doi.org/10.1016/j.scitotenv.2020.141129

PCE, P. C. f. t. E. (2015). *The state of air quality in New Zealand.*

Pearce, J. L., Beringer, J., Nicholls, N., Hyndman, R. J., & Tapper, N. J. (2011). Quantifying the influence of local meteorology on air quality using generalized additive models. *Atmospheric Environment 45*, 1328-1336.

Pearce, J. L., Beringer, J., Nicholls, N., Hyndman, R. J., Uotila, P., & Tapper, N. J. (2011, 2011/01/01/). Investigating the influence of synoptic-scale meteorology on air quality using self-organizing maps and generalized additive modelling. *Atmospheric Environment, 45*(1), 128-136. https://doi.org/https://doi.org/10.1016/j.atmosenv.2010.09.032

Pedregal, D. J., & Young, P. C. (2008). Development of Improved Adaptive Approaches to Electricity Demand Forecasting. *Journal of the Operational Research Society 59*(8), 1066–1076.

Pfaff, B., Zivot, E., & Stigler, M. (2016). *Package 'urca'*. In *Unit Root and Cointegration Tests for Time Series Data* (Version 1.3-0) [Package]. CRAN.

Pohlert, T. (2018). trend: Non-Parametric Trend Tests and Change-Point Detection. In *R package version 1.1.1*. https://CRAN.R-project.org/package=trend.

Querol, X., Pérez, N., Reche, C., Ealo, M., Ripoll, A., Tur, J., Pandolfi, M., Pey, J., Salvador, P., Moreno, T., & Alastuey, A. (2019, 2019/10/10/). African dust and air quality over Spain: Is it only dust that matters? *Science of The Total Environment, 686*, 737-752. https://doi.org/https://doi.org/10.1016/j.scitotenv.2019.05.349

Ramsay, T. O., Burnett, R. T., & Krewski, D. (2003). The Effect of Concurvity in Generalized Additive Models Linking Mortality to Ambient Particulate Matter. *Epidemiology, 14*(1), 18-23. https://www.jstor.org/stable/3703274 (Lippincott Williams & Wilkins)

RESSTE, R. N. (2017). Spatio-temporal data with R *Journal de la Société Française de Statistique, 158*(3), 124-158.

Reyes, I. V., Fedyushkina, I. V., Skvortsov, V. S., & Filimonov, D. (2013). Prediction of progesterone receptor inhibition b y high"performance neural network algorithm.

Rosato, A., Panella , M., Araneo , R., & Andreotti, A. (2019). A Neural Network Based Prediction System of Distributed Generation for the Management of Microgrids. *IEEE TRANSACTIONS ON INDUSTRY APPLICATIONS, 55*(6), 7092-7012. https://doi.org/ 10.1109/TIA.2019.2916758

Rutkowska, A. (2014). roperties of the Cox–Stuart Test for Trend in Application to Hydrological Series: The Simulation Study. *Communications in Statistics - Simulation and Computation* 565-579.

Sagheer, A., & Kotb, M. (2019). Unsupervised Pre-training of a Deep LSTM-based Stacked Autoencoder for Multivariate Time Series Forecasting Problems. *Scientific Reports, 9*. https://doi.org/https://doi.org/10.1038/s41598-019-55320-6

Said, S. E., & Dickey, D. A. (1984). Testing for Unit Roots in Autoregressive-Moving Average Models of Unknown Order. *Biometrika, 71*(3), 599-607. https://doi.org/10.2307/2336570

Salmond, J. A., N. Dirks, K. N., Fiddes, S., Pezza, A., Talbot, N., Scarfe, J., Renwick, J., & Petersen, J. (2015). A climatological analysis of the incidence of brown haze in Auckland, New Zealand. *Intrnational Journal of climatology*. https://doi.org/ https://doi.org/10.1002/joc.4509

Salomon, V., & Smithson, J. (2015). *Inventory of emissions to air in the Canterbury airsheds – industrial and commercial emissions– 2014 update*.

Sayegh, A., Said Munir, & Habeebullah, T. M. (2014). Comparing the Performance of Statistical Models for Predicting PM10 Concentrations. *Aerosol and Air Quality Research, 14*, 653–665.

Scarrott, C., Reale, M., & Newell, J. (2009). *Statistical estimation and testing of trends in PM10 concentrations: Is Christchurch city likely to meet the NES target for PM10 concentrations in 2013?*

Schlink, U., Dorling, S., Pelikan, E., Nunnari, G., Cawley, G., Junninen, H., Greig, A., Foxall, R., Eben, K., Chatterton, T., Vondracek, J., Richter, M., Dostal, M., Bertucco, L., Kolehmainen, M., & Doyle, M. (2003). A Rigorous Inter-Comparison of Ground-Level Ozone Predictions. *Atmospheric Environment, 37*(23), 3237-3253. https://doi.org/10.1016/S1352-2310(03)00330-3

Seinfeld, J. H., & Pandis, S. N. (1998). *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. J. Wiley.

Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. *he American Statistical Association, 63*(324), 1379–1389.

Senaratne, I., & Shooter, D. (2004, 2004/06/01/). Elemental composition in source identification of brown haze in Auckland, New Zealand. *Atmospheric Environment, 38*(19), 3049-3059. https://doi.org/https://doi.org/10.1016/j.atmosenv.2004.02.046

Shuang, L., Zhai, L., Zou, B., Sang, H., & Fang, X. (2017). A Generalized Additive Model Combining Principal Component Analysis for PM2.5 Concentration Estimation. *International Journal of Geo-Information*.

Silva, I. N., Spatti, D. H., Flauzino, R. A., Liboni, L. H. B., & Alves, S. F. (2017). Artificial Neural Network Architectures and Training Processes. In I. N. Silva, D. H. Spatti, R. A. Flauzino, L. H. B. Liboni, & S. F. Alves (Eds.), *Artificial Neural Networks A Practical Course* (1 ed., pp. 21-28). Springer International Publishing.

Silva, L. T., & Mendes, J. F. G. (2011). A New Air Quality Index for Cities. In F. Nejadkoorki (Ed.), *Advanced Air Pollution*. https://doi.org/DOI: 10.5772/710

Skøien, J. O., & Blöschl, G. (2005). Spatio-temporal geostatistical analyses of runoff and precipitation. In P. Renard, H. Demougeot-Renard & R. Froidevaux (Eds.). Geostatistics for Environmental ApplicationsProceedings of the Fifth European Conference on goeostatistics for environmental applications, Berlin.

Smith, M. J., Goodchild, M. F., & Longley, P. A. (2007). Geostatistical Interpolation Methods. In *Geospatial Analysis: A Comprehensive Guide* (Second Edition ed.). Troubador Publishing Ltd.

Spadavecchia, L., & Williams, M. (2009, 2009/06/15/). Can spatio-temporal geostatistical methods improve high resolution regionalisation of meteorological variables? *Agricultural and Forest Meteorology, 149*(6), 1105-1117. https://doi.org/https://doi.org/10.1016/j.agrformet.2009.01.008

Sramka, M., Slovak, M., Tuckova, J., & Stodulka, P. (2019). Improving clinical refractive results of cataract surgery by machine learning. *PeerJ, 7*, e7202-e7202. https://doi.org/10.7717/peerj.7202

Sridhar, S., & Metcalfe, J. (2018). *Auckland Air Emissions Inventory 2016 –Transport (Revised)*

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., Hoogh, K. D., Donato, F. D., Gariazzo, C., Lyapustin, A., Michelozzi, P., Renzi, M., Scortichini, M., Shtein, A., ViegiItai, G., & Schwartz, K. (2019). Estimation of daily PM10and PM2.5concentrations in Italy, 2013–2015,using a spatiotemporal land-use random-forest model. *Environment International, 124*, 170-179. https://doi.org/https://doi.org/10.1016/j.envint.2019.01.016

Stafoggia, M., Johansson, C., Glantz, P., Renzi, M., Shtein, A., Hoogh, K. D., Kloog, I., Davoli, M., Michelozzi, P., & Bellander, T. (2020). A Random Forest Approach to Estimate Daily Particulate Matter, Nitrogen Dioxide, and Ozone at Fine Spatial Resolution in Sweden. *Atmosphere, 11*(239). https://doi.org/doi:10.3390/atmos11030239

STAR. (2014). *Hydro-Estimator - Digital Global Data Archive*. https://www.star.nesdis.noaa.gov/smcd/emb/ff/digGlobalData.php

Stark, H., & Woods, J. W. (2012). Random Processes. In *Probability, Statistics, and Random Variables for Engineers*. Pearson Education Inc.

Stats. (2019). Tren Assessment-Technial Information. *stat*. Retrieved 01 22, from http://archive.stats.govt.nz/browse_for_stats/environment/environmental-reporting-series/environmental-indicators/Home/About/trend-assessment-technical-information.aspx

StatSoft, I. (2013). How To Identify Patterns in Time Series Data: Time Series Analysis. In *Electronic Statistics Textbook*. . StatSoft. http://www.statsoft.com/textbook/

Steinbach, M., Karypis, G., & Kumar, V. (2000). A comparison of document clustering techniques. *KDD Workshop on Text Mining*.

Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society. Series B (Methodological), 36*(2), 111-147.

Sumesh, R. K., Rajeevan , K., Resmi, E. A., & Unnikrishnan, C. K. (2017). Particulate Matter Concentrations in the Southern Tip of India: Temporal Variation, Meteorological Influences, and Source Identification. *Earth Systems and Environment 13*.

Sun, W., & Sun, J. (2017). Prediction of carbon dioxide emissions based on principal component analysis with regularized extreme learning machine: The case of China. *Environmental and Engineers Research, 22*(3), 302-311. https://doi.org/https://doi.org/10.4491/eer.2016.153

Sun, Z., & Zhu, D. (2019). Exposure to outdoor air pollution and its human health outcomes: A scoping review. *PLoS ONE, 14*(5). https://doi.org/https://doi.org/10.1371/journal.pone.0216550

Suzuki, N. M., & Taylor, B. (2003, 10 20). *Particulate matter in BC: a report on PM10 and PM2.5 mass concentrations up to 2000*. British Columbia Ministry of Water, Land and Air Protection, and the Pacific and Yukon Region of Environment Canada.

Talbot, N., Reid, N., & Crimmins, P. (2017). *Auckland Ambient Air Quality Trends for PM2.5 and PM10 – 2006-2015.*

Taufik, M. R., Rosanti, E., Eka Prasetya, T. A., & Wijayanti Septiarini, T. (2020, 2020/03). Prediction algorithms to forecast air pollution in Delhi India on a decade. *Journal of Physics: Conference Series, 1511*, 012052. https://doi.org/10.1088/1742-6596/1511/1/012052

Theil, H. (1950). A rank-invariant method of linear and polynomial regression analysis. *Nederlandse Akademie Wetenchappen, A*(53), 386–392.

Thomson, D. J. (1994, 19-22 April 1994). Jackknifing multiple-window spectra. Proceedings of ICASSP '94. IEEE International Conference on Acoustics, Speech and Signal Processing,

Thorsteinsson, T., Jóhannsson, T., Stohl, A., & Kristiansen, N. I. (2012). High levels of particulate matter in Iceland due to direct ash emissions by the Eyjafjallajökull eruption and resuspension of deposited ash. *Journal of Geophysical Research Atmospheres, 117*(B9).

Tobias, S., & Carlson, J. E. (1969, 1969/07/01). BRIEF REPORT: BARTLETT'S TEST OF SPHERICITY AND CHANCE FINDINGS IN FACTOR ANALYSIS. *Multivariate Behavioral Research, 4*(3), 375-377. https://doi.org/10.1207/s15327906mbr0403_8

Tobler, W. R. (1970). A Computer Movie Simulating Urban Growth in the Detroit Region. Economic Geography International Geographical Union. Commission on Quantitative Methods,

Valencia, R., Sanchez, G., & Diaz, I. (2006). A general regression neural network for modeling the behavior of pm10 concentration level in santa marta, colombia. *ARPN Journal of Engineering and Applied Sciences, 11*(11), 7085-7092.

Van Zoest, V., Osei, F. B., Hoek, G., & Stein, A. (2020, 2020/05/03). Spatio-temporal regression kriging for modelling urban NO2 concentrations. *International Journal of Geographical Information Science, 34*(5), 851-865. https://doi.org/10.1080/13658816.2019.1667501

Wang, S. (2012). *Online Monitoring and Prediction of Complex Time Series Events from Nonstationary Time Series Data* The State University of New Jersey].

Wang, S., & Chaovalitwongse, W. (2011). Evaluating and Comparing Forecasting Models. In *Wiley Encyclopedia of Operations Research and Management Science*. John Wiley & Sons, Inc. https://doi.org/10.1002/9780470400531.eorms0307

Wang, W. (2006). *Stochasticity, Nonlinearity and Forecasting of Streamflow Processes.* Nanjing University, China].

Wikle, C., Zammit-Mangion, A., & Cressie, N. (2019). Descriptive Spatio-Temporal Statistical Models. In *Spatio-Temporal Statistics with R*. Taylor & Francis Group.

Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019a). Descriptive Spatio-Temporal Statistical Models. In *Spatio-Temporal Statistics with R*. Chapman & Hall/CRC.

Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019b). Goals of Spatio-Temporal Statistics. In *Spatio-Temporal Statistics with R*. Chapman & Hall/CRC.

Wikle, C. K., Zammit-Mangion, A., & Cressie, N. (2019c). Spatio-Temporal Statistical Models. In *Spatio-Temporal Statistics with R* (pp. 79-80). Chapman & Hall/CRC.

Willmott, C. J., Robeson, S. M., & Matsuura, K. (2011). A refined index of model performance. *International Journal of Climatology, 32*(13), 2088-2094. https://doi.org/ https://doi.org/10.1002/joc.2419

Wilton, E. (2015). *Napier, Hastings and Havelock North Air Emission Inventory*.

Wilton, E., & Caldwell, J. (2018). *Ambient air quality monitoring report for the Waikato region - 2016* (9793067).

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Chapter 10 - Deep learning. In I. H. Witten, E. Frank, M. A. Hall, & C. J. Pal (Eds.), *Data Mining (Fourth Edition)* (pp. 417-466). Morgan Kaufmann. https://doi.org/https://doi.org/10.1016/B978-0-12-804291-5.00010-6

Wood, S. N. (2013). GAM concurvity measures. *mgcv R package version 1.8-28*.

Wood, S. N. (2017a). GAMs IN PRACTICE: mgcv. In *Generalized Additive Models : An Introduction with R* (2 ed., pp. 325-403). CRC Press LLC.

Wood, S. N. (2017b). *Generalized Additive Models: An Introduction with R, Second Edition*. Chapman and Hall/CRC.

Wood, S. N. (2017c). Introducing GAMs. In *Generalized additive models : an introduction with R*. Boca Raton, Florida ; London, [England] ; New York : CRC Press.

Wood, S. N. (2018). GAM concurvity measures. *R Documentation*. Retrieved 11 02, from https://stat.ethz.ch/R-manual/R-devel/library/mgcv/html/concurvity.html

Wood, S. N. (2018). Generalized Additive Model Selection. *R Documentation*. Retrieved 02 18, from https://www.rdocumentation.org/packages/mgcv/versions/1.8-27/topics/gam

Wood, S. N., Pya, N., & Säfken, B. (2016, 2016/10/01). Smoothing Parameter and Model Selection for General Smooth Models. *Journal of the American Statistical Association, 111*(516), 1548-1563. https://doi.org/10.1080/01621459.2016.1180986

Wu, X., Wang, Y., He, S., & Wu, Z. (2020). PM2.5 / PM10 ratio prediction based on a long short-term memory neural network in Wuhan, China. *Geoscientific Model Development (GMD), 13*, 1499–1511. https://doi.org/https://doi.org/10.5194/gmd-13-1499-2020

Wu, Z. Z., & Zhang, S. (2019). Study on the spatial–temporal change characteristics and influence factors of fog and haze pollution based on GAM. *Neural Computing and Applications, 31*, 1619–1631.

Xayasouk, T., Lee, H., & Lee, G. (2020). Air Pollution Prediction Using Long Short-Term Memory (LSTM) and Deep Autoencoder (DAE) Models. *Sustainability 12(6), 12*(6). https://doi.org/10.3390/su12062570

Yadav, A., Jha, C. K., & Sharan, A. (2020, 2020/01/01/). Optimizing LSTM for time series prediction in Indian stock market. *Procedia Computer Science, 167*, 2091-2100. https://doi.org/https://doi.org/10.1016/j.procs.2020.03.257

Yanosky, J. D., Paciorek, C. J., Laden, F., Hart, J. E., Puett, R. C., Liao, D., & Suh, H. H. (2014, 2014/08/05). Spatio-temporal modeling of particulate air pollution in the conterminous United States using geographic and meteorological predictors. *Environmental Health, 13*(1), 63. https://doi.org/10.1186/1476-069X-13-63

Ye, L., Zhou, J., Zeng, X., & Tayyab, M. (2015). Hydrological Mann-Kendal Multivariate Trends Analysis in the Upper Yangtze. *Geoscience and Environment Protection*, 34-39.

Yu, H., & Wang, C. (2013). Quantile-Based Bayesian Maximum Entropy Approach for Spatiotemporal Modeling of Ambient Air Quality Levels. *Environmental Sciences Technology, 47*(3), 1416-1424.

Yue, S., Pilon, P., Phinney, B., & Cavadias, G. (2002). The Influence of Autocorrelation on the Ability to Detect Trend in Hydrological Series. *Hydrological Processes 16*, 1807–1829. https://doi.org/hyp.1095

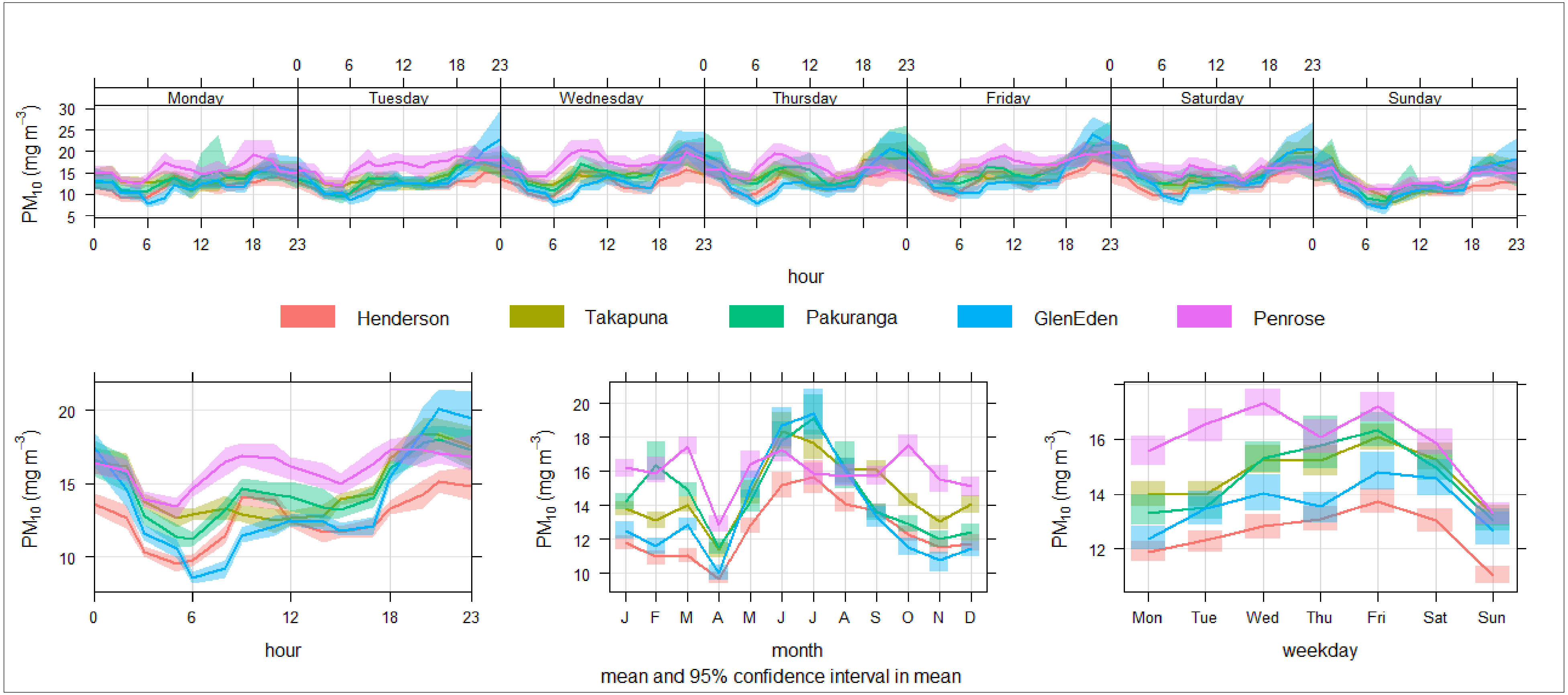
Zamo, M., & Naveau, P. (2018). Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts. *Mathematical Geosciences, 50*, 209–234.

Zeileis, A., & Hothorn, T. (2002). Diagnostic Checking in Regression Relationships. *R News, 3*(2), 7-10. https://CRAN.R-project.org/doc/Rnews/

Zhang, H., Zhang, S., Wang, P., Qin, Y., & Wang, H. (2017). Forecasting of PM 10 time series using wavelet analysis and wavelet-ARMA model in Taiyuan, China. *Journal of the Air & Waste Management Association, 67*(7), 776-788. https://doi.org/https://doi.org/10.1080/10962247.2017.1292968

Appendix A

1) Temporal variation of PM_{10} for all urban sites



Appendix B

1) Harmonic Regression Model Details

Glen Eden

```
Series: msts_train

Regression with ARIMA(1,0,2) errors

Box Cox transformation: lambda= -0.9999242

Coefficients:

         ar1      ma1      ma2  intercept     S1-7    C1-7   S2-
7

      0.8180  -0.3723  -0.2517     0.9135  -0.0011  0.0023  9e-0
4

s.e.  0.0569   0.0636   0.0396     0.0015   0.0012  0.0012  1e-0
3

       C2-7   S3-7   C3-7   S1-365   C1-365

      0.000  9e-04  8e-04  -0.0036  -0.0167

s.e.  0.001  7e-04  7e-04   0.0021   0.0021


sigma^2 estimated as 0.00095:  log likelihood=3768.49

AIC=-7510.98   AICc=-7510.78   BIC=-7439.36
```

Henderson

```
Series: msts_train

Regression with ARIMA(1,0,2) errors

Box Cox transformation: lambda= 0.1355062

Coefficients:

         ar1      ma1      ma2  intercept     S1-7    C1-7    S2
-7

      0.7563  -0.2306  -0.2266     3.0280  -0.0510  0.0218  0.01
64
```

```
s.e.  0.0673   0.0753   0.0479     0.0204   0.0162  0.0162  0.01
18
        C2-7    S3-7     C3-7  S1-365   C1-365
      0.0167  0.0200  -0.0036  0.0166  -0.1572
s.e.  0.0119  0.0081   0.0081  0.0289   0.0288


sigma^2 estimated as 0.1547:  log likelihood=-880.93
AIC=1787.86   AICc=1788.06   BIC=1859.48
```

Pakuranga

```
Series: msts_train
Regression with ARIMA(3,0,1) errors
Box Cox transformation: lambda= -0.9498489
Coefficients:
         ar1      ar2     ar3      ma1  intercept     S1-7    C1
-7
      1.2442  -0.4435  0.0950  -0.7438     0.9603  -0.0030  0.00
19
s.e.  0.0969   0.0556  0.0243   0.0955     0.0017   0.0012  0.00
12
       S2-7    C2-7    S3-7    C3-7   S1-365   C1-365
      9e-04  0.0019  0.0017  -7e-04  -0.0021  -0.0116
s.e.  9e-04  0.0009  0.0006   6e-04   0.0024   0.0024


sigma^2 estimated as 0.0008656:  log likelihood=3853.97
AIC=-7679.93   AICc=-7679.7   BIC=-7602.8
```

Patumahoe

```
Series: msts_train
Regression with ARIMA(3,0,0) errors
```

```
Box Cox transformation: lambda= 1.745064

Coefficients:

         ar1      ar2     ar3  intercept     S1-7    C1-7    S2-
7

      0.5027  -0.0337  0.0737    45.5444  -1.8453  0.6573  0.769
2

s.e.  0.0235   0.0263  0.0235     1.4922   1.1465  1.1464  0.856
0

         C2-7    S3-7     C3-7  S1-365  C1-365

      -0.2532  1.2164  -1.1790  1.0747  8.1250

s.e.   0.8564  0.6337   0.6343  2.1103  2.1075


sigma^2 estimated as 857:  log likelihood=-8751

AIC=17528   AICc=17528.2   BIC=17599.63
```

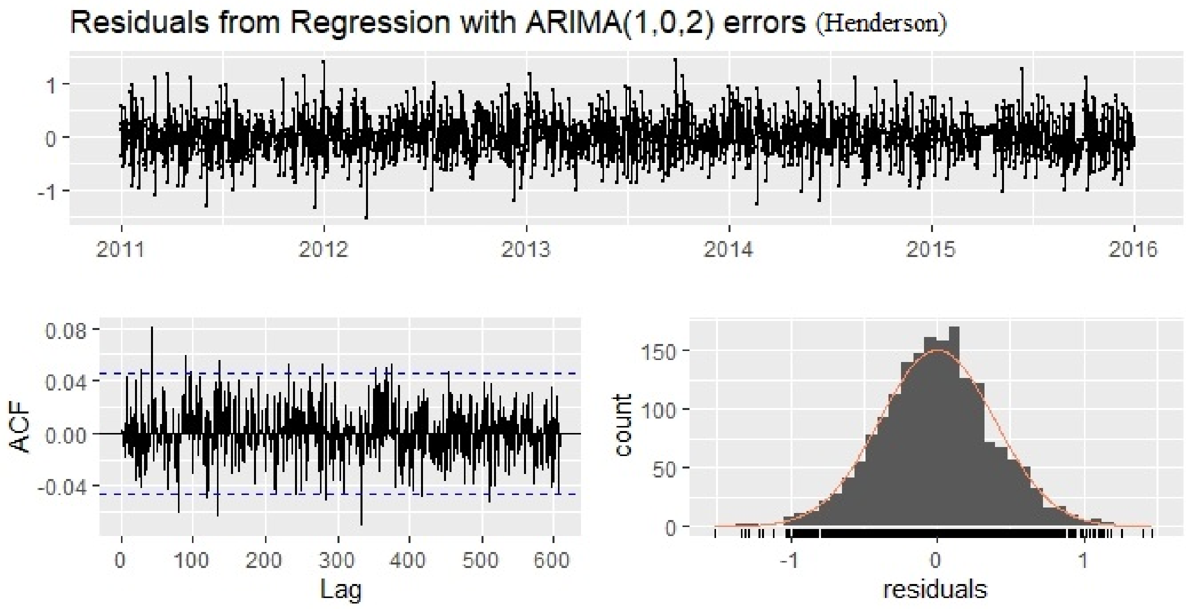
Penrose

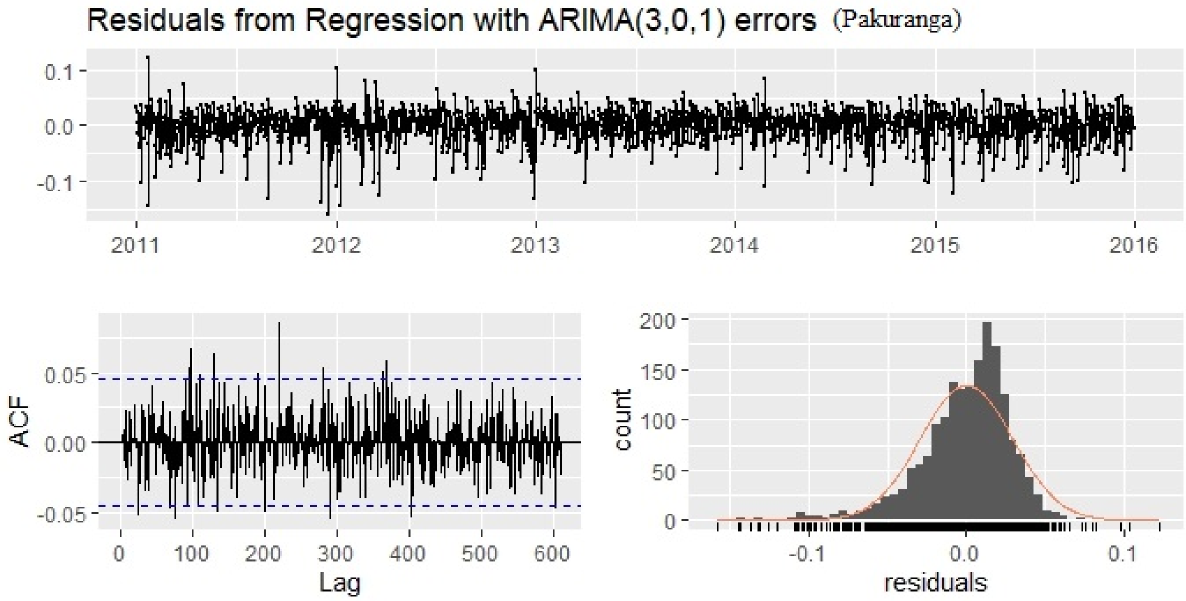
```
Series: msts_train

Regression with ARIMA(3,0,2) errors

Box Cox transformation: lambda= 1.226444

Coefficients:

         ar1      ar2      ar3      ma1      ma2  intercept
S1-7

      1.1517  -0.1967  -0.0147  -0.7004  -0.1647    22.7266  -1.
7708

s.e.  0.5381   0.7526   0.2265   0.5369   0.5168     0.4626   0.
3660

        C1-7    S2-7    C2-7    S3-7     C3-7  S1-365   C1-365

      0.0568  0.0195  0.9996  0.6025  -0.4426  0.5780  -1.5193

s.e.  0.3661  0.2562  0.2564  0.1988   0.1990  0.6485   0.6457


sigma^2 estimated as 77.82:  log likelihood=-6559.61

AIC=13149.22   AICc=13149.49   BIC=13231.87
```

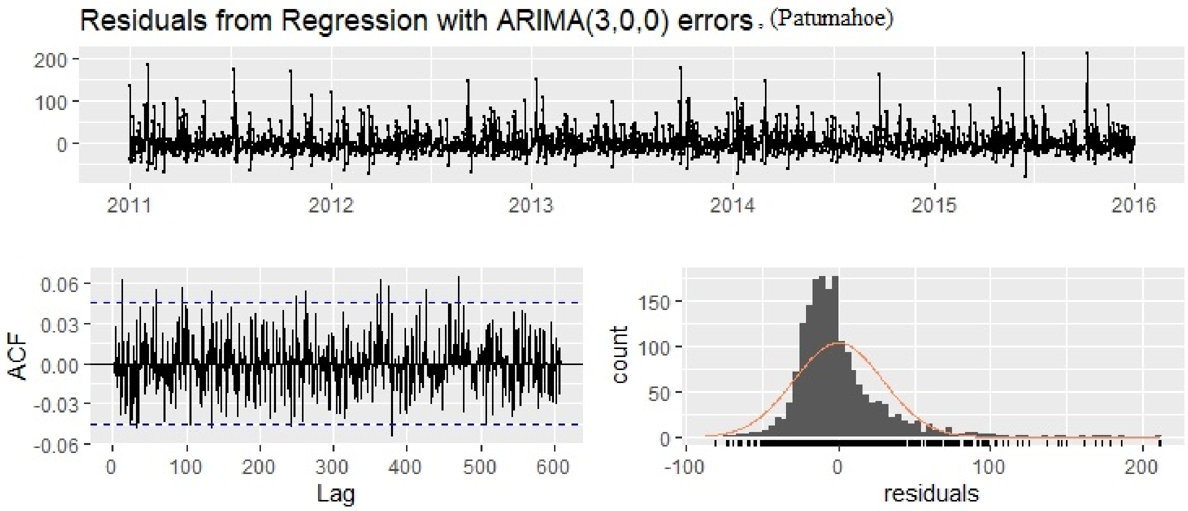
Takapuna

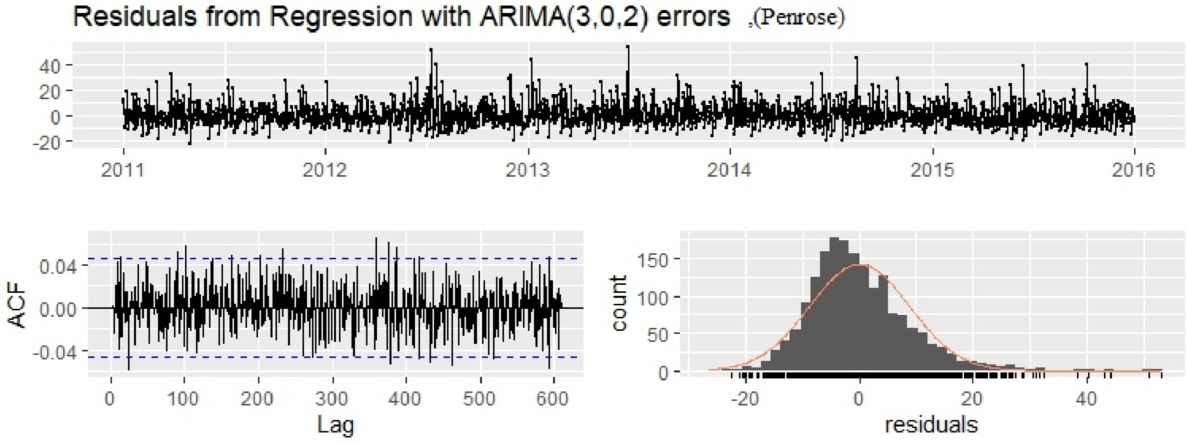
```
Series: msts_train
Regression with ARIMA(5,1,0) errors
Box Cox transformation: lambda= -0.0861932


Coefficients:
          ar1      ar2      ar3      ar4      ar5     S1-7    C1
-7
      -0.4148  -0.4287  -0.3184  -0.1961  -0.1286  -0.0506  0.00
40
s.e.   0.0232   0.0248   0.0257   0.0247   0.0232   0.0097  0.00
97
        S2-7    C2-7    S3-7    C3-7   S1-365   C1-365
      0.0034  0.0221  0.0125  0.0001  -0.0016  -0.0544
s.e.  0.0069  0.0069  0.0048  0.0048   0.1880   0.1876


sigma^2 estimated as 0.05918:  log likelihood=-3.53
AIC=35.06   AICc=35.29   BIC=112.19
```

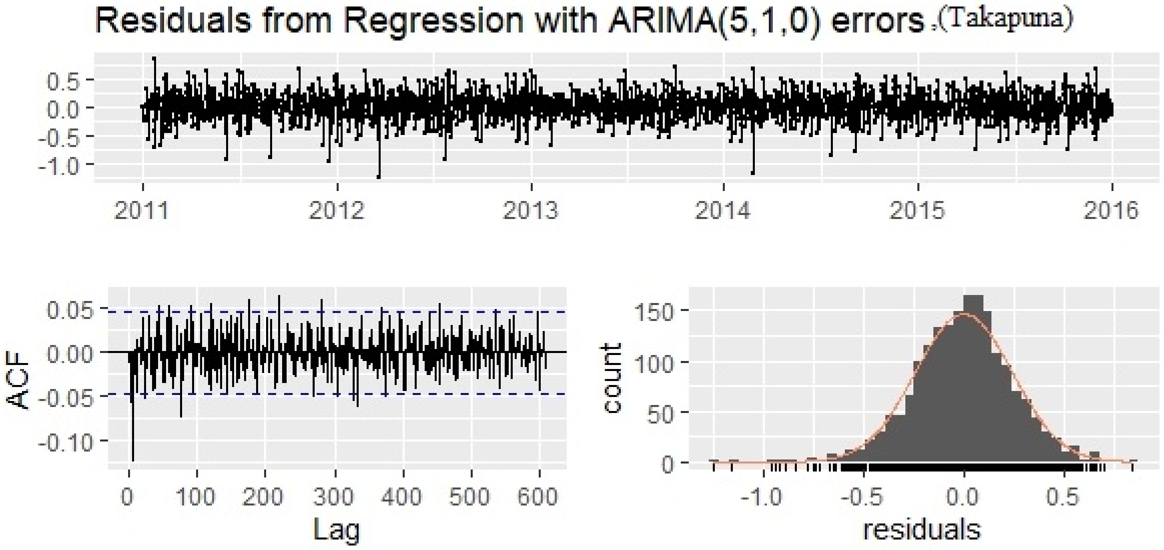
2) Plots of Residuals from Obtained Harmonic Regression Models


Residuals from Regression with ARIMA(1,0,2) errors (Henderson)

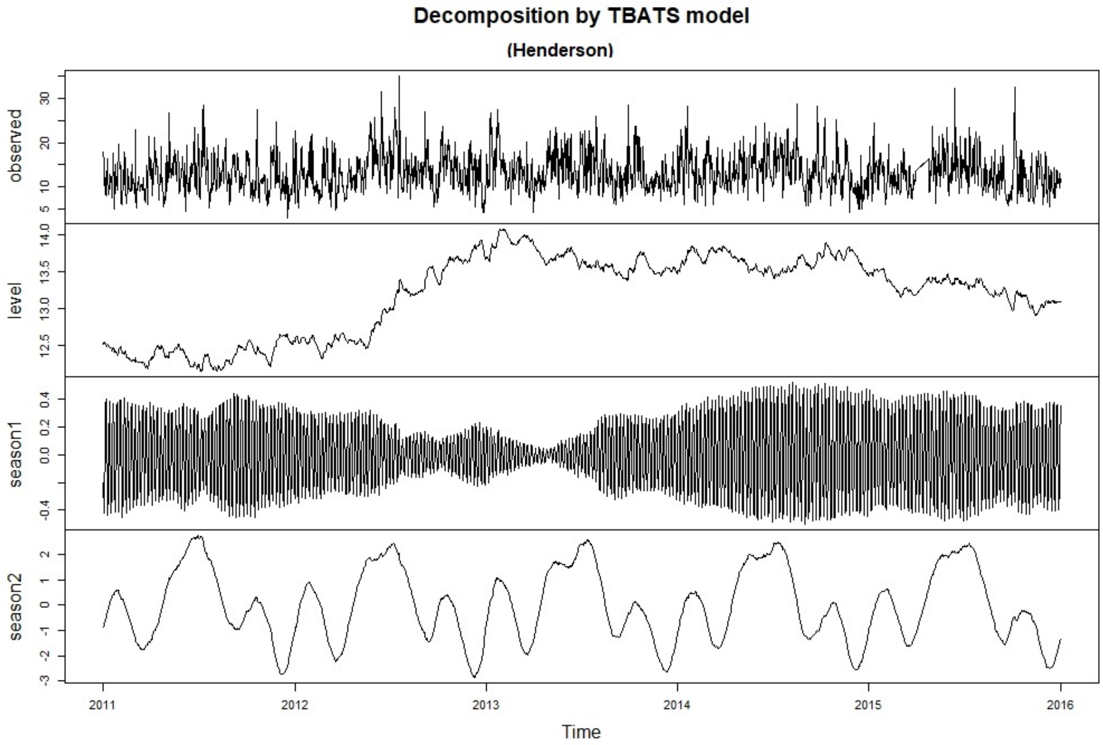

Residuals from Regression with ARIMA(3,0,1) errors (Pakuranga)

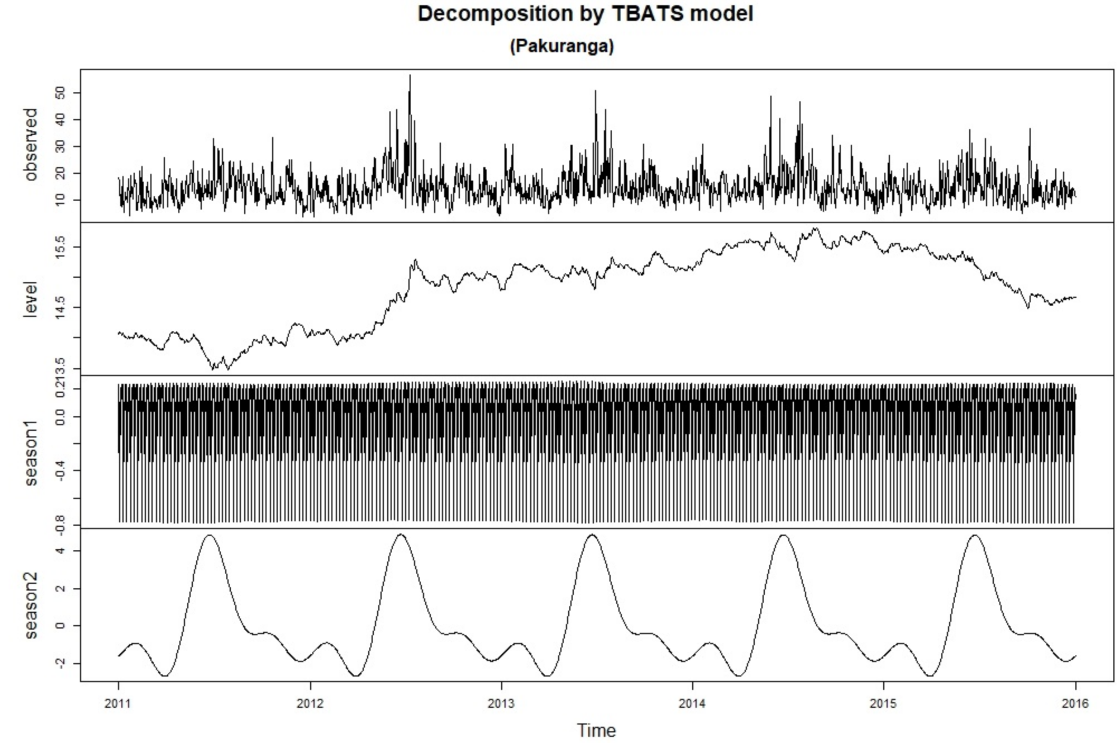


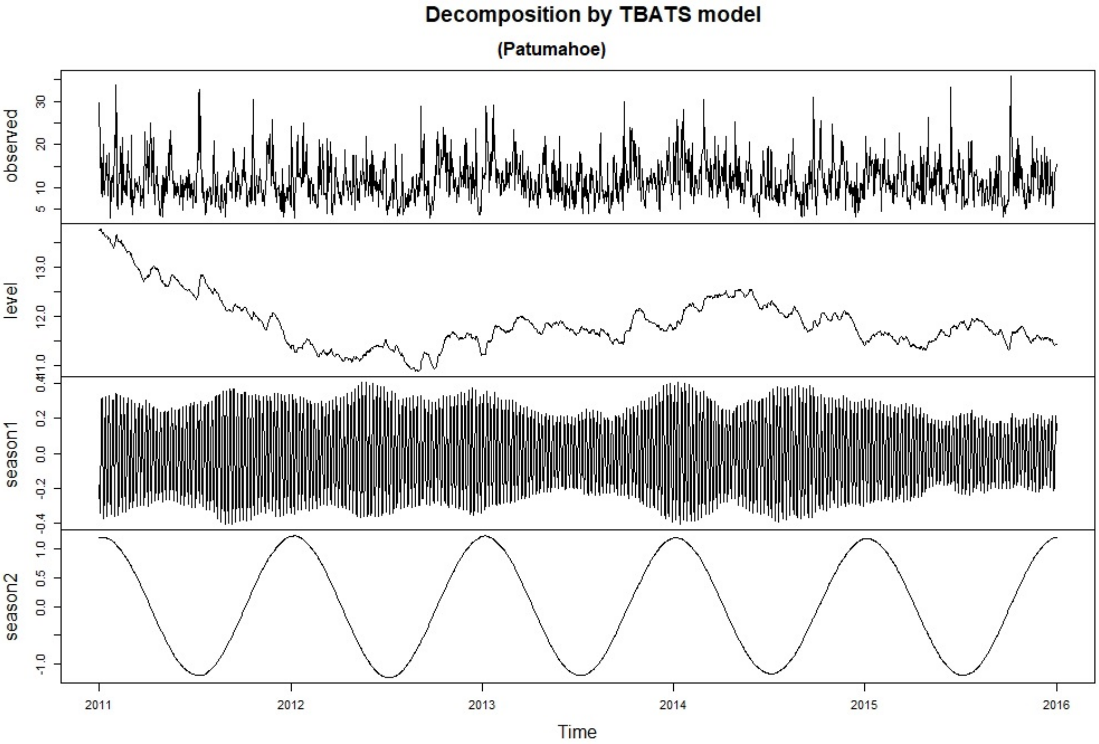


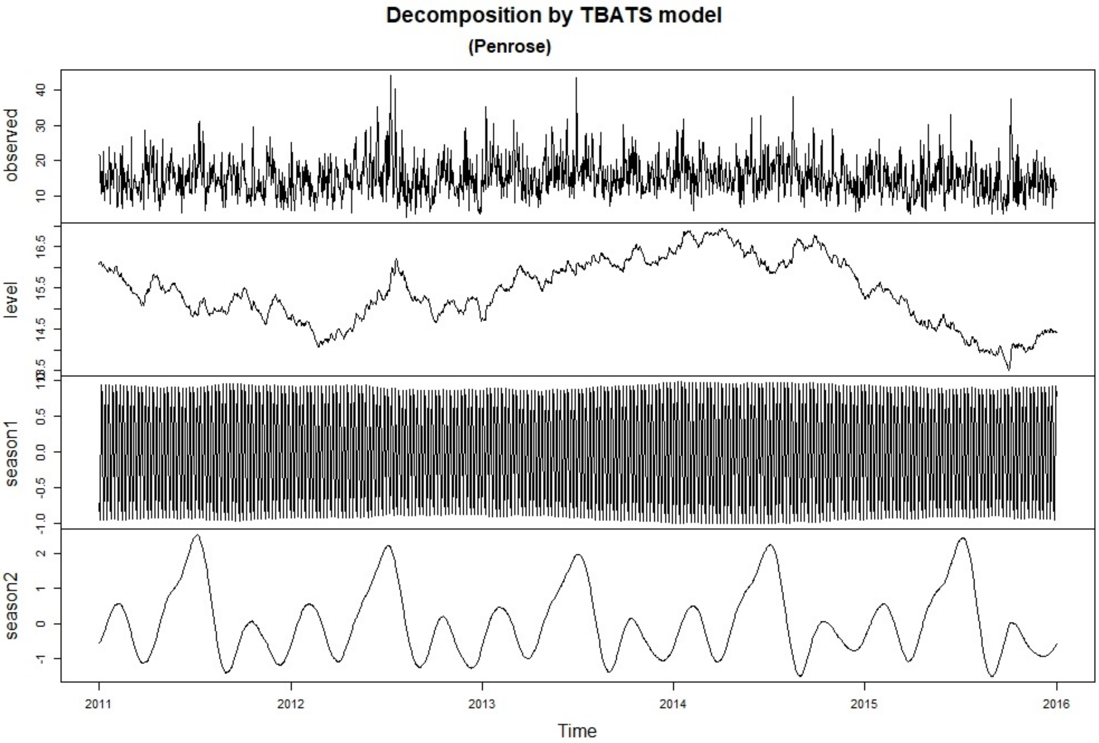


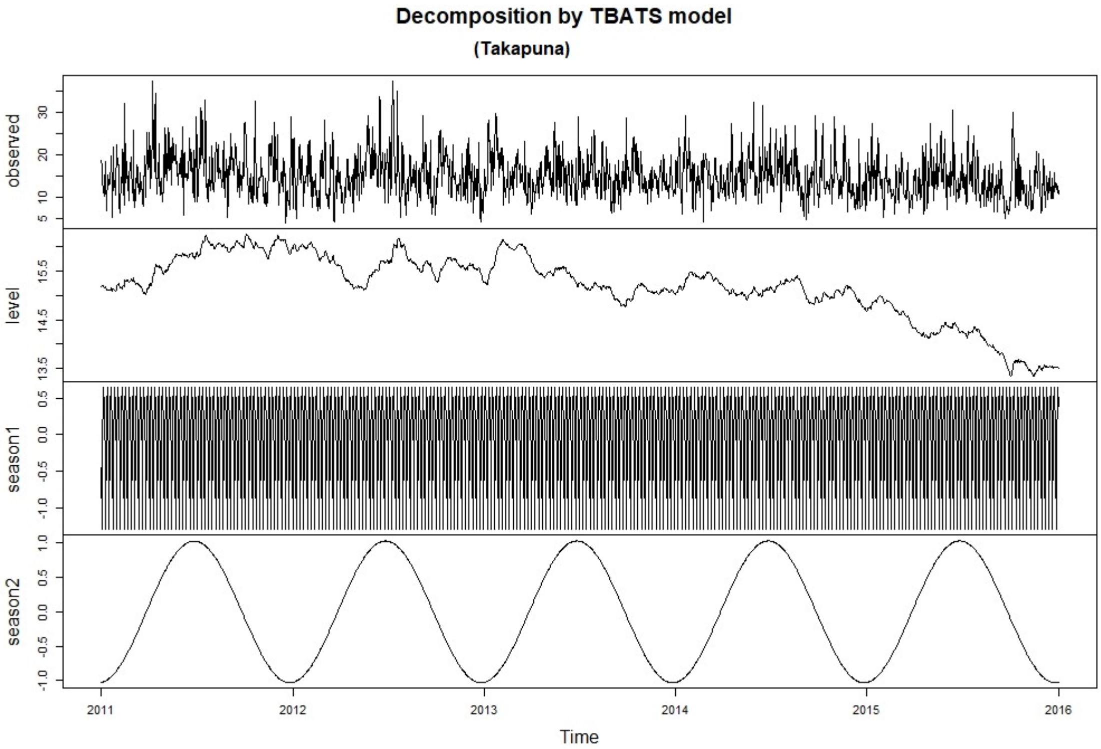
3) TBATS Decomposition Plots


Decomposition by TBATS model
(Henderson)
observed
5
10
20
30
level
12.5
13.0
13.5
14.0
season1
-0.4
0.0
0.2
0.4
season2
-3
-2
-1
0
1
2
2011
2012
2013
2014
2015
2016
Time


Decomposition by TBATS model
(Pakuranga)
observed
10
20
30
40
50
level
13.5
14.5
15.5
season1
-0.8
-0.4
0.0
0.2
season2
-2
0
2
4
2011
2012
2013
2014
2015
2016
Time

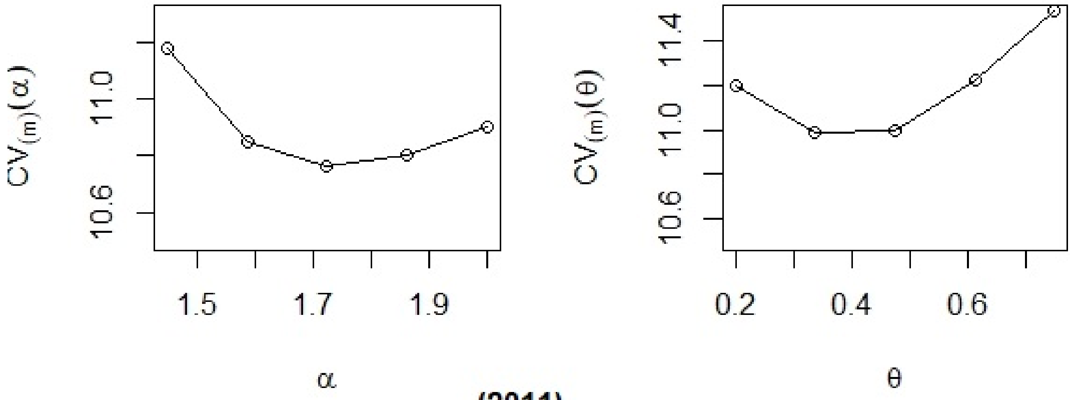

Decomposition by TBATS model
(Patumahoe)
observed
5
10
20
30
level
11.0
12.0
13.0
season1
-0.4
-0.2
0.0
0.2
0.4
season2
-1.0
0.0
0.5
1.0
2011
2012
2013
2014
2015
2016
Time


Decomposition by TBATS model
(Penrose)
observed
10
20
30
40
level
13.5
14.5
15.5
16.5
season1
-1.0
-0.5
0.0
0.5
1.0
season2
-1
0
1
2
2011
2012
2013
2014
2015
2016
Time

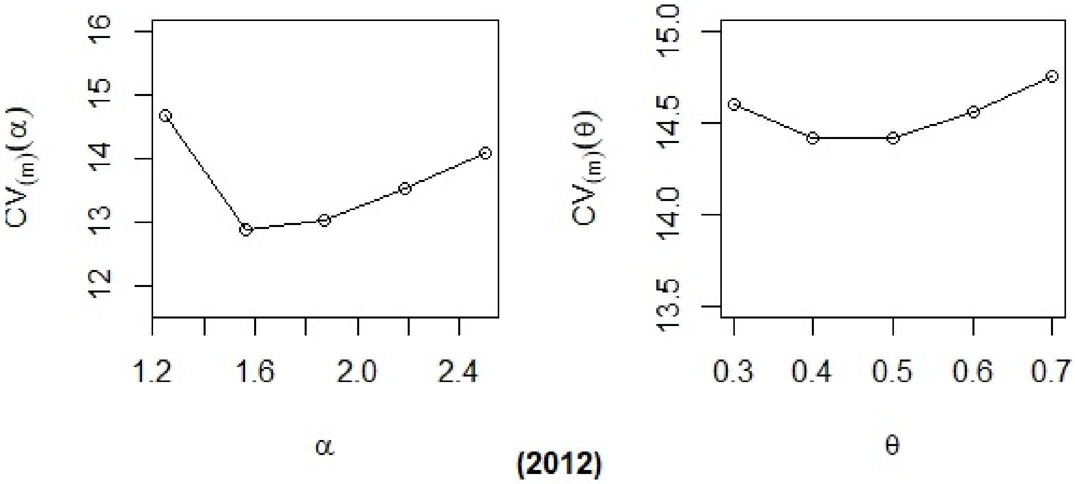

Decomposition by TBATS model
(Takapuna)
observed
5
10
20
30
level
13.5
14.5
15.5
season1
-1.0
-0.5
0.0
0.5
season2
-1.0
-0.5
0.0
0.5
1.0
2011
2012
2013
2014
2015
2016
Time

Appendix C

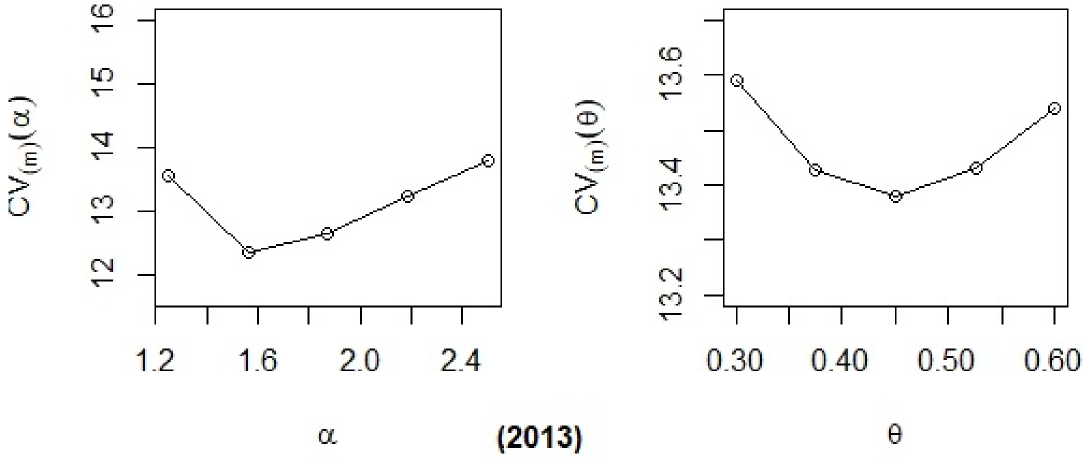
1) **The LOOCV score on different range of α and θ for IDW prediction (a) and Gaussian kernel prediction (b) of PM$_{10}$**



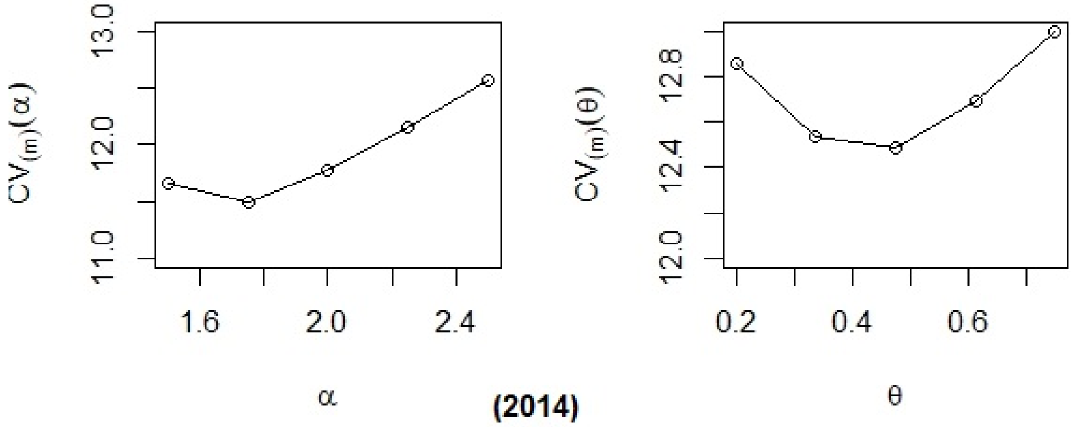
(2011)



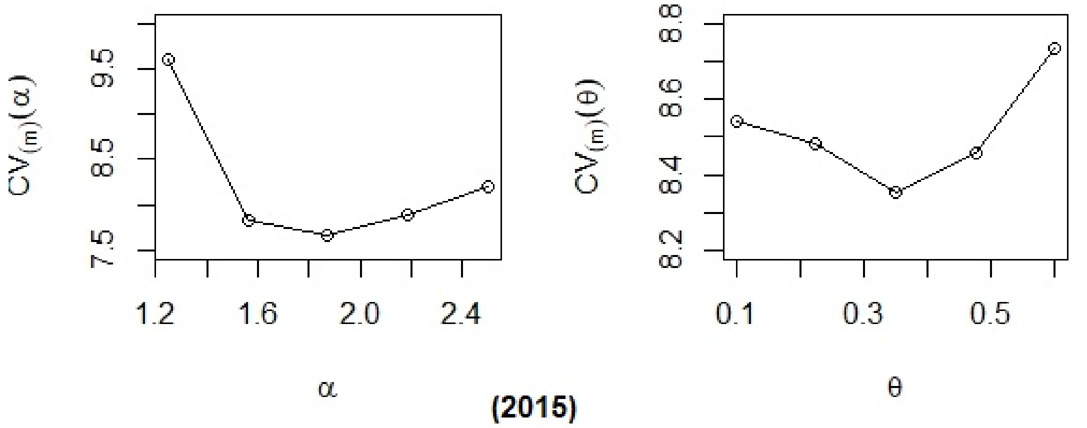
(2012)



(2013)



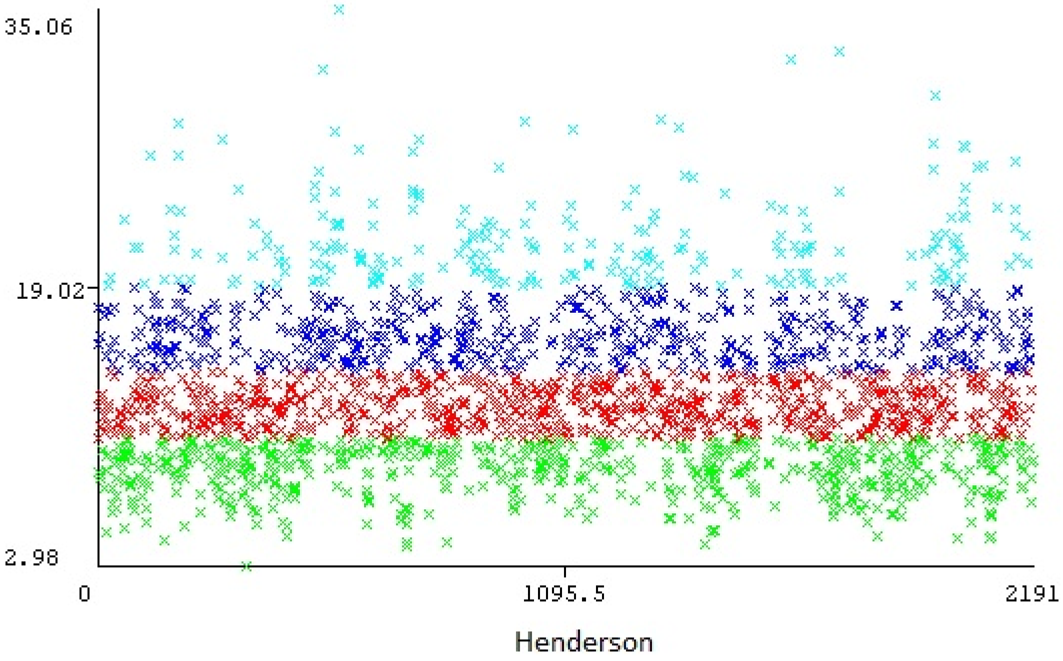
(2014)

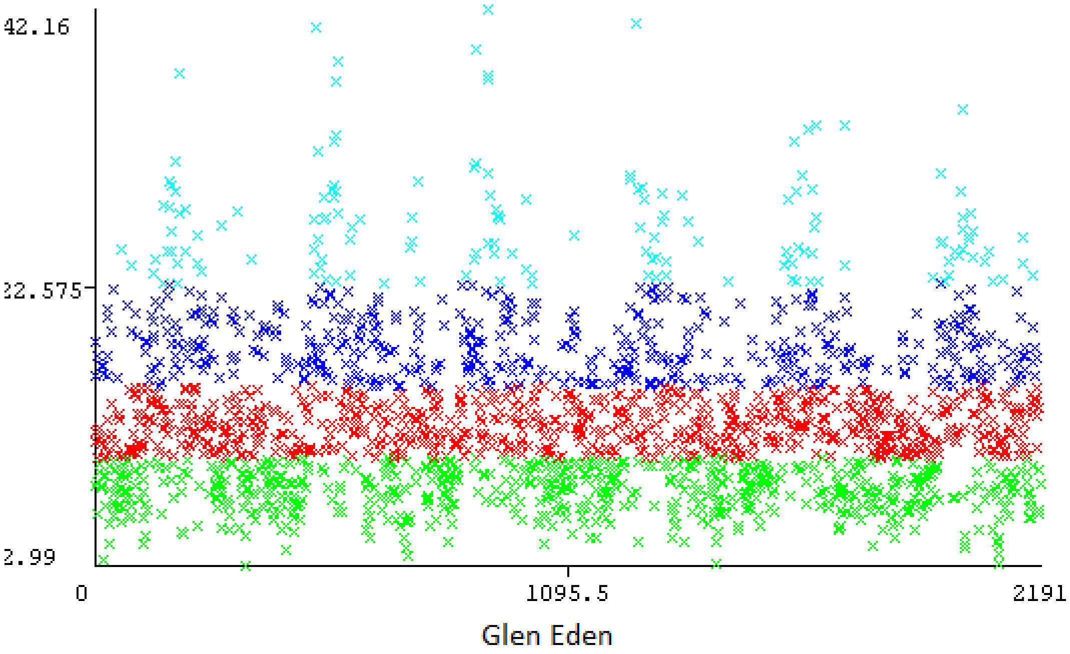


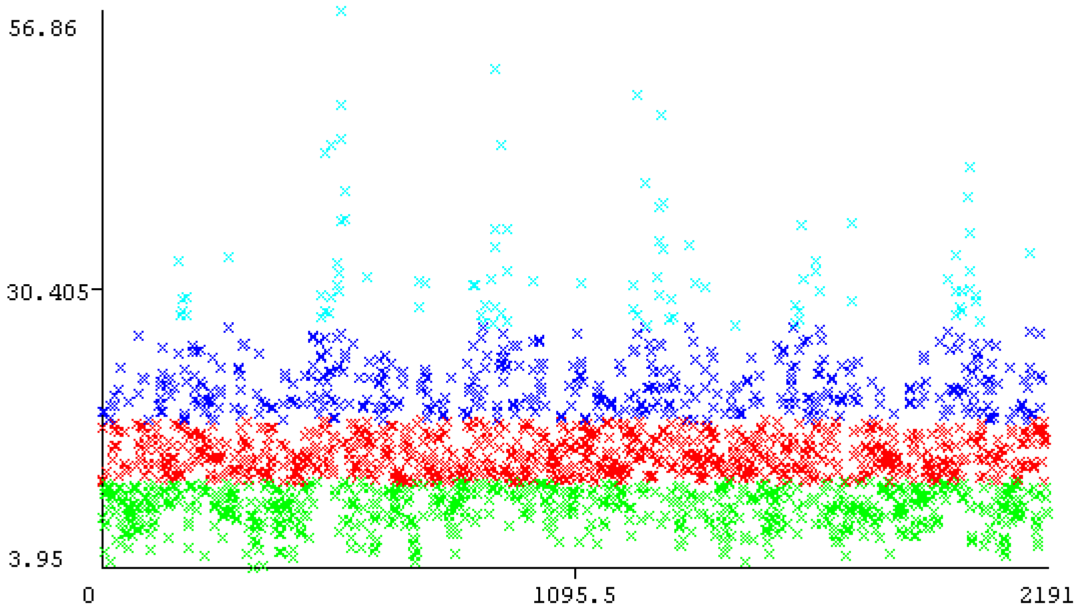
(2015)

Appendix D

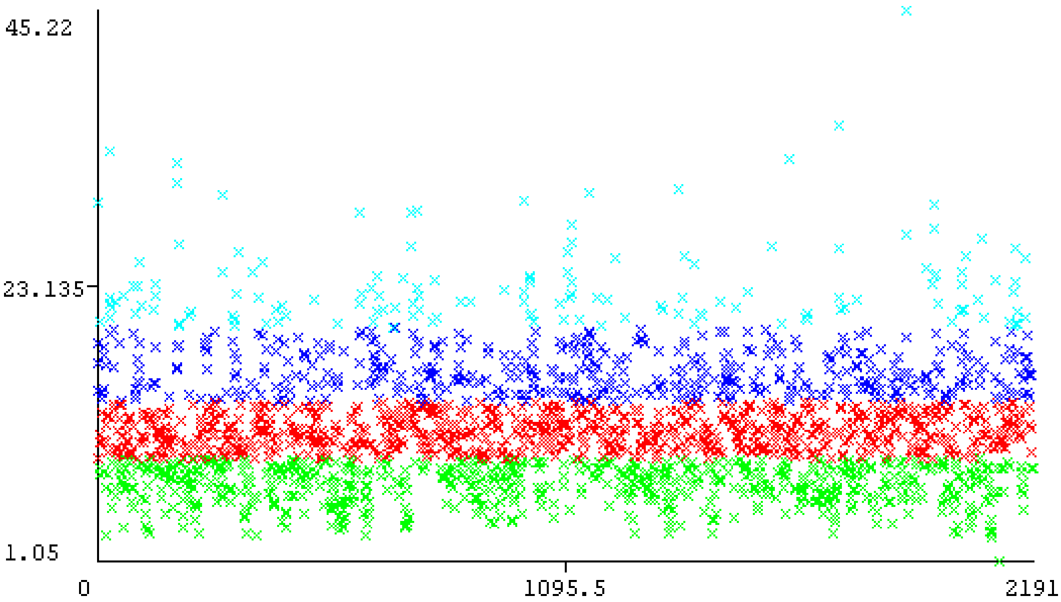
Site-specific K-mean clustering results



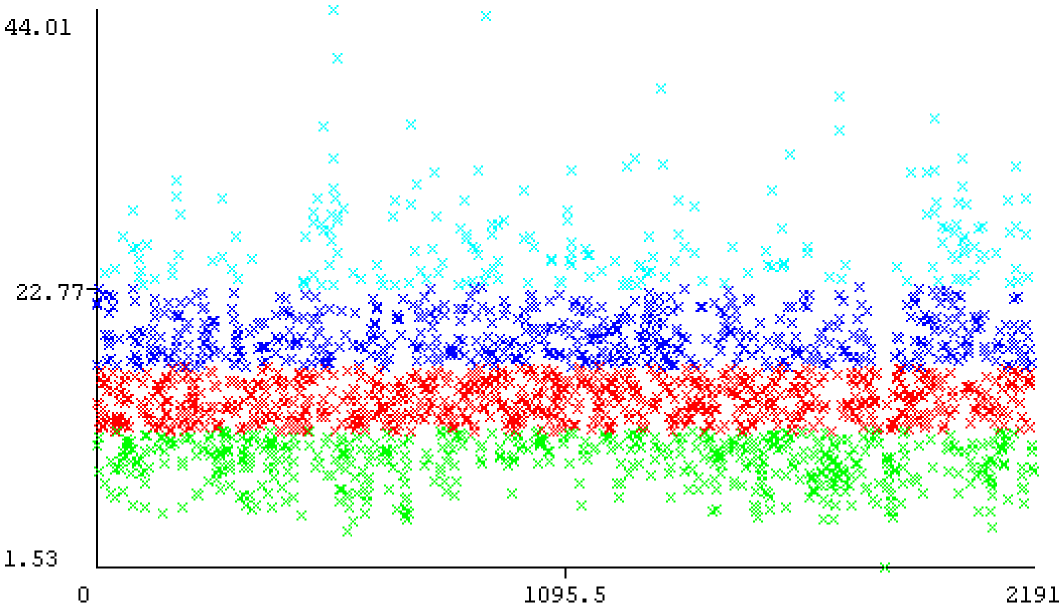




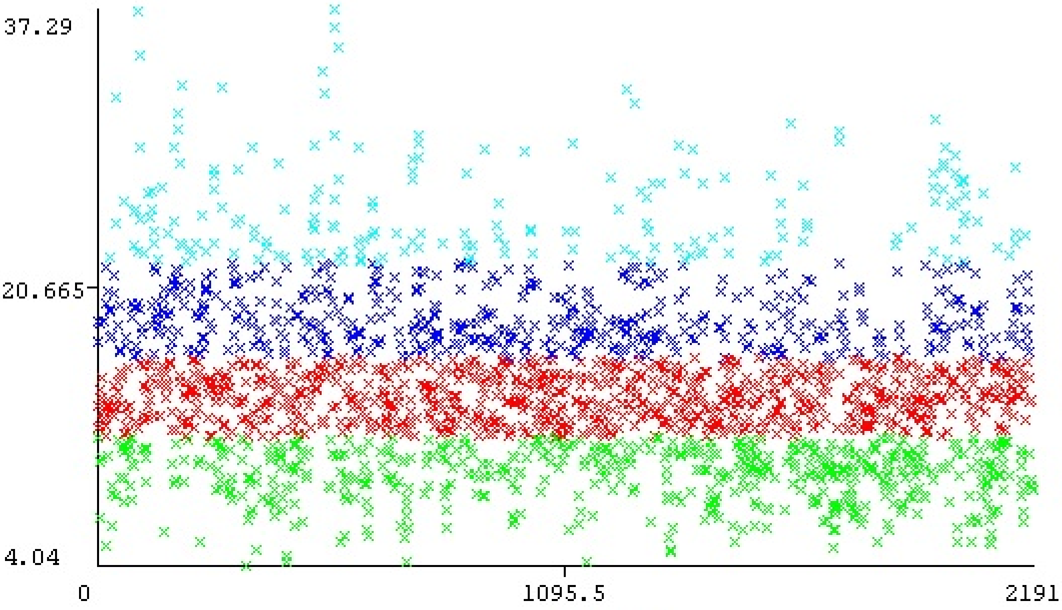
Pakuranga



Patumahoe



Penrose



Takapuna

Appendix E (Image copyrights)

1) **Figure 2.1:** PLoS ONE Copyright: © 2019 Sun, Zhu. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

2) **Figure 2.2:** Ministry for the Environment MfE & Stats NZ, 2018 Crown copyright ©.
Unless indicated otherwise for specific items or collections of content (either below or within specific items or collections), this copyright material is licensed for re-use under the Creative Commons Attribution 4.0 International licence.

In essence, you are free to copy, distribute and adapt the material, as long as you attribute it to the Ministry for the Environment and abide by the other licence terms.

Please note that this licence does not apply to any logos, emblems and trade marks on the website or to the website's design elements or to any photography and imagery. Those specific items may not be re-used without express permission.

3) **Figure 2.3:** Ministry for the Environment (MfE) Crown copyright ©.
Unless indicated otherwise for specific items or collections of content (either below or within specific items or collections), this copyright material is licensed for re-use under the Creative Commons Attribution 4.0 International licence.

In essence, you are free to copy, distribute and adapt the material, as long as you attribute it to the Ministry for the Environment and abide by the other licence terms.

Please note that this licence does not apply to any logos, emblems and trade marks on the website or to the website's design elements or to any photography and imagery. Those specific items may not be re-used without express permission.

4) **Figure 2.4 and Figure 2.5:**

From: Shahir Masri <sfm392@mail.harvard.edu>
Sent: Monday, 12 April 2021 4:57 PM
To: Sara Zandi <sara.zandi@aut.ac.nz>
Subject: Re: Permission on Using Your Journal Paper Figures

Hello Sara,

Thank you for your email. Yes, you may use the figures as long as they cite our original paper. Good luck with your thesis!

Kind regards,

Shahir

Shahir Masri, Sc.D., M.S.
Author of *Beyond Debate: Answers to 50 Misconceptions on Climate Change*

Assistant Specialist in Air Pollution Exposure Assessment & Epidemiology

University of California, Irvine

On Sun, Apr 11, 2021 at 5:47 PM Sara Zandi <sara.zandi@aut.ac.nz> wrote:

Dear Shahir,

My name is Sara Zandi and I am undertaking my PhD program in Auckland University of Technology (AUT), New Zealand. I am writing to ask for your permission on using two of the Figures (Figure 1. Relative contributions of coarse and fine particles to total PM10 mass. and Figure 2. Mass closure results for (a) fine and (b) coarse particles.) from the published journal article" Composition and sources of fine and coarse particles collected during 2002–2010 in Boston, MA" in my PhD thesis.

As part of our University's copyright rules, I am allowed to use the Figures (with citation) subject to the author's permission.

Could you please advise.

Kind regard,
Sara Zandi
Auckland University of Technology
New Zealand

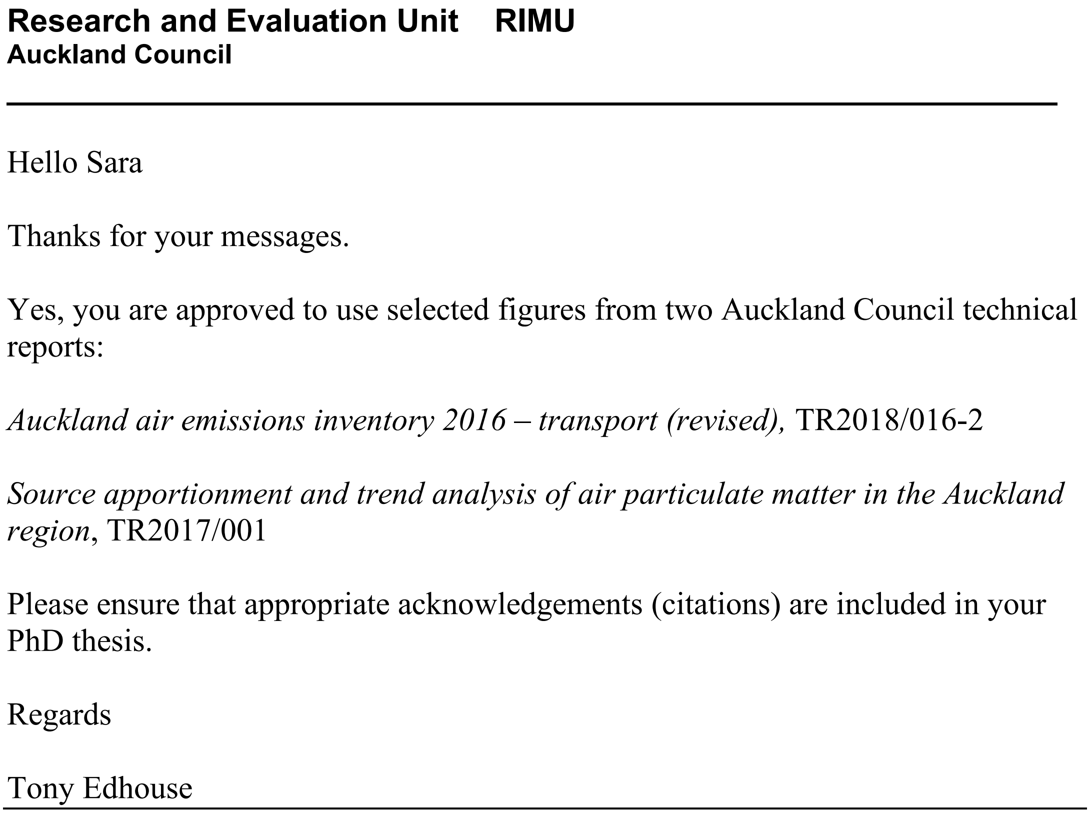
5) **Figure 2.6**

No special permission is required to reuse all or part of article published by MDPI, including figures and tables. For articles published under an open access Creative Common CC BY license, any part of the article may be reused without permission provided that the original article is clearly cited. Reuse of an article does not imply endorsement by the authors or MDPI.

6) **Figure 2.7** and **Figure 2.8**: Copyright notice © European Environment Agency, 2020 Reproduction is authorised provided the source is acknowledged.

7) **Figure 2.9**

From: Tony Edhouse <Tony.Edhouse@aucklandcouncil.govt.nz>
Sent: Thursday, 15 April 2021 12:45 PM
To: Sara Zandi <sara.zandi@aut.ac.nz>
Subject: Approval for use of figures from Auckland Council technical reports



Research and Evaluation Unit   RIMU
Auckland Council

Hello Sara

Thanks for your messages.

Yes, you are approved to use selected figures from two Auckland Council technical reports:

Auckland air emissions inventory 2016 – transport (revised), TR2018/016-2

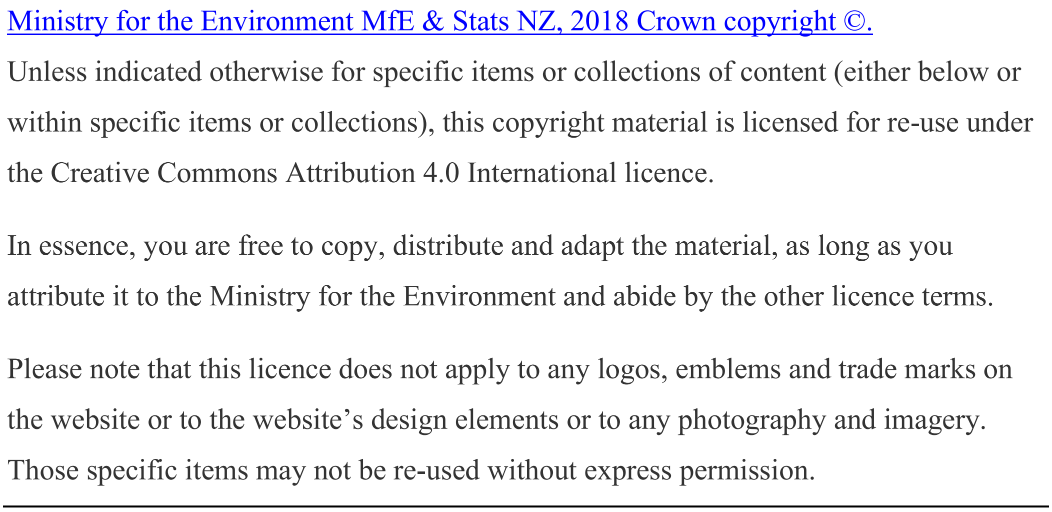
Source apportionment and trend analysis of air particulate matter in the Auckland region, TR2017/001

Please ensure that appropriate acknowledgements (citations) are included in your PhD thesis.

Regards

Tony Edhouse

8) Figure 2.10



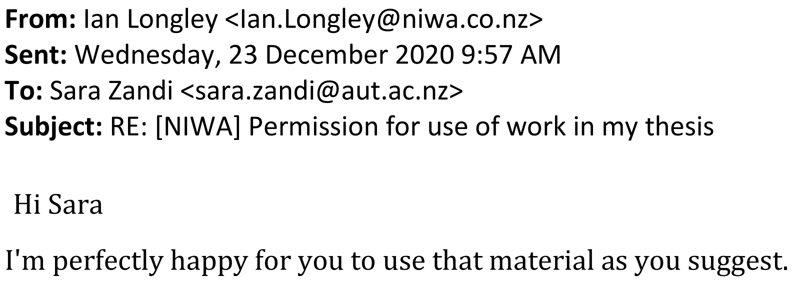
[Ministry for the Environment MfE & Stats NZ, 2018 Crown copyright ©.](#)

Unless indicated otherwise for specific items or collections of content (either below or within specific items or collections), this copyright material is licensed for re-use under the Creative Commons Attribution 4.0 International licence.

In essence, you are free to copy, distribute and adapt the material, as long as you attribute it to the Ministry for the Environment and abide by the other licence terms.

Please note that this licence does not apply to any logos, emblems and trade marks on the website or to the website's design elements or to any photography and imagery. Those specific items may not be re-used without express permission.

9) Figure 2.15



From: Ian Longley <Ian.Longley@niwa.co.nz>
Sent: Wednesday, 23 December 2020 9:57 AM
To: Sara Zandi <sara.zandi@aut.ac.nz>
Subject: RE: [NIWA] Permission for use of work in my thesis

Hi Sara

I'm perfectly happy for you to use that material as you suggest.

All the best

Ian

-----Original Message-----

From: webmaster@niwa.co.nz <webmaster@niwa.co.nz> On Behalf Of szandi@aut.ac.nz

Sent: Wednesday, 23 December 2020 9:15 AM

To: Ian Longley <Ian.Longley@niwa.co.nz>

Subject: [NIWA] Permission for use of work in my thesis

Hello Dr Ian Longley,

Sara Zandi (szandi@aut.ac.nz) has sent you a message via your contact form (https://aus01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fniwa.co.nz%2Fuser%2F449%2Fcontact&amp;data=04%7C01%7CIan.Longley%40niwa.co.nz%7C8d73bac2c7ef400d500b08d8a6b652e1%7C41caed736a0c468aba499ff6aafd1c77%7C0%7C0%7C637442649346834215%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C2000&amp;sdata=tbYv0i5BolM6ktfMG%2FMWa9QE3PsI4qZ2J1OkDP9RhR4%3D&amp;reserved=0) at NIWA.

this is a message forwarded from contact form on the NIWA website

Message:

Dear Ian Longley,

I am a Doctoral student at Auckland University of Technology and am writing a thesis on PM10 concentration modeling in urban area for my PhD degree.

I am writing to request permission for the following work, for which I believe you hold the copyright, to be included in my thesis:

Longley, I. (2020, 30/04/2020). Pollution levels soar in Level 3, says NIWA.

A digital copy will be made available online via the University's digital repository Tuwhera. This is an open access research repository for scholarly work, intended to make research accessible to as wide an audience as possible.

I am seeking from you a non-exclusive licence to include these materials in my thesis. The materials will be fully and correctly referenced.

If you agree, I should be very grateful if you would reply to me via email.

If you do not agree, or if you do not hold the copyright in this work, would you please let me know.

I can most quickly be reached by email. Thank you for your assistance. I look forward to hearing from you.

Yours sincerely,

Sara Zandi

10) Figure 7.1

From: Mark Fischer <mfischer@niu.edu>
Sent: Wednesday, 20 November 2019 4:27 PM
To: Sara Zandi <sara.zandi@aut.ac.nz>
Subject: {Spam?} Re: Permission on Using Your Journal Paper Figures

Hello Sara,

You are more than welcome to use that figure in your thesis. You may want to contact the journal and request permission from them too, as they are the official copyright holder on that paper. You may need to speak with your advisor or a university representative to determine whether you actually need to do that.

If for some reason you need or would like to have the original Adobe Illustrator version of the figure, just let me know and I can send it to you. Having that file might make it easier to edit the figure to your needs. You are welcome to do that too, as long as you cite the original paper.

Best of luck on your research.

—Mark

Professor Mark P. Fischer, Ph.D.
Chair, Geology and Environmental Geosciences
Northern Illinois University
DeKalb, IL 60115-2828, USA
Phone: 815.753.0523
FAX: 815.753.1945

On Nov 19, 2019, at 2:40 PM, Sara Zandi <sara.zandi@aut.ac.nz> wrote:

Dear Mark,

My name is Sara Zandi and I am undertaking my PhD program in Auckland University of Technology (AUT), New Zealand. I am writing to ask for your permission on using one of the Figures (*Fig. 7. Schematic illustration of the steps involved in the semivariogram modeling*) from your published journal article" *Directional semivariogram analysis to identify and rank controls on the spatial variability of fracture networks" in* my PhD thesis.

As part of our University's copyright rules, I am allowed to use the Figures (with citation) subject to the author's permission.

Could you please advise.
Kind regards,
Sara Zandi

11) Figure 8.1 and Figure 8.2

From: Christopher Olah <christopherolah.co@gmail.com>

Sent: Saturday, 19 September 2020 2:55 AM

To: Sara Zandi <sara.zandi@aut.ac.nz>

Subject: Re: Your permission on adopting the LSTM images in my thesis

Dear Sara,

You are very welcome to use my LSTM diagrams. Please cite the blog post.

All the best with your writing your thesis.

Chris

On Thu, Sep 17, 2020 at 11:50 PM Sara Zandi <sara.zandi@aut.ac.nz> wrote:

Dear Christopher,

My name is Sara Zandi and I am undertaking my PhD study at AUT University. I found your post on Understanding LSTM Networks (Posted on August 27, 2015) very useful, so I am writing to ask for permission of adopting the LSTM images in my thesis. As part of AUT University referencing regulation I must get your permission on using graphs and images. Could you please let me know if I get your permission and if you have used these images in other publications so I can cite them? Otherwise, I will cite your blog as online resource.

Kind regards,
Sara Zandi