Data Pre-processing to Identify Environmental Risk Factors Associated with Diabetes

by

Lakmini Wijesekara

Principal Supervisor: Dr. Liwan Liyanage

Co-supervisor: A/Prof. Michael O'conner

A thesis submitted for the degree of

Doctor of Philosophy

School of Computer, Data and Mathematical Sciences

Western Sydney University

July, 2023

Statement of Authentication

I, Lakmini Wijesekara, declare that the work presented in this thesis titled, “Data Pre-processing to Identify Environmental Risk Factors Associated with Diabetes” is, to the best of my knowledge and belief, original except when the due reference has been made in the text of this thesis. I hereby declare that I have not submitted this material, either in full or part, for a degree at this or any other institution.

Signed: 

Date: 11/07/2023

Dedicated to my beloved family

Acknowledgements

First and foremost, I would like to express my most profound appreciation to Western Sydney University, Australia, for providing me with an opportunity to pursue my PhD with the Western Sydney University Postgraduate Research Scholarship. I would also like to extend my deepest gratitude to the Department of Statistics, University of Colombo for the support provided to get the PhD placement at WSU. Also, I am extremely grateful to the AHEAD (Accelerating Higher Education Expansion and Development) operation of Sri Lanka funded by World Bank for providing me with the financial support for the period until I got the WSU scholarship.

I cannot begin to express my thanks to my principal supervisor Dr Liwan Liyanage. I am extremely grateful for her invaluable advice and immense support during my PhD. My gratitude extends to my co-supervisor Dr Michael O'Conner. Moreover, my gratitude extends to A/Prof. Dongmo Zhang, the HDR director of the School of Computer, Data and Mathematical Sciences, Western Sydney University and Dr. Omar Mubin, the acting HDR director for his advice and support. I also thank the Research HDR Coordinator and the Librarian of the school. I would like to thank my friends and colleagues for their encouragement. Last but not least, I express my gratitude to my family, especially my husband for his patience and scarification throughout the duration of the PhD. Without their tremendous understanding and encouragement, it would be impossible to complete my PhD.

Contents

List of Figures

List of Tables

"The world is one big data problem."

Dr. Andrew McAfee

Publications

1. Wijesekara, L., and Liyanage, L. (2020). Data exploration and pre-processing techniques on air pollution and meteorological data in Sydney region. In Proceedings: International Conference on Environmental and Medical Statistics, 9-10 January 2020, Postgraduate Institute of Science, University of Peradeniya, Sri Lanka (pp. 30-30).
   https://hdl.handle.net/1959.7/uws:55976

2. Wijesekara, W. M. L. K. N., and Liyanage, L. (2020). Comparison of imputation methods for missing values in air pollution data: Case study on Sydney air quality index. In Future of Information and Communication Conference (pp. 257-269). Springer, Cham.
   https://doi.org/10.1007/978-3-030-39442-4_20

3. Wijesekara, L., and Liyanage, L. (2020). Modelling Environmental Impact on Public Health using Machine Learning: Case Study on Asthma. In 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA) (pp. 1-7). IEEE.
   https://doi.org/10.1109/CITISIA50690.2020.9397488

4. Wijesekara, L., and Liyanage, L. (2021). Air quality data pre-processing: A novel algorithm to impute missing values in univariate time series. In 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 996-1001). IEEE.
   https://doi.org/10.1109/ICTAI52525.2021.00159

5. Wijesekara, L., and Liyanage, L. (2021). Imputing Large Gaps of High-resolution

Environment Temperature. In 2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS) (pp. 74-79). IEEE.
https://doi.org/10.1109/ICIIS53135.2021.9660672

6. Wijesekara, L., Nanthakumaran, P., and Liyanage, L. (2023). Space and Time Data Exploration of Air Quality Based on PM10 Sensor Data in Greater Sydney 2015-2021. In International Conference on Sensing Technology (pp. 295-308). Cham: Springer Nature Switzerland.
https://doi.org/10.1007/978-3-031-29871-4_30

7. Wijesekara, L., and Liyanage, L. (2023). Mind the Large Gap: Novel Algorithm Using Seasonal Decomposition and Elastic Net Regression to Impute Large Intervals of Missing Data in Air Quality Data. Atmosphere, 14(2), 355.
https://doi.org/10.3390/atmos14020355

Posters

- Poster titled "Missing Piece Prevents Seeing the Big Picture: Dealing with missing values in air pollution and weather to identify environmental risk factors associated with diabetes" presented at the Westmead Research and Innovation Conference on 18-19 August 2022 in Sydney.
https://github.com/Lakminikw/westmeadcon2022/blob/main/poster.pdf

Other

- Visualise Your Thesis (VYT) Competition 2022
Video Entry: https://www.westernsydney.edu.au/schools/grs/the_student_experience/visualise_your_thesis/2022_vyt_competition
One of the finalists in the Western Sydney University

- Western Sydney University Junior Researchers Conference 2022
Video Entry: https://www.youtube.com/watch?v=XcgnP-qQp3E&t=15s&ab_channel=CDMSOpenDay2022
Received the People's Choice Award for the Best Lightning Talk

Abstract

Lakmini Wijesekara

Data Pre-processing to Identify Environmental Risk Factors Associated with Diabetes

Diabetes has become a substantial burden on global health. According to reports from the World Health Organization (WHO), diabetes was the ninth leading cause of death in 2019, with an estimated 1.5 million deaths directly caused by diabetes. "Halting the rise in diabetes and obesity by 2025" is one of the goals of the Global Action Plan for the Prevention and Control of Non-Communicable Diseases (NCDs). More than 1.2 million Australians currently live with diabetes. Hospitalisations due to diabetes in the Western Sydney area are growing compared to other regions in Sydney. Reducing the impact of diabetes in Western Sydney is one of the goals of the Western Sydney University's diabetes research team.

Genetics, diet, obesity, and lack of exercise play a major role in the development of type II diabetes. Additionally, environmental conditions are also linked to type II diabetes. The aim of this research is to identify the environmental conditions associated with diabetes. To achieve this, the research study utilises hospital-admitted patient data in NSW integrated with weather, pollution, and demographic data. The environmental variables (air pollution and weather) change over time and space, necessitating spatiotemporal data analysis to identify associations. Moreover, the environmental variables are measured using sensors, and they often contain large gaps of missing values due to sensor failures. Therefore, enhanced methodologies in data cleaning and

imputation are needed to facilitate research using this data. Hence, the objectives of this study are twofold: first, to develop a data cleaning and imputation framework with improved methodologies to clean and pre-process the environmental data, and second, to identify environmental conditions associated with diabetes.

This study develops a novel data-cleaning framework that streamlines the practice of data analysis and visualisation, specifically for studying environmental factors such as climate change monitoring and the effects of weather and pollution. The framework is designed to efficiently handle data collected by remote sensors, enabling more accurate and comprehensive analyses of environmental phenomena that would otherwise not be possible. The study initially focuses on the Sydney Region, identifies missing data patterns, and utilises established imputation methods. It assesses the performance of existing techniques and finds that Kalman smoothing on structural time series models outperforms other methods. However, when dealing with larger gaps in missing data, none of the existing methods yield satisfactory results. To address this, the study proposes enhanced methodologies for filling substantial gaps in environmental datasets. The first proposed algorithm employs regularized regression models to fill large gaps in air quality data using a univariate approach. It is then extended to incorporate seasonal patterns and expand its applicability to weather data with similar patterns. Furthermore, the algorithm is enhanced by incorporating other correlated variables to accurately fill substantial gaps in environmental variables. To evaluate the performance of the algorithm, missing values are artificially generated and then imputed using both the proposed method and a chosen set of existing methods. Various scenarios are considered, with missing percentages ranging from 10% to 50%, and gap sizes spanning from 50 to 500. To ensure reliable results, each scenario is repeated multiple times, and the positions of missing values are randomly changed in each repetition to avoid any misleading results due to chance. The evaluation includes a comparison of Root Mean Squared Errors between the original and imputed series, as well as a visual inspection of the actual and imputed data. Consistently, the algorithm presented in this thesis outperforms other methods in imputing large gaps. This algorithm is applicable for filling large gaps in air pollution and weather

data, facilitating downstream analysis.

The study also utilises Quasi-Poisson regression models to investigate the relationship between hospital admissions for type II diabetes-related issues and environmental factors. Type II diabetes hospital admissions show a positive relationship with the humidity level of the environment. However, further research is needed to make causal inferences. The methodologies used in this study have the potential to be applied to other diseases influenced by environmental conditions. Additionally, the findings of this research hold significant value for policymakers, offering valuable insights for the effective management and reduction of the impact of diabetes.

Chapter 1

Introduction

This chapter introduces the background of the thesis, objectives, significance and an overview of the rest of the thesis.

1.1 Background

1.1.1 Diabetes Burden

Diabetes is identified as one of the most common chronic diseases in the world. When the body's ability to produce insulin hormone is impaired, or the function of insulin is impaired, the level of blood sugar increases. This causes diabetes. The two most common forms of diabetes are type I and type II. Type I diabetes occurs when the pancreas fails to produce insulin which is essential for the breakdown of glucose. In this situation, external insulin is needed for treatment. This type of diabetes is identified primarily in children and adolescents. Type II diabetes occurs when the body cannot respond to the insulin produced by the pancreas or no insulin is produced. This type of diabetes is much more common in adults and accounts for around 90% of all diabetes cases worldwide. Even though diabetes can be treated, it can lead to complications such as kidney failure, heart disease, nerve damage, vision loss, leg amputation and stroke (World Health Organization, 2016).

Diabetes has become a substantial burden for global health. As of 2014, there were 422 million adults living with diabetes. "Halt the rise in diabetes and obesity in 2025" is one of the goals of the Global Action Plan for the Prevention and Control of Non-Communicable Diseases (NCDs) (World Health Organization, 2016). In Australia, more than 1.2 million individuals are suffering from diabetes, and over 2 million are estimated to be at high risk of developing diabetes (Sainsbury et al., 2018). The economic impact of diabetes in Australia is substantial, with an estimated annual cost of AUD14.6 billion (Lee et al., 2013). Therefore, it is crucial to implement the necessary measures to reduce the burden of diabetes in Australia. Moreover, the number of hospital admissions due to diabetes-related issues in Western Sydney is almost twice that of the rest of the Sydney Region. Consequently, reducing the impact of diabetes in Western Sydney is one of the goals of the Research and Innovation Plan 2018-2020, at Western Sydney University under the Health and Well-being research theme (Western Sydney University, 2019).

1.1.2 Diabetes risk factors

Type I diabetes is an auto-immune disorder, whereas type II diabetes is influenced by numerous risk factors. Genetics, diet, obesity and sedentary lifestyle are some of them which play a major role in developing type II diabetes. While these factors are often discussed and investigated, some other factors are underinvestigated. Environmental pollution, stress, chemical exposure, ethnicity, and socio-economic status are some of them (Joshi & Shrestha, 2010). Figure 1.1 summarises the factors associated with developing type II diabetes.

A plethora of research has been conducted to find the socioeconomic and lifestyle factors affecting diabetes. However, the research on associations of diabetes with environmental factors, especially air pollution and weather, is still in its infancy. Therefore, the focus of this research is to find out air pollution and weather effects on type II diabetes. Throughout this thesis, 'environmental factors' refers to ambient air pollution and weather factors.

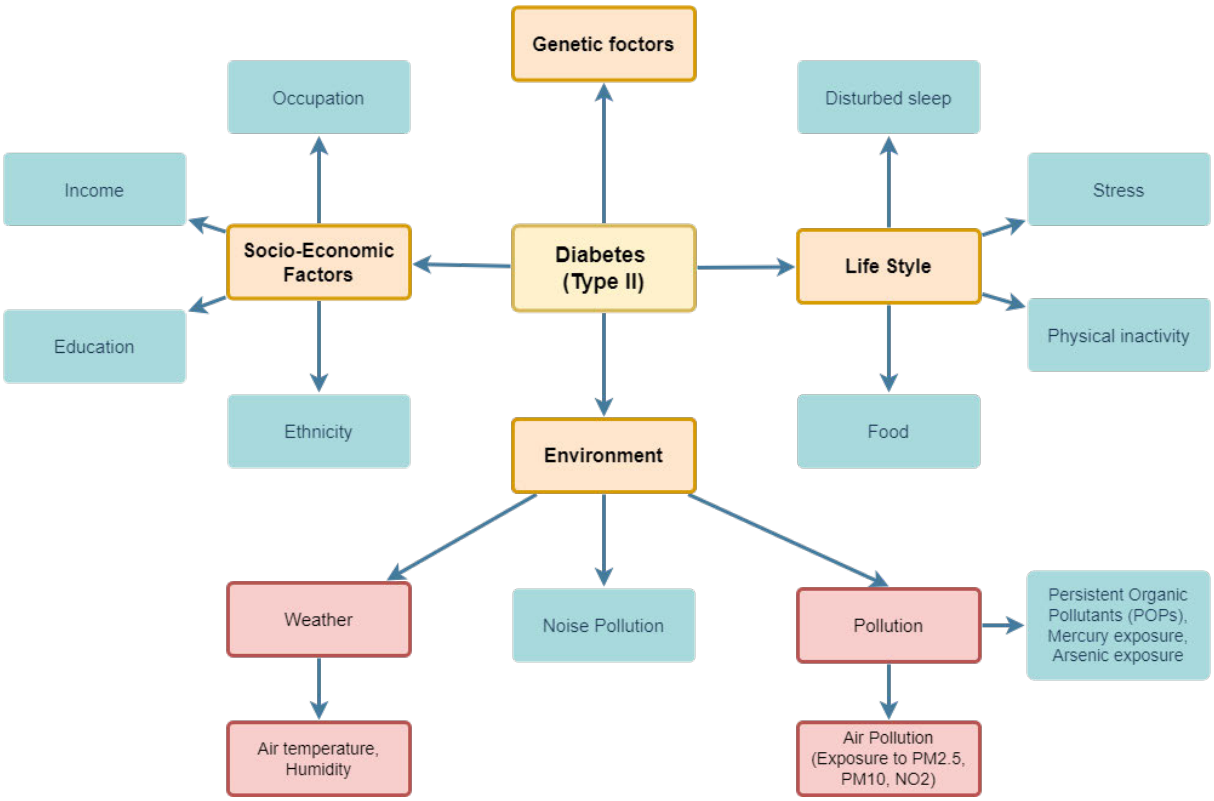


FIGURE 1.1: Factors associated with type II diabetes. Compliled from (Joshi & Shrestha, 2010) (Ling & Groop, 2009) (Rajagopalan & Brook, 2012) (Yang et al., 2020a) (Blauw et al., 2017) (Thiering & Heinrich, 2015) (Zanobetti et al., 2014) (Tyrovolas et al., 2014) and (Montonen et al., 2005)

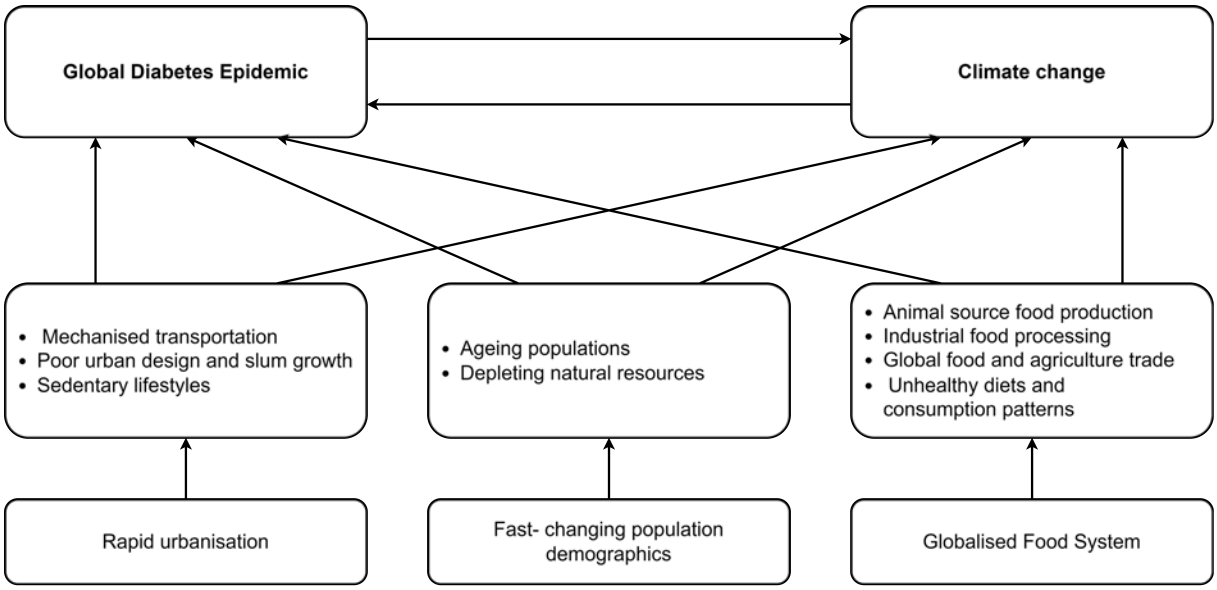


FIGURE 1.2: : Interconnections between Type II Diabetes and Climate change. Source: (International Diabetes Federation, 2012)

1.1.3 Diabetes and Climate Change

While genetics, diet, obesity and lack of exercise play a significant role in developing diabetes, climate change also exhibits links with diabetes (International Diabetes Federation, 2012). Type II diabetes has direct as well as indirect links with climate change as summarized in Figure 1.2.

The process of food production from field to dining table is becoming more expensive, more environmentally harmful in terms of greenhouse gas (GHG) and diminishing its natural value (Colagiuri, 2013). A higher prevalence of obesity is associated with high carbon-intensive health systems and greenhouse gas emissions. Since obesity is a principal driver of diabetes, these associations hold for diabetes as well. Extreme climatic events such as heat waves can cause an increased risk of morbidity and mortality in those with diabetes. Moreover, increasing temperatures can cause heatstroke and cardiovascular diseases, attributing complications to patients with diabetes. Further, extreme climatic events cause enormous threats to food security. Hence the number of people with malnutrition increases, leading to a high prevalence of type II diabetes. Also, climate change directly affects health infrastructure. Rapid urbanisation, fast-changing population demographics and globalised food systems are common

driving factors of the global diabetes epidemic and climate change (International Diabetes Federation, 2012).

Physical inactivity and inappropriate diets have a straightforward relationship with type II diabetes. On the other hand, those two activities may have a strong relationship with the activities such as inactive transport, sedentary work and mass-manufactured food, which lead to high carbon footprints (Colagiuri, 2013). Therefore, a high carbon footprint has an indirect link with type II diabetes. Even though this is not a causal relationship, a significant association can be expected.

A healthy diet, rich in fruits and vegetables, is central to preventing type II diabetes and is friendly to the environment. The wrong food production methods damage the soil. Hence plant-based food becomes poor in quality. In order to take the requirements of nutrients, one has to consume more food than in the past. This leads to overeating and obesity and hence type II diabetes. Therefore, land mix use is another example that is leading to both climate change and obesity.

1.1.4 Missing values problem in environmental data

Air pollution weather data are usually collected through sensors, and it is a common problem to have missing values in these data due to failures of sensors. Missing values of these data may lead to underestimating or overestimating the associated health effects. The focus of this research lies on the New South Wales state of Australia, with particular attention given to the Sydney region. The missing value problem is persistent in the environmental data in the NSW as well. Figure 1.3 depicts the number of missing values in air quality index data in the Sydney region collected through the NSW monitoring network from 1994 to 2018. More red cells in the heat map highlight the problem of missing data. After 2013, the problem seemed to be reduced with the advancements in sensing technologies. However, missing values are still persistent. Figure 1.4 further illustrates the percentage of missing values in each monitoring site for the period from 1994 to 2018. Detailed analysis of the distributions of missing values over time for each variable at each monitoring site in the whole NSW, as well as statistics of missing values, are given in Appendix A. What is striking from this analysis

is that the problem of missing is not ignorable in order to carry out any downstream analysis.

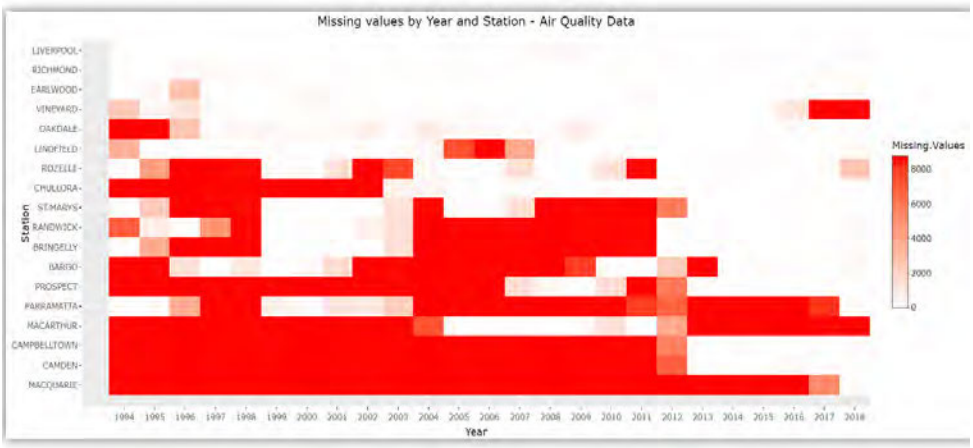


FIGURE 1.3: Heat map of the number of missing values by stations from 1994-2018 in the Sydney region (central-east, north-west, south-west)

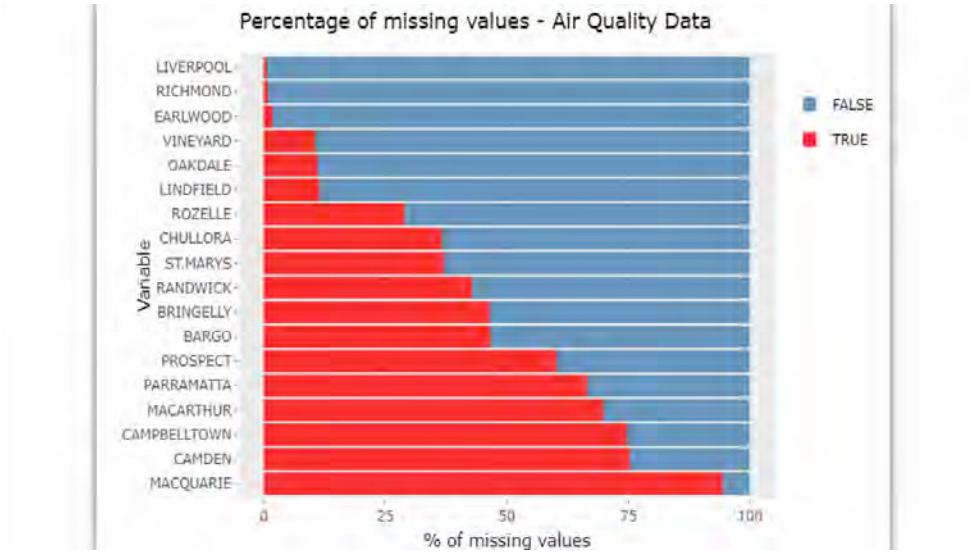


FIGURE 1.4: Percentage of missing values of the stations from 1994-2018. False: Non-missing, True: Missing

1.1.5 Missing Mechanism

Depending on the apparent reason for being missing, there are three types of missing mechanisms. They can be identified as Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR) (Rubin, 1976). Let

$Y = (y_1, ..., y_n)^T$ denote a vector of data with both observed (Y_{obs}) and missing values (Y_{mis}) with probability density function f_θ. The objective is to make inferences about the vector of the unknown parameter θ. The missing indicator, $M = (M_1, ..., M_n)^T$ defines a binary variable that denotes whether the value of a variable is observed or missing (i.e. $M_i = 0$ if value y_i is observed and $M_i = 1$ if the value is missing). The conditional distribution of M given the complete data $Y, f(M|Y, \phi)$ describes the missing data mechanism where ϕ denotes the vector of unknown parameters that describe the probability of missing data (Rantou, 2017).

MCAR:

$$P(M|Y_{obs}, Y_{mis}, \phi) = P(M|\phi) \; for \; all \; Y, \phi \tag{1.1}$$

MAR:

$$P(M|Y_{obs}, Y_{mis}, \phi) = P(M|Y_{obs}, \phi) \; for \; all \; Y_{mis}, \phi \tag{1.2}$$

MNAR:

$$P(M|Y, \phi) = P(M|Y_{obs}, Y_{mis}, \phi) \tag{1.3}$$

In the MCAR scenario, the missingness occurs entirely randomly. That is, independent of both observable and unobservable parameters of interest. In MAR, there is a systematic relationship between the propensity of a value to be missing and the observed data while in MNAR, there is a relationship between the propensity of a value to be missing and its unobserved value. However, this terminology is criticised as MAR does not mean that the missing data are distributed at random. MCAR and MAR situations are considered as 'ignorable' while MNAR is considered 'non-ignorable' (Rubin, 1976; Rubright et al., 2014). In this context, 'ignorable' means that it is unnecessary to make any particular assumptions about how the data are missing to recover missing values (Nakagawa & Freckleton, 2011). Figure 1.5 gives a schematic representation of the three mechanisms of missing data. These missing mechanisms are further explained in Table 1.1 using examples in general and time series settings.

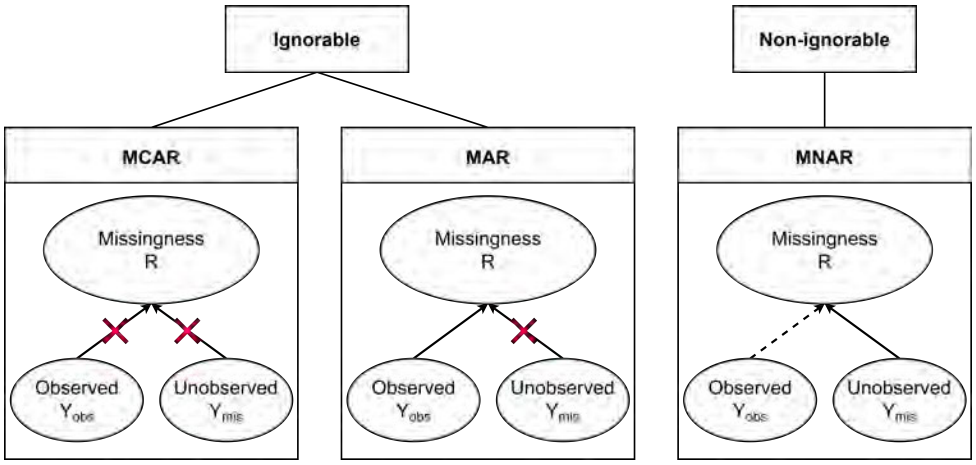


FIGURE 1.5: A schematic representation of the three classes (mechanisms) of missing data (i.e. MCAR, MAR and MNAR) in relation to the observed values/variables (Y obs), unobserved values/variables (Y mis), missingness (R) and ignorability. (Adopted from Nakagawa and Freckleton, 2011)

TABLE 1.1: Examples for each missing mechanism under general setting and time series setting

Missing Mechanism	**General**	**Time series**
MCAR	Tube containing a blood sample of an individual who is under study is broken by accident so that the blood variables cannot be measured, Questionnaire of a survey participant is lost (Donders et al., 2006)	Sensor data are recorded from a field test and sent via radio signals to be recorded. The transmission fails on random occasions due to unknown reasons (Moritz et al., 2015)
MAR	Income of an individual who is under study can be missing when the level of income is relatively high (Donders et al., 2006)	Particular sensor machine is shutdown on some weekends for maintenance the data are more likely to be missing on weekends (Moritz et al., 2015)
MNAR	When predicting the outcome of a diagnostic test based on some patient characteristics, the test results are known for all the diseased subjects whereas unknown for a random sample of no-diseased subjects (Donders et al., 2006)	Temperature sensor fails to record values when the temperature is over 50^0C (Moritz et al., 2015)

1.2 Research Objectives

The study has two-fold aims, finding solutions to the application domain while developing improved methods contributing to the disciplines, Artificial Intelligence (Pattern Recognition and Data mining), Statistics and Data Science. The objectives can be presented as follows.

1.2.1 Objective 1:Theoretical contribution

- To develop a framework to identify and analyse missing data in environmental data (whether and pollution data).
- To develop improved algorithms to clean and impute environmental data with demonstrable accuracy (whether and pollution data).

1.2.2 Objective 2: Application in Population Health (Diabetes)

- To identify associations of environmental conditions on type II diabetes.

1.3 Significance of the Research

1.3.1 Objective 1: Theoretical Contribution

Considerations about missing values can rarely be seen in spatiotemporal analysis literature. So far, the well-established data imputation techniques are still insufficient to deal with spatiotemporal datasets. This will further explain throughout the thesis. The data pre-processing techniques suggested to use throughout this research will be helpful not only in the analysis of environment variables but also in other time series datasets which display similar characteristics.

1.3.2 Objective 2: Application

This study aims to identify environmental risk factors associated with diabetes. As there are links between environmental variables and diabetes, research needs to be carried out to explain those links. Only a few studies have been carried out attempting

to explain these relations. "Halt the rise in diabetes and obesity in 2025" is one of the goals in the Global Action Plan for the Prevention and Control of Non-Communicable Diseases (NCDs), and reducing the impact of diabetes in Western Sydney is one of the goals in Western Sydney University Research and Innovation Plan under the Health and Wellbeing research theme. Since diabetes has become a burden to Australia, it is useful to investigate the links between the environment and diabetes in NSW. The findings of this research may provide the necessary information for policymakers to reduce the burden of diabetes. Also, it could be useful in holistic approaches to prevent the adverse effects of climate change. According to the literature, spatiotemporal analysis of type II diabetes in Australia has not been carried out. Therefore, this study will contribute to the knowledge of diabetes in NSW as well as the links between diabetes and the environment. Moreover, the findings of this research may be useful for the general public, especially for whom are at risk of developing diabetes.

Further, the chief executive of the International Diabetes Federation mentions that overcoming diabetes epidemiology is a long-term goal which may need intergenerational attention with holistic approaches (Guardian, 2012). Therefore, the findings of this research will be helpful in holistic approaches linking health and environmental issues to make policies to overcome those.

1.4 Thesis Outline

Figure 1.6 illustrates the organisation of the thesis.

Chapter 1. Introduction

This chapter explains the background of the research, its objectives and the significance of the research.

Chapter 2: Literature Review

This chapter includes a review of the literature. It explains the methods for missing values imputation, links between diabetes and environmental variables and the available spatiotemporal analysis methods.

Chapter 3: Data and Methodology

This chapter describes the data and explains the methodology used in achieving the research objectives.

Chapter 4: Sensor Data Cleaning Framework

This chapter describes more about pollution and weather data and explains the data cleaning framework.

Chapter 5: Missing Value Imputation: Univariate Approcahes

This chapter includes a comparison of time series imputation methods on air quality data. Also, it gives two improved algorithms to impute missing values; one for non-seasonal air quality variables and another one for seasonal weather variables.

Chapter 6: Missing Value Imputation: Multivariate Approach

This chapter proposes an algorithm to impute large gaps in air quality variables incorporating the use o correlated variables,

Chapter 7: Air Quality Data Analysis and Clustering

This chapter includes an analysis of air quality variables in the Sydney region

Chapter 8: Application of Machine Learning in Health: Case Study with Asthma

This chapter presents an application of machine learning methods in predicting the risk of asthma based on air pollution and weather. Even though this chapter is not directly related to the rest of the thesis, the methods may apply to diabetes as well. This analysis was done before receiving the diabetes dataset to investigate the applicability of the machine learning models in health data in general.

Chapter 9: Spatiotemporal Data Analysis on Diabetes

This chapter includes the analysis carried out to find the associations of diabetes with air pollution and weather.

Chapter 10: Discussion and Conclusions

This chapter will discuss the overall findings and contributions of the research with identified problems and further research possibilities. Then it will provide a conclusion based on the identified environmental risk factors and the potential recommendations to reduce the risk of diabetes burden.

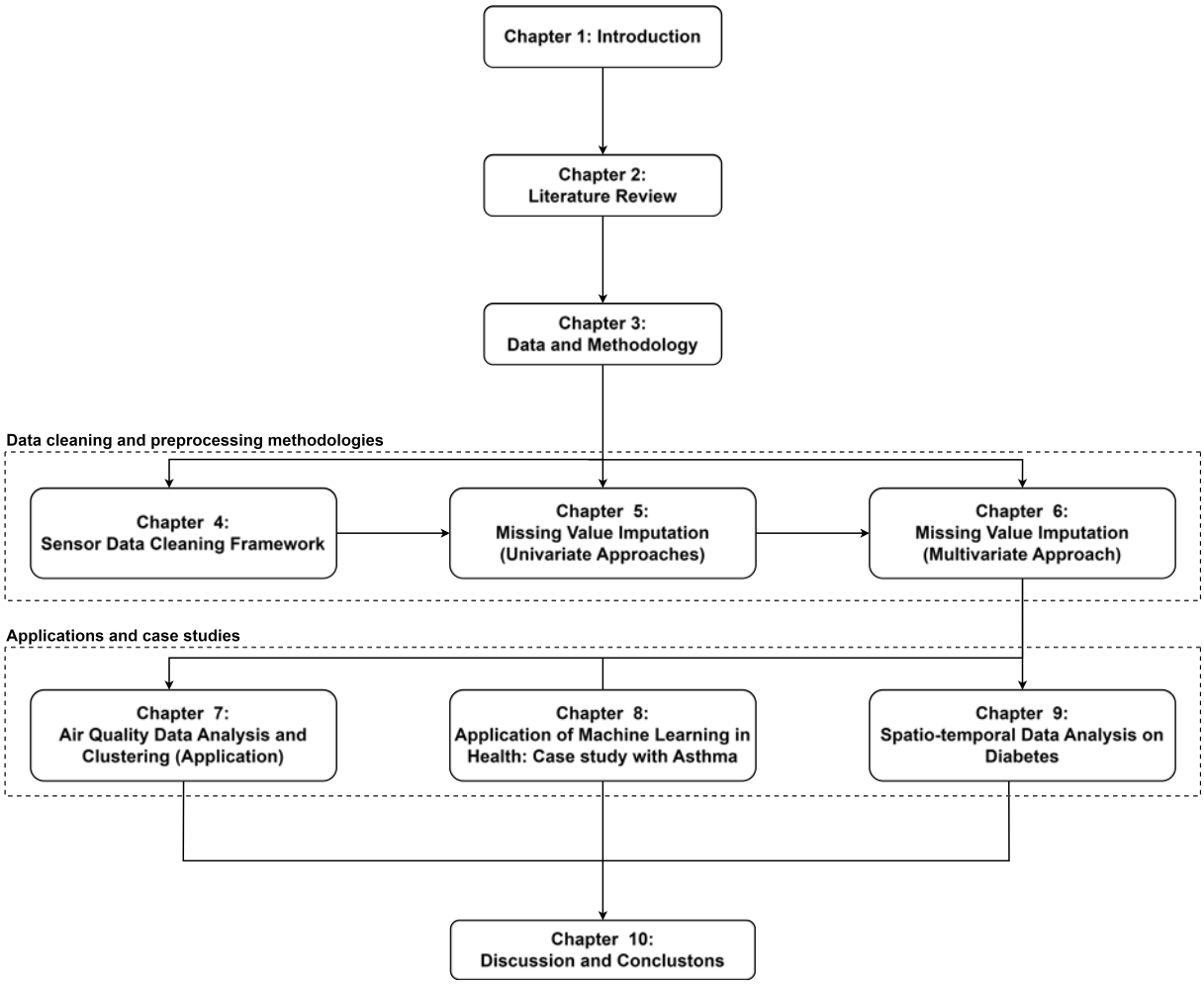


FIGURE 1.6: Overview of thesis structure

Chapter 2

Literature Review

This chapter explains the literature related to the research objectives. First, it explains the literature related to the problem of missing values, particularly in time series and then explains the literature related to the links between diabetes and environmental variables. Also, it includes a detailed discussion about the methods and applications of spatiotemporal models.

2.1 Objective 1: Theoretical

2.1.1 Identifying Missing Mechanism

There are three types of missing mechanisms, MCAR, MAR and MNAR, as stated in the Introduction chapter. The MCAR mechanism can be tested by comparing missing data subgroups using a series of independent t-tests (Dixon, 1988; Rantou, 2017). When one variable(y) has missing data in a multivariate dataset, comparing the distribution of fully observed variables between the groups (one group corresponding to missing y values and the other group corresponding to non-missing y values) helps to identify whether the missing mechanism is MCAR or not. Significant differences between means support the hypothesis that the data are not MCAR. This method is exhaustive as it has to conduct a massive number of t-tests to make sure that the dataset is MCAR, especially when the dataset is with a large number of variables (Little, 1988). Little (1988) proposed a single global test to check for the MCAR assumption, which is currently being widely used (Little, 1988; Moritz et al., 2015). The null distribution of

this test is a sum of functions of independent F statistics for small samples whereas for large samples it becomes an asymptotic chi-squared distribution. This test is most appropriate when the variables are quantitative. It assumes multivariate normality and is hence sensitive to the departures from normality (Little, 1988). Also, it may produce weak results when observed and missing groups are unbalanced and when the sample size is small (Little, 1988; Nakagawa, 2015).

Even though there are methods to detect the MCAR mechanism, manual analysis of patterns of the data and domain knowledge is needed to detect MAR and MNAR. Visualisation of missing data is still essential to identify and make assumptions about the missing mechanism. There is no method to distinguish MAR and MNAR (Nakagawa & Freckleton, 2008). Most studies are based on MCAR or MAR, and assumptions need to be made when the data are MNAR. Imputation algorithms usually perform better when the missing mechanism is ignorable, that is when MCAR or MAR. When it is non-ignorable (MNAR), special models are needed to identify why the data are missing and guess the likely values.

Identifying the missing mechanism is helpful in two ways. Firstly to select a proper imputation method and secondly to simulate data with similar missing mechanisms so that different imputation techniques can be compared using true data (Moritz et al., 2015).

2.1.2 Missing value handling techniques

The most convenient and frequent way to handle missing data is to remove them and analyse them using available observations. This is known as complete case analysis (Donders et al., 2006; Nakagawa, 2015), which is applicable in most of the situations where the missing mechanism is MCAR, and the percentage of missing values is as low as 5%. However, this will lead to incorrect parameter estimates if the missing mechanism is MAR or MNAR and creates estimation bias reducing the statistical power of the analysis (Nakagawa & Freckleton, 2008).

Other solutions for missing data are data augmentation and imputation. Data augmentation uses the information gained through assumptions and augments the parameter estimates, while imputation substitutes some values for missing observations. Maximum Likelihood (ML), expectation-maximization (EM), and Markov Chain Monte Carlo (MCMC) are some of the model-based data augmented procedures (Nakagawa & Freckleton, 2008). ML procedure assumes multivariate normal distribution for both target and predictor variables (Nakagawa & Freckleton, 2008).

Imputation is based on the idea that any observation can be replaced by a randomly selected observation from the same population of interest. It substitutes a missing value of a variable with a value drawn from an estimate of the distribution of this variable (Donders et al., 2006). Imputation can be further divided into two categories as single imputation and multiple imputation.

In single imputation, an estimation of the distribution of the variable with missing values will be made based on the available data. Randomly drawn values from this estimated distribution will be used to fill in the missing values. If the estimated distribution is identical to the true distribution of the variable, a single imputation procedure is equivalent to the direct replacement of a missing value by a randomly drawn value from the same population of interest. Single imputation uses one substitute for the missing observation and, hence, produces erroneous results about the uncertainty of parameter estimates. It tends to produce unbiased estimates of the population parameters; however, it leads to underestimating standard errors of the estimates and overestimating precision (Donders et al., 2006). Single imputation sometimes produces inaccurate estimates of parameter uncertainty even when ML, EM and MCMC procedures are used (Nakagawa & Freckleton, 2008).

In contrast, multiple imputation produces several substitutes based on random draws from several different underlying distributions of the variable with missing data. Therefore, it creates several completed datasets, and those datasets are then analysed separately, and parameter estimates are aggregated (Rubin, 1976; Donders et al., 2006). Standard errors of the parameter estimates are also aggregated and can be considered

as a measure of the uncertainty of the estimate. The standard deviation of the different parameter estimates corresponding to different datasets reflects the differences among imputed datasets. The same thing reflects the uncertainty in the estimated underlying distribution of the variable (Donders et al., 2006). One of the major advantages of the multiple imputation procedure is that it provides information about the impact of missing values on parameter estimates. This is recommended especially when the model selection is based on an information criterion such as Akaike's Information Criterion (AIC) (Nakagawa & Freckleton, 2008).

Figure 2.1 shows the schematic representation of single and multiple imputation methods.

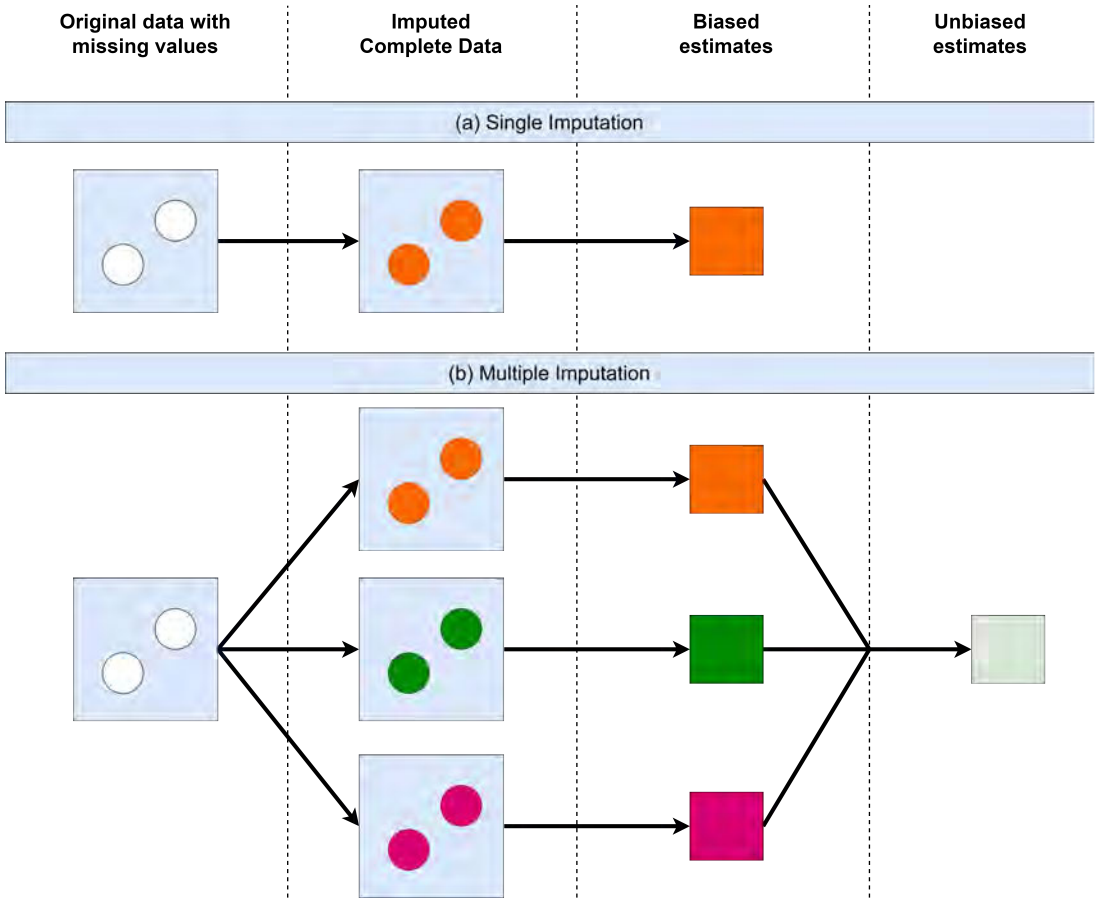


FIGURE 2.1: Illustration of imputation methods. Adopted from Nakagawa and Freckleton (2008).

2.1.3 Missing values in time series

In the case of univariate time series, time is considered as an implicit variable. Incorporating time series characteristics creates effective univariate imputation methods. Interpolation with seasonal Kalman filter from the zoo R-package or a linear interpolation on seasonal Leoss decomposed data from the forecast R-package are effective methods of dealing with missing data in univariate time series context (Moritz et al., 2015). Also, Kalman smoothing on structural time series models and Kalman smoothing on ARIMA (Auto Regressive Integrated Moving Average) models from the ImputeTS R-package appeared to perform well in MCAR situations (Wijesekara & Liyanage, 2020c). The mean imputation method has been widely used and performed well in most of the situations where the percentage of missing values is as low as 5% and especially in the single imputations (Norazian et al., 2008; Zakaria & Noor, 2018).

Widely used other methods to deal with time series missing data, in general, include Nearest Neighbor (Junger & De Leon, 2015; Zakaria & Noor, 2018; Junninen et al., 2004), Regression-based methods (Zakaria & Noor, 2018; Junninen et al., 2004), Self-Organizing Maps (SOM) and Multi-Layer Perceptron (MLP) (Junninen et al., 2004). When the data can be formulated as a multivariate normal time series, the Expectation-Maximization (EM) based methods appeared to perform well (Junger & De Leon, 2015). Moreover, attempts have been made to impute missing values by combining the power of neural networks and fuzzy logic in handling missing air quality data (Lei & Wan, 2010; Shahbazi et al., 2018). Recently, it can be seen that deep learning techniques such as Long Short Term Memory (LSTM) Recurrent Neural Networks are also used in missing value imputations in air quality data (Yuan et al., 2018). Most of these methods perform well when the percentage of missing values is as low as 5%.

Recovering a large interval of missing data which is very common in many real-world datasets, is still challenging. Traditional methods produce highly biased results as the gap size of the missing data increases (Chandrasekaran et al., 2016). Seasonal Moving Window Algorithm (SMWA) (Chandrasekaran et al., 2016) performs well with large gaps in the case of a univariate time series. This algorithm, first, decomposes the univariate

time series using STL decomposition (Seasonal-Trend decomposition procedure based on Loess) (Cleveland et al., 1990). Then, the trend component is linearly interpolated, and other components are fitted with the best-identified pattern from the past data. Finally, the recovered decomposed series are combined and transformed into a complete dataset. This SMWA algorithm works for both trend and seasonal types of data. Even though it is not possible to suggest a universal approach to deal with missing data, the need for recommending suitable methodologies based on the nature of the data is persisted.

2.2 Objective 2: Application

2.2.1 Diabetes and Weather

Researchers have found empirical evidence of the relationship between diabetes and environmental temperature. Booth et al. (2017) show that, the higher the temperature, the higher the chance of developing gestational diabetes mellitus, based on 12 years dataset in the Greater Toronto Area (Booth et al., 2017). They used Generalized Estimating Equations for the analysis and adjusted their findings for maternal age, parity, neighbourhood income quintile, world region and year. Further, they have found that each 10^0C increase in the 30-day average temperature is associated with 1.06 times higher odds of gestational diabetes (Booth et al., 2017).

Blauw et al. (2017) also have shown evidence for the relationship between diabetes and outdoor temperature by using the meta-regression method. After adjusting for the age, the diabetes incidence in the USA increases by 0.314 per 1000 for a 1^0C increase in temperature. Further, the worldwide glucose intolerant prevalence increases by 0.170 per 1000 for a 1^0C increase in temperature after adjusting for the obesity (Blauw et al., 2017). This method can be extended to any region by not only considering the temperature but also considering other pollution and weather variables to identify the overall impact.

There is a hypothesis that brown fat (which burns fat to generate heat) has a role to play in the linkage between diabetes and temperature. On the other hand, there are

arguments stating that brown fat does not play any considerable role in the above mechanism and there may be lots of other factors associated with the diabetes incidence (Howard, 2017). This makes it harder to explain the linkage between diabetes and temperature. However, the take-home message is that there is a significant correlation (association) between climate change and diabetes which need to be further investigated.

Further, some studies have found positive relations between type II diabetes and humidity. Multi-stage analysis based on convenient sampling from the elderly population was carried out from 2005 to 2011. Those who live in high-humidity areas tend to have a high proportion of people with diabetes. (Tyrovolas et al., 2014). Another study based on a group of elderly people followed up for five years also shows a positive relationship between type II diabetes and humidity.

2.2.2 Diabetes and Air Pollution

Evidence for the association between air pollution and diabetes also have been investigated during the last decade. It is found that fine particulate matter, PM2.5 and diabetes prevalence have a positive relationship in the USA, based on a study carried out by Pearson et al. (2010). They have used multivariate regression models and found out that for a $10\mu g/m^3$ increase in PM2.5 exposure, the diabetes prevalence rate increases by 1% (Pearson et al., 2010). Bowe et al. (2018) also carried out a longitudinal cohort study to show the positive relationship between type II diabetes and PM2.5 air pollution. They have used survival models in their analysis adjusting for sociodemographic and health characteristics (Bowe et al., 2018). Moreover, several other cohort studies also have shown higher type II diabetes risks associated with PM2.5 (Chen et al., 2013; Hansen et al., 2016; To et al., 2015 as cited in Dendup et al., 2018). In contrast, some studies have found negative relationships between type II diabetes and PM2.5 (Zanobetti et al., 2014). This study is a repeated measure study of sixty-four people with diabetes followed up to five years. Greater risk of type II diabetes is associated not only with particulate matter exposure but also with higher levels of NO exposure (Coogan et al., 2012; Krämer et al., 2010; Morland et al., 2006 as cited in

Dendup et al., 2018). Moreover, Ozone is also linked to an increased risk of type II diabetes (Jerrett et al., 2017). Hence, exploring the relationship between air pollutants on the risk of type II diabetes in Australia becomes a compelling motivation, considering that no studies have examined this relationship to the best of the author's knowledge.

2.2.3 Spatiotemporal Data Analysis

Spatiotemporal analysis is the analysis of data collected across time and space. An event that occurs at a certain time at a certain location describes the spatiotemporal properties. Measurements of climate variables are usually recorded at different time points at different locations. Therefore, these variables are spatiotemporal in nature. Spatiotemporal analysis is an emerging research area with a wide range of applications in fields such as Epidemiology, Ecology, Meteorology, Econometrics and Forestry. With the advancement of computational techniques, data on large spatial and temporal dimensions are available. This section investigates the existing methods and techniques in spatiotemporal data analysis.

A variety of methods have been used for spatiotemporal data for predictive purposes. Dadvand et al. (2011) has modelled weekly levels of ambient Black Smoke (BS) in Northeast London using a spatiotemporal model. Their model is based on Fanshawe's method of two stages, firstly to separate temporal trends and secondly, to predict BS levels at locations using a linear model with the result of the first stage and some other spatial features as predictors. They have used log transformation to stabilise the variance and random walk to smooth the seasonal patterns in the first stage to predict average weekly BS levels. Then, the regional weekly average BS levels have been modelled using the above-predicted values and some other spatial variables such as traffic, industrial activity, and population density as linear predictors (Dadvand et al., 2011). Even though their model has performed well with a high R2 value and the model assumptions were met, more measures need to be taken to justify its general applicability. Cross Validation approaches may be used to enrich the model validity. Spatio-Temporal Generalized Additive Models (ST-GAMs) are effective in identifying associated factors and predicting events. This model can account for a variety of data

types, such as spatial, temporal, geographic, and demographic data (Wang & Brown, 2012). Probabilistic approaches to model spatiotemporal patterns have been widely used in various fields such as criminology to identify crime rates and predict crimes (Law et al., 2014; Ratcliffe, 2002; Wang & Brown, 2012) and to predict biomass in ecosystems (Bénié et al., 2005).

In addition to predictive purposes, spatiotemporal data has been used with the intention of clustering, visualisation and information retrieval. Geographical Information System (GIS) based analyses have been carried out to identify clusters in many situations. Moran's I method of spatial autocorrelation, Getis-Ord Gi* statistics and point Kernel density functions were used to cluster road accidents in a South Indian city (Prasannakumar et al., 2011). Visualisation of spatiotemporal data has rapidly evolved with the contribution of Google Earth. Using Google Earth to visualise geographical output together with the power of Java3D-based tools to account for associations and variable distributions, has served as a milestone in visualising spatiotemporal patterns (Compieta et al., 2007). Traditionally, raster-based representation for location-based queries and vector-based representation for feature-based queries have been used for information retrieval. Later an Event-based Spatio-Temporal Data Model (ESTDM) has been developed to describe the temporal patterns along with the spatial patterns (Peuquet & Duan, 1995).

STARIMA model (Space-time Autoregressive Integrated Moving Average) can be used to model a single variable which changes over time and space. It is best suited for datasets with large temporal or spatial dimensions. It involves three stages; identification, estimation and diagnostic checking. Extensions of STARIMA models can be used to model instantaneous spatial terms and exogenous variables. Bayesian Vector Autoregressive (BVAR) models can be used for small spatial-scaled problems. Vector Autoregressive Moving Average (VARMA) models can be used in multivariate situations. However, this model is complex as the number of parameters to be estimated is very high (Kamarianakis & Prastacos, 2003). Seemingly Unrelated Regression (SUR) models also have been used in spatial time series. In these models, each location has a regression equation and covariance matrix which models geographical relationships.

In contrast, Spatial SUR models have regression equations for a certain period. This is used when the spatial dimension is larger than the temporal dimension (Kamarianakis & Prastacos, 2003).

Also, some researchers have tried to model and describe time series by using granular computing and fuzzy rules (Hryniewicz & Kaczmarek, 2016; Pedrycz et al., 2014; Roychowdhury & Pedrycz, 2002; Tu et al., 2015). These approaches use linguistic decision rules at different granularities like the human brain conceptualises the world. Recently the use of Deep Learning approaches also can be seen to model spatiotemporal modelling especially when dealing with nonlinear high dimensional predictors (Dixon et al., 2019).

Spatiotemporal analysis of disease rates is an evolving area which leads to identifying the geographical patterns of morbidity and mortality. Bayes methods were widely used in the analysis of disease rates. An attempt has been made to compare type I diabetes and leukaemia in the context of spatiotemporal similarities and dissimilarities by using the Bayesian approach (Manda et al., 2009). A similar approach was used to investigate the relationship between mortality and the incidence of Malaria in Western Kenya (Khagayi et al., 2017). Markov Chain Monte Carlo (MCMC) methods along with the Bayesian approach have been used to incorporate temporal effects and spatiotemporal interactions (Waller et al., 1997). Moreover, Bayesian Hierarchical Model (BHM) with a hidden dynamical Markov random field is also effectively used to model spatiotemporal sudden-infant-death rates (Zhuang & Cressie, 2012). Generalised Additive Mixed Models also have been used in this context with Conditionally Autoregressive (CAR) smoothing in the spatial dimension and B-spline smoothing in the temporal dimension to model diseases (Torabi, 2013). Some studies have used Integrated Nested Laplace Approximation (INLA) for Bayesian parameter estimation in spatiotemporal modelling (Martínez-Bello et al., 2018). Also, the spatiotemporal models were used to explain patterns of infectious diseases over time by using Negative Binomial models with both seasonal and autoregressive components (Sharmin & Rayhan, 2012).

Although the modelling of spatiotemporal data has evolved through applying various

stochastic and deterministic models, not much attention has been given to model validation. In most of the models, the whole data set has been used for the model building and hence model validation has been overly performed. However, when dealing with spatiotemporal modelling with environment variables, it is very important to consider the model's validity for a previously unseen data set. Underestimated predictive errors may lead to building excessive confidence in the models and therefore, applying those models to new scenarios might incur unnecessary costs. Few studies have looked into this and suggested improved cross-validation (CV) approaches to prevent underestimated predictive errors. However, this needs to be done carefully taking into consideration the temporal and spatial autocorrelation of the data. Also, consideration has to be given in the modelling when the model assumptions are not satisfied. Non-random cross-validation approaches by blocking spatial, temporal group and hierarchical variables to increase model validity have been suggested recently (Roberts et al., 2017). A similar approach along with a forward feature selection method has been used to improve the performance of spatiotemporal machine learning models (Meyer et al., 2018). However, further development of cross-validation approaches may be useful.

According to the available literature, it is clear that methods to handle multi-variable spatiotemporal data are very limited and complex. The existence of multi-collinearity makes the process even more complex. Therefore, modelling pollution, meteorological, demography and health variables with spatiotemporal properties will produce novel knowledge.

Chapter 3

Data and Methodology

This chapter provides a concise summary of the data used for the thesis and overall methodology. To ensure clarity, the subsequent chapters will delve into specific methods as and when necessary.

3.1 Data

This research involves secondary data collected through different sources.

3.1.1 Air pollution and Weather data

Air pollution and weather data recorded using sensors from 1994 to 2018 at 18 monitoring sites in the Sydney region were downloaded manually from the website of the Department of Planning and Environment, NSW Government. A list of variables is given in Table 3.1. Subsequently, air pollution and weather data for the entire NSW area, covering the period from 2003 to 2020, were downloaded through an API (Application Programming Interface) available on the same website (Planning and Environment, 2020; Riley et al., 2020).

3.1.2 Demography data

Population, gender, and age data (2016 census) for different statistical areas in the NSW state were downloaded from the website of the Australian Bureau of Statistics (ABS) (Australian Bureau of Statistics, 2017).

TABLE 3.1: Air pollution and weather variables collected at NSW monitoring sites

Pollutant variables:	Meteorological variables:
Ozone - O3 (hourly)	Wind speed
Ozone - O3 (rolling 4 hour)	Wind direction
Nitric oxide - NO	Sigma Theta (SD1)
Nitrogen dioxide - NO2	Air temperature
Visibility - nephelometer	Relative humidity
Carbon monoxide - CO (hourly)	Global solar radiation
Carbon monoxide - CO (rolling 8 hours)	Rainfall
Sulfur dioxide - SO2	
Particles - PM10	
Particles - PM2.5	
Ammonia	

3.1.3 Spatial data

Shapefiles containing the geometry of different statistical areas were downloaded through the website of the ABS (Australian Bureau of Statistics, 2021).

3.1.4 Health data

NSW Admitted Patient Data Collection (APDC) was accessed through the data integration platform (Liyanage & Liyanage, 2010) in Western Sydney University. This data set was requested and obtained from the Centre for Epidemiology and Evidence of the NSW Ministry of Health.

3.2 Methodology

Figure 3.1 illustrates the methodology used in achieving the objectives of this research.

First, air pollution and weather data were collected through web scraping and using API. These data were explored extensively to identify the patterns of missing data. This process was exhausting as there were a lot of variables and sites. Also, it is important to visualise when and where the missing data occur with different frequencies (hourly, daily, annually, etc.). Therefore a dashboard was developed connecting the API. Then the nature of the missingness was explored and the applicability of the existing methods in handling missing data was evaluated. When the performance of

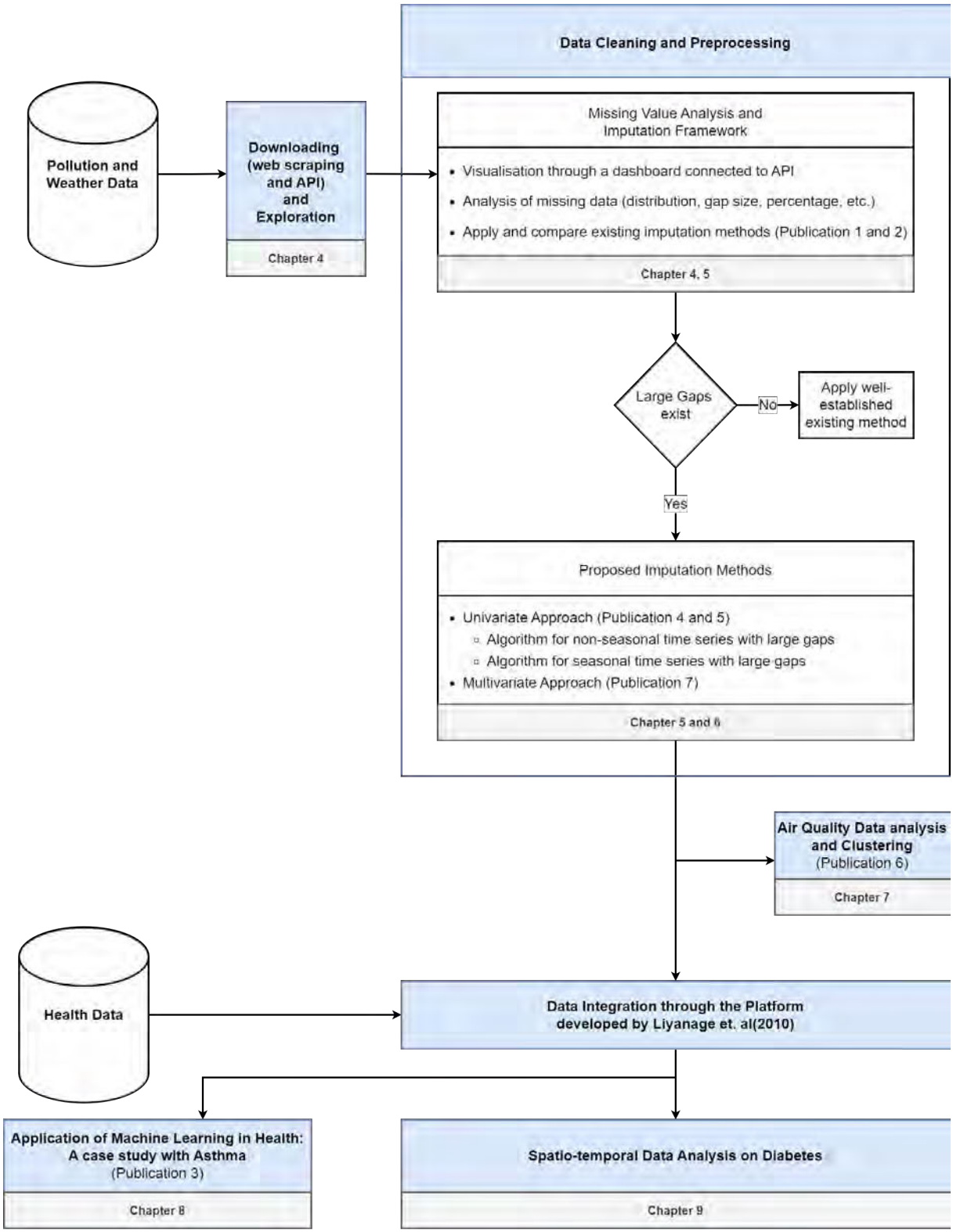


FIGURE 3.1: Overview of thesis structure

existing methods was not appealing, new algorithms were developed and evaluated. Then, an air quality data analysis was carried out by applying those methods. Finally, cleaned data were linked with health data and a spatiotemporal analysis was carried out for the type II hospital admissions for the period of 2013-2018. Quasi-Poison regression models were used to identify the associations of environmental variables with type II diabetes. More details of the data cleaning process will be given in the proceeding chapters with the relevant methods.

Chapter 4

Sensor Data Cleaning Framework

This chapter provides more details of the air pollution and weather data used in this thesis.

4.1 Data collection

Pollution and Weather data were collected through the website of the Department of Planning and Environment, NSW Government. As part of the Enhance Air Quality Website and Data Delivery (EWADD) project, air quality and meteorological data from the Air Quality Monitoring Network are being made available to download. There are three facilities to download the data.

- Data download facility: This provides a graphical interface to search and download historical data.
- Application Programming Interface (API): This provides the capability to stream data as well as search and download historical data.
- Air quality data explorer: This is a map-based data search and download platform.

Initially, only the data download facility was available. Air pollution and weather variable data for the Sydney Region were downloaded manually using this facility. Later, the data for the NSW region were downloaded using the API facility. Figure 4.1 shows the locations of all the monitoring sites belonging to the NSW monitoring network.

However, some of the sites do not measure all the variables and some are not functioning.

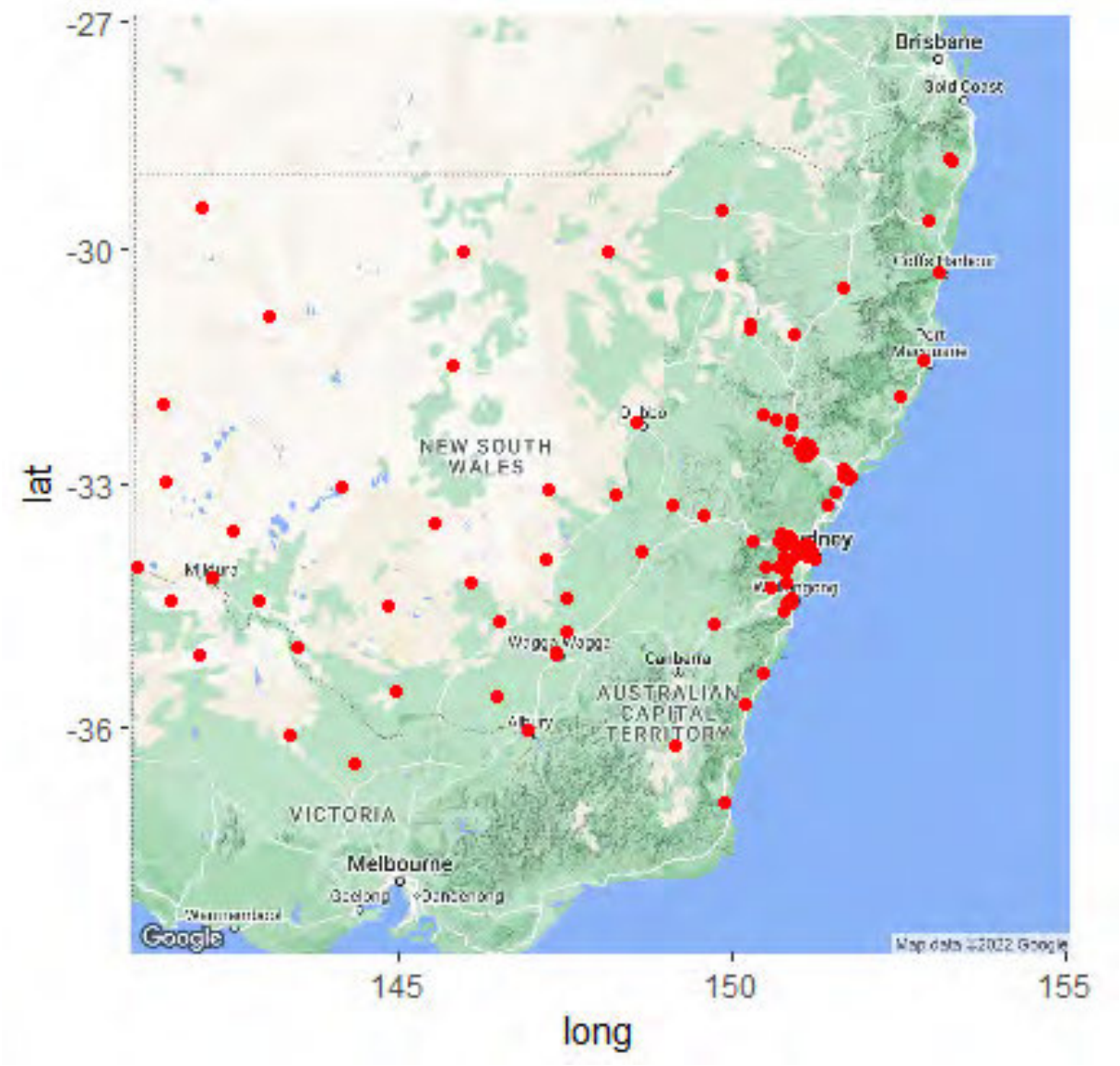


FIGURE 4.1: NSW Pollution Sites

4.2 Data Cleaning Framework

Visualisation of air pollution and weather data was required to understand the nature of the missingness. For this purpose, a dashboard was created with several facilities. This framework obtains NSW pollution data from an Azure Cloud Data Warehouse of the Department of Planning, Industry and Environment (DPIE) through an Application

Programming Interface (API) (State of NSW and Department of Planning, Industry and Environment, 2020) developed by DPIE (Figure 4.2). The development of this framework has been done using R programming language (R Core Team, 2020) and Shiny R package (Chang et al., 2020).

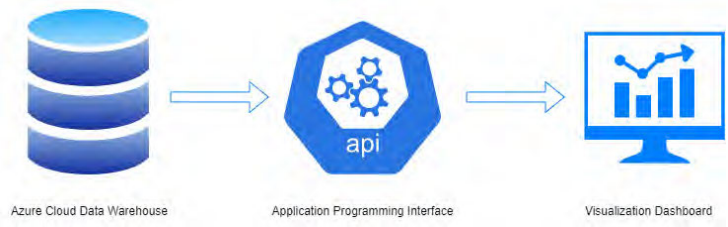


FIGURE 4.2: Pollution and Weather Data Collection

4.2.1 Basic visualisation facility

Figure 4.3a shows a screenshot of the framework. First, it allows the user to select the time series (pollution/weather variables) through a panel of drop-down lists on the left side, as shown in Figure 4.3a. Description of the user inputs are given in Table 4.1

TABLE 4.1: Description of the user inputs

Site details	Select region	List of regions in NSW
	Select station	List of sites in the selected region
Variable details	Select variable	List of pollution variables
	Select category	List of available categories (averages, maximum, exceedances etc.)
	Select sub-category	List of available sub-categories (hourly, daily, monthly, annual etc.
	Select frequency	List of available frequencies (24-average, 8-hour rolling average etc.)
Select time range	A calendar to select the start and end of the series	

It visualises the selected series in the Imputation tab of the framework as an interactive time series plot with a time slider so that the user can have a closer look at any interesting patterns as shown in Figure 4.3b. It also provides some useful statistics about the selected series. The missing percentage gives the total number of missing values as a percentage of the total number of observations in the selected time series. The longest gap indicates the size of the largest gap with consecutive missing values while the most frequent gap shows the size of the most frequently occurring gap size. In addition, it shows some useful graphs in the Missingness tab which is very helpful in the data understanding and preparation process. The top left bar chart in Figure 4.3b shows the number of occurrences of each gap size in red-coloured bars while the

total number of missing observations under each gap size in blue-coloured bars. The top right chart shows the interactive original series. The bottom left graph shows the positions of missing observations in red coloured vertical lines while the bottom right chart shows a histogram of missing values for successive intervals. These plots were created using imputeTS package (Moritz & Bartz-Beielstein, 2017a).

4.2.2 Basic Imputation facility

This framework provides a set of well-established imputation methods from imputeTS package as it is proven to be more efficient in the recent literature(Moritz & Bartz-Beielstein, 2017a; Moritz et al., 2015; Wijesekara & Liyanage, 2020c) to impute univariate time series variables. It currently includes methods namely mean imputation, linear interpolation, spline interpolation, exponentially weighted moving average, Kalman smoothing on structural time series models and Kalman smoothing on Autoregressive Integrated Moving Average(ARIMA) models. The imputed values can be seen in the dashboard directly below the original series in red colour while the observed values are shown in blue colour (Figure 4.4 and Figure 4.5).

Performance Evaluation

One of the major challenges in imputation is that there is no way to measure the accuracy of the imputed series as the actual values are unknown. Typically comparisons of imputation techniques are done using a reference series with no missing values and artificially creating missing values. Then the missing values are replaced using several methods and their performances are compared. However, the imputation method is highly dependent on the nature of the data. Even though the ground truth is unknown, this framework provides a facility to compare the performance of different methods using a reference series of similar nature. Depending on the percentage of missing values of the selected time series, it will create missing values in the reference series and apply the imputation technique and calculate the Mean Absolute Percentage Error (MAPE) as a measure of performance. To create the missing values in the reference series the simulation procedure suggested by Moritz et al. (Moritz et al., 2015) for

their comparison of imputation methods in univariate time series was applied using an exponential distribution. The rate parameter of the exponential distribution is the rate of missing values in the selected time series. The missing mechanism is assumed to be Missing Completely at Random (MCAR). Even though this may not provide an accurate measure of the performance, this can be used to compare the performance of several methods. Moreover, as it shows the graphical representation of the imputed series, the user can decide on a suitable imputation method rather than just using a single method and relying on that. This helps researchers to utilise their valuable time on their original research question reducing the time and effort in the data cleaning process.

Downloading the complete data

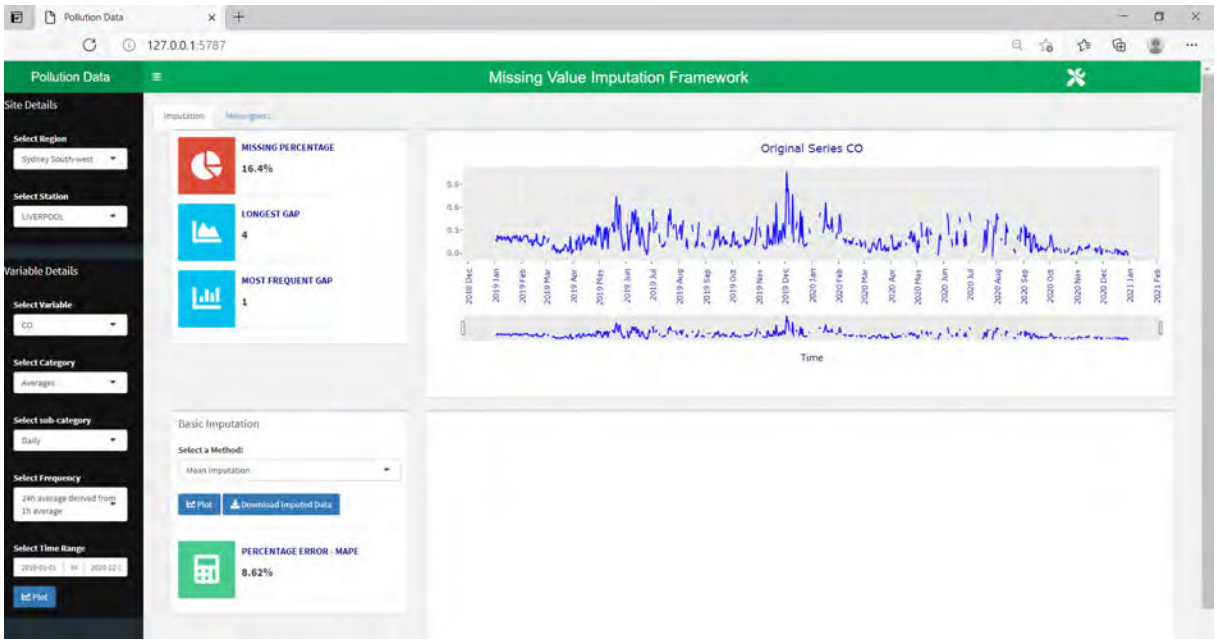
After investigating the missing mechanism visually, and comparing the applicability of some of the well-established methods, the user can download the imputed dataset from this framework. The project can be accessible via,
https://gitops.westernsydney.edu.au/18570263/data-cleaning-framework.git.

Limitation

As, this framework is directly linked with the API of NSW weather and pollution data, this cannot be used for other datasets. However, it is expected to extend this to incorporate those facilities in the future. Currently, it includes a few well-established time series imputation methods. It is expected to implement new algorithms in the future.

4.3 Contribution

The main contribution of this chapter is the development of a data-cleaning framework. It connects the existing data imputation algorithms to clean data and it is available to use by any researcher who uses Sydney air pollution and weather data. Through the graphical user interface, one can apply existing methods without having to learn R codes. It also provides the capability to visualise patterns of missing data, so that one



(a) Screen shot of the developed framework: Visualisation

(b) Screen shot of the developed framework: Missingness

FIGURE 4.3: Visualisation facility of the framework

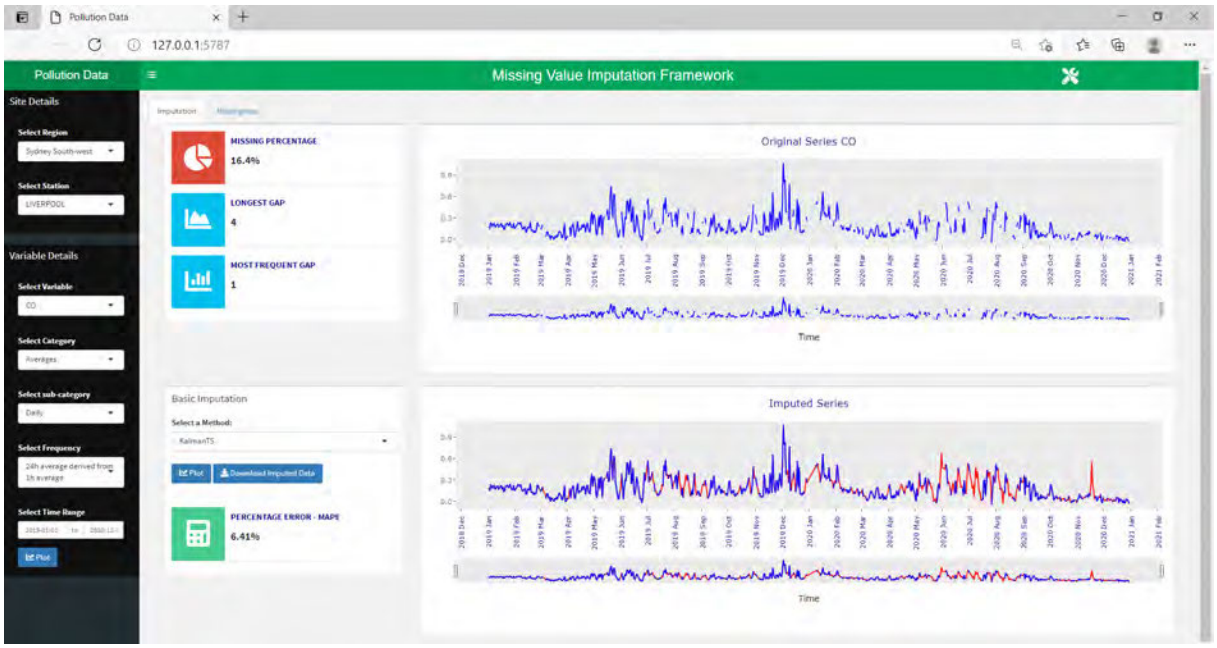
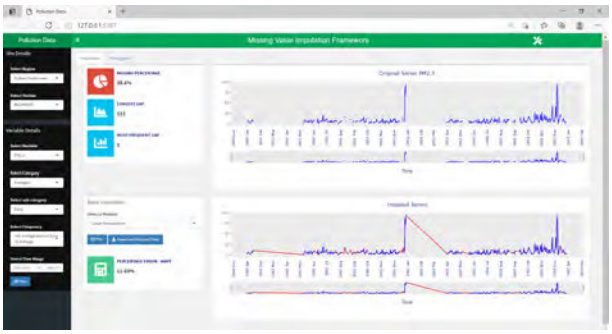


FIGURE 4.4: Spline interpolation

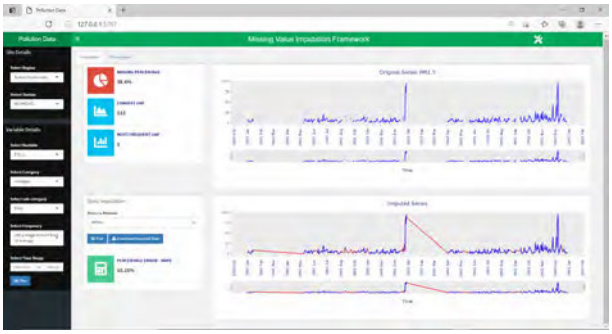
can analyse them prior to selecting a methodology. Moreover, after applying a method, one can visually as well as statistically compare the performances. This framework is demonstrated using NSW weather and pollution data as explained in section 4.2 and it can be extended to other data in general.

(a) Mean imputation



(b) Linear interpolation

(c) Spline interpolation



(d) Kalman smoothing on ARIMA models

(e) Kalman smoothing on structural time series models

(f) Exponentially weighted moving average

FIGURE 4.5: Deferent types of imputation methods

Chapter 5

Missing Value Imputation: Univariate Approaches

This chapter includes a comparison of time series imputation methods on air quality data. Also, it gives two improved algorithms to impute missing values; one for non-seasonal air quality variables and another one for seasonal weather variables. This chapter is based on the following three publications.

- Section 5.1:

  Wijesekara, W. M. L. K. N., and Liyanage, L. (2020, March). Comparison of imputation methods for missing values in air pollution data: Case study on Sydney air quality index. In Future of Information and Communication Conference (pp. 257-269). Springer, Cham.

  https://doi.org/10.1007/978-3-030-39442-4_20

- Section 5.2:

  Wijesekara, L., and Liyanage, L. (2021, November). Air quality data pre-processing: A novel algorithm to impute missing values in univariate time series. In 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 996-1001). IEEE.

  https://doi.org/10.1109/ICTAI52525.2021.00159

- Section 5.3:

  Wijesekara, L., and Liyanage, L. (2021). Imputing Large Gaps of High-resolution

Environment Temperature. In 2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS) (pp. 74-79). IEEE.
https://doi.org/10.1109/ICIIS53135.2021.9660672

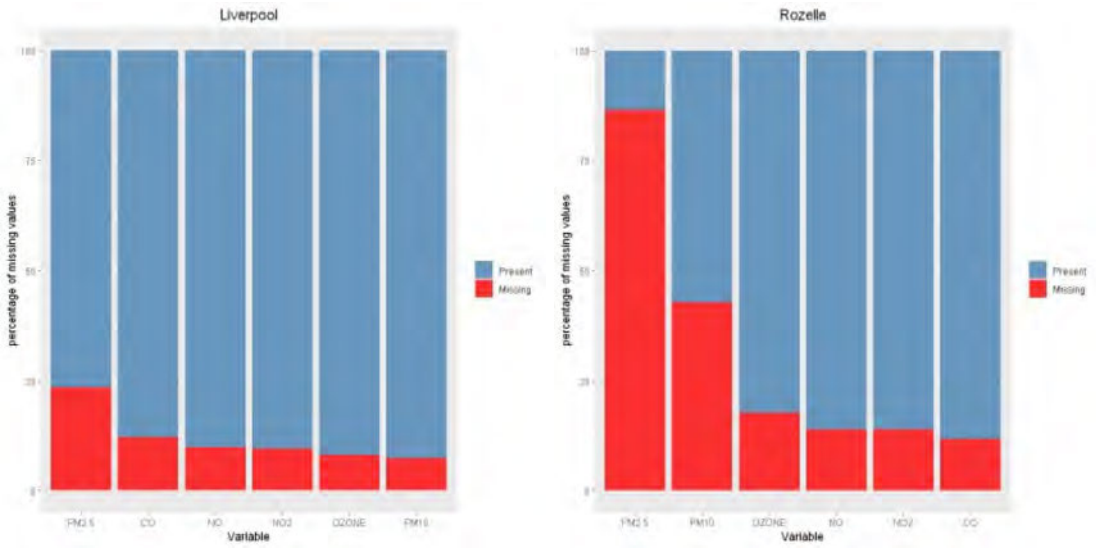
5.1 Comaparison of Imputation methods: Case Study on Sydney Air Quality Index

5.1.1 Introduction

Air quality data is widely used in models for various purposes including assessing the impact of air quality on health and well-being. It is common to expect a large number of missing values in data sources collected using sensors. Missing values of air quality data may lead to underestimating the associated health effects. In general, missing values create problems by introducing a substantial amount of bias and reducing the efficiency of analysis (Nakagawa & Freckleton, 2008). Although many methods have been developed for missing value imputations, methods for time series data are still in their infancy. The inherent nature of auto-correlation, trend, seasonality and cyclic effects have made the process more challenging. However, it is necessary to impute missing values in certain situations to make more accurate predictions. Therefore, it is an important area of research. In this study, we focus on the air pollution data in the Sydney region of Australia. Figure 5.1 summarises the percentage of missing data on air pollutant variables at two monitoring stations (Liverpool and Rozelle) from 1994 to 2018.

Substantial percentages of missing values are present in almost all pollutant variables. Liverpool station has lower percentages of missing data than Rozelle. Even though only two stations are presented here, all other stations also showed a considerable percentage of missing values which we cannot disregard. Therefore, a proper mechanism to impute these missing values is essential before any type of modelling.

In this section, we discuss six well-established methods of dealing with missing values in a univariate time series context and compare their performance on imputing



(a) Liverpool (b) Rozelle

FIGURE 5.1: Missing value percentages of pollutant variables in two monitoring stations in Sydney

missing values for air quality data in the Sydney region. The methods discussed here are Mean Imputation, Spline Interpolation, Simple Moving Average, Exponentially Weighted Moving Average, Kalman Smoothing on Structural Time Series Models and Kalman Smoothing on ARIMA models. The performances of these methods were compared with three performance measures; Mean Squared Error (MSE), Coefficient of Determination (R2) and Index of Agreement (d). The objective of this section is to identify the best available imputing method for air quality data in the Sydney region.

5.1.2 Related research

A variety of methods ranging from simple methods such as mean imputation to advanced methods such as Long Short Term Memory (LSTM) Recurrent Neural Networks have been applied to impute missing values in the context of air pollution data. Mean imputation methods have performed well in most situations where the percentage of missing values is as low as 5% and especially in single imputations (Norazian et al., 2008; Zakaria & Noor, 2018). Other widely used methods include interpolations (linear, quadratic and cubic), Nearest Neighbor(NN) (Junninen et al., 2004; Norazian et al., 2008), Regression-based methods (Junninen et al., 2004; Wyzga, 1973), Self-Organizing Maps (SOM) and Multi-Layer Perceptron (MLP) (Junninen et

al., 2004). When the data can be formulated as a multivariate normal time series, the Expectation-Maximization (EM) based methods appeared to perform well (Junger & De Leon, 2015). Moreover, attempts have been made by combining the power of the neural network and fuzzy logic in handling missing air quality data (Lei & Wan, 2010; Shahbazi et al., 2018). These methods have been recommended for nonlinear and complex phenomena. One such method is the hybrid approach of Multiple Imputation (MI) and Adaptive Nero-Fuzzy Inference System (ANFIS). Recently, it can be seen that deep learning techniques such as LSTM Recurrent Neural Networks are also used in missing value imputations in air quality data (Yuan et al., 2018). However, the simplest technique, mean imputation is still dominant in this area and it is considered the best method in some scenarios.

The most commonly used performance measures for comparing missing data imputation methods are Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and the Coefficient of Determination (R2). The Index of Agreement (d) also has been used in some studies. There is no universal method to measure the performance of imputation techniques. Methods widely depend on the nature of data and the distribution of missing values.

There are three types of missing data mechanisms identified as Missing Completely at Random (MCAR), Missing at Random (MAT) and Missing Not at Random (MNAR) (Rubright et al., 2014). In the MCAR scenario, the missingness is independent of both observable and unobservable parameters of interest. In MAR, there is a systematic relationship between the propensity of a value to be missing and the observed data while in MNAR there is a relationship between the propensity of a value to be missing and its unobserved value.

5.1.3 Methodology

Mean Imputation

This is the most commonly used single imputation technique where the missing values are replaced with the mean value of the variable. The mean of a series of values

$y_1, y_2, ..., y_n$ is given by

$$\bar{y} = \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right). \tag{5.1}$$

Spline Interpolation

For $n+1$ pairs of observations $(t_i, y_i) : i = 0, 1, ..., n)$, the shape of the spline is modelled by interpolating between all the pairs of observations (t_{i-1}, y_{i-1}) and (t_i, y_i) with polynomials

$$y = q_i(t), i = 1, 2, \ldots, n. \tag{5.2}$$

Simple Moving Average

Simple moving average for a series Y at t is given by the unweighted mean of the previous n observations:

$$\bar{y}_{ma} = \frac{1}{n}\left(\sum_{i=0}^{n-1} y_{t-(n-i)}\right). \tag{5.3}$$

Exponentially Weighted Moving Average

Exponentially weighted moving average for a series Y at any time t may be calculated recursively by

$$s_t = \begin{cases} y_1, & t = 1, \\ \alpha \cdot y_t + (1-\alpha).s_{t-1}, & t > 1, \end{cases} \tag{5.4}$$

where α denotes the degree of weighting, representing the decreasing effect of y variable with respect to time.

Autoregressive Integrated Moving Average (ARIMA) model

When the process is stationary, an Autoregressive Moving Average model ARMA(p,q) can be defined as

$$y_t = c + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}, \tag{5.5}$$

where φ_is are the autoregressive parameters, θ_js are the moving average parameters to be estimated, and ε_t is a white noise with mean zero and constant variance (σ^2).

Time series which needs to be differenced to be stationary is said to be an "integrated" version of a stationary series. In ARIMA(p,d,q) model, the number of autoregressive terms, the number of non-seasonal differences and the number of lagged forecast errors in the prediction equation are denoted by p, d and q respectively.

Structural Time Series Models

All linear time series have a state space representation. This representation relates the disturbance vector ε_t to the observation vector y_t via a Markov process α_t. A convenient expression of the state space form is

$$\begin{aligned} y_t &= \mathbf{Z}_t\alpha_t + \varepsilon_t, \quad \varepsilon_t \sim N\left(\mathbf{0}, \mathbf{H}_t\right), \\ \alpha_t &= \mathbf{T}_t\alpha_{t-1} + \boldsymbol{R}_t\eta_t, \quad \eta_t \sim \left(\mathbf{0}, \mathbf{Q}_t\right), t = 1, \ldots n, \end{aligned} \tag{5.6}$$

where y_t is a $p \times 1$ vector of observations and α_t is an unobserved $m \times 1$ vector called the state vector. The system matrices Z_t, T_t and R_t have dimensions $p \times m$, $m \times m$ and $m \times g$ respectively. The disturbance terms ε_t and η_t are assumed to be serially independent and independent of each other at all time points. The matrix H_t has dimension $p \times p$ with rank p, and the matrix Q_t has dimension $g \times g$ with rank $g \leq m$ (Abril, 2011).

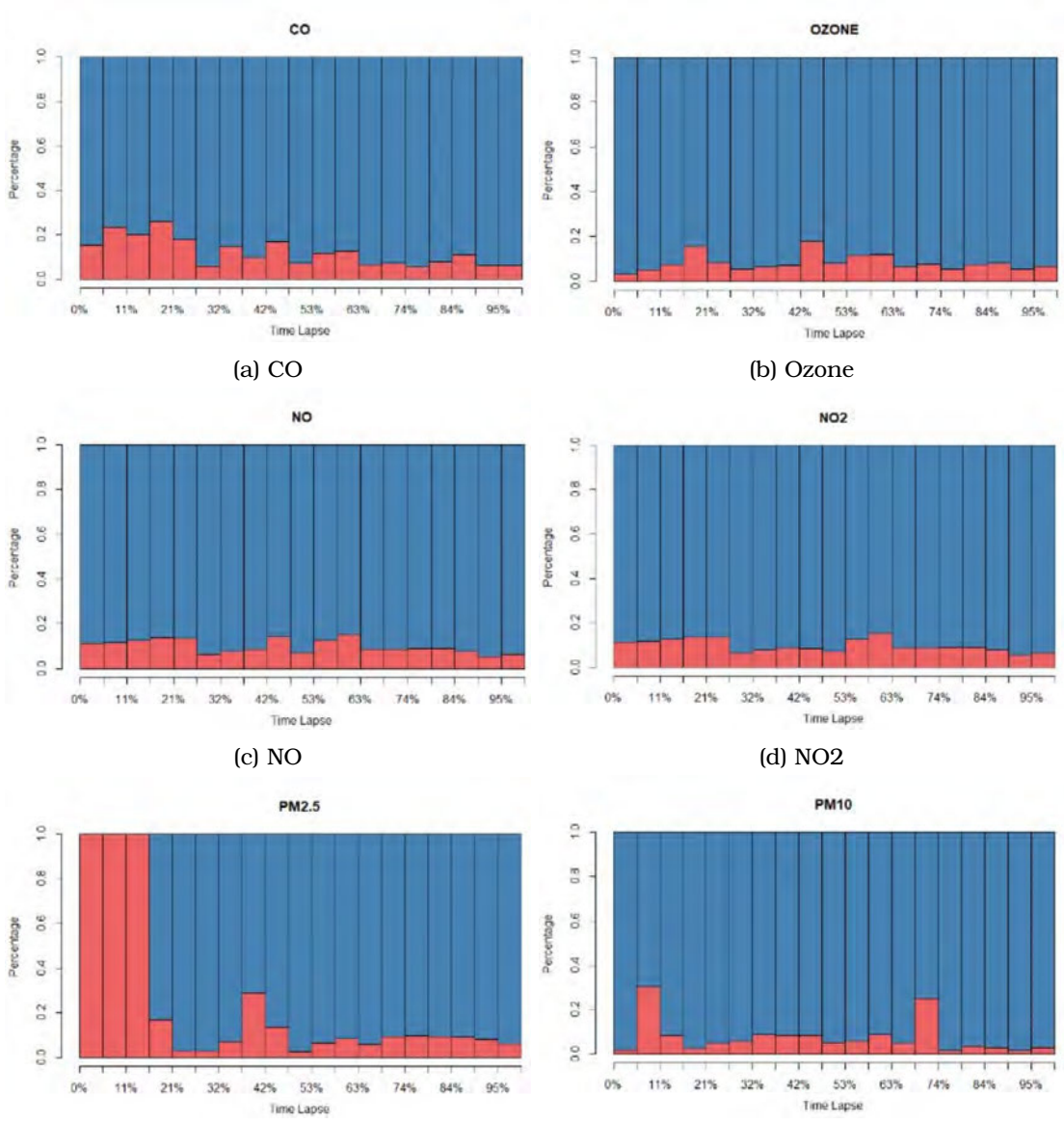
Kalman Smoothing

Kalman filter calculates the mean and variance of the unobserved state, given the observations. This filter is a recursive algorithm; the current best estimate is updated whenever a new observation is obtained. Kalman Smoothing takes the form of a backwards recursion and it can be used to compute smoothed estimator of the disturbance vector (Abril, 2011).

Data and Approach

This data includes air pollutant variables measured hourly at each of the monitoring stations in the Sydney region. Figure 5.2 depicts how the missing values of air pollutants at the Liverpool monitoring station are distributed across the period from 1994 to 2018. The red coloured bars represent the percentage of missing values and the

blue colour bars represent the percentage of observed values over a given period. The x-axis represents the cumulative percentage of values in the time series. As can be seen, the missingness is almost random across the time except for the first few years in PM2.5 (fine particulate matter 2.5 micrometres or less in diameter) and CO. Therefore, in this study, simulations were carried out considering the MCAR mechanism.



(a) CO (b) Ozone

(c) NO (d) NO2

(e) PM2.5 (f) PM10

FIGURE 5.2: Percentage of missing values of air pollutants at Liverpool station over the time from 1994 to 2018

To select a reference time series to carry out missing value simulations and to use as the ground truth in the comparison of imputation methods, the Air Quality Index

(a standard index calculated by incorporating all the air pollutants) recorded at each station was considered. Figure 5.3 shows the heat map of the number of missing values in the hourly air quality index for each monitoring station from 1994-01-01 01:00:00 AEST to 2018-12-31 24:00:00 AEST.

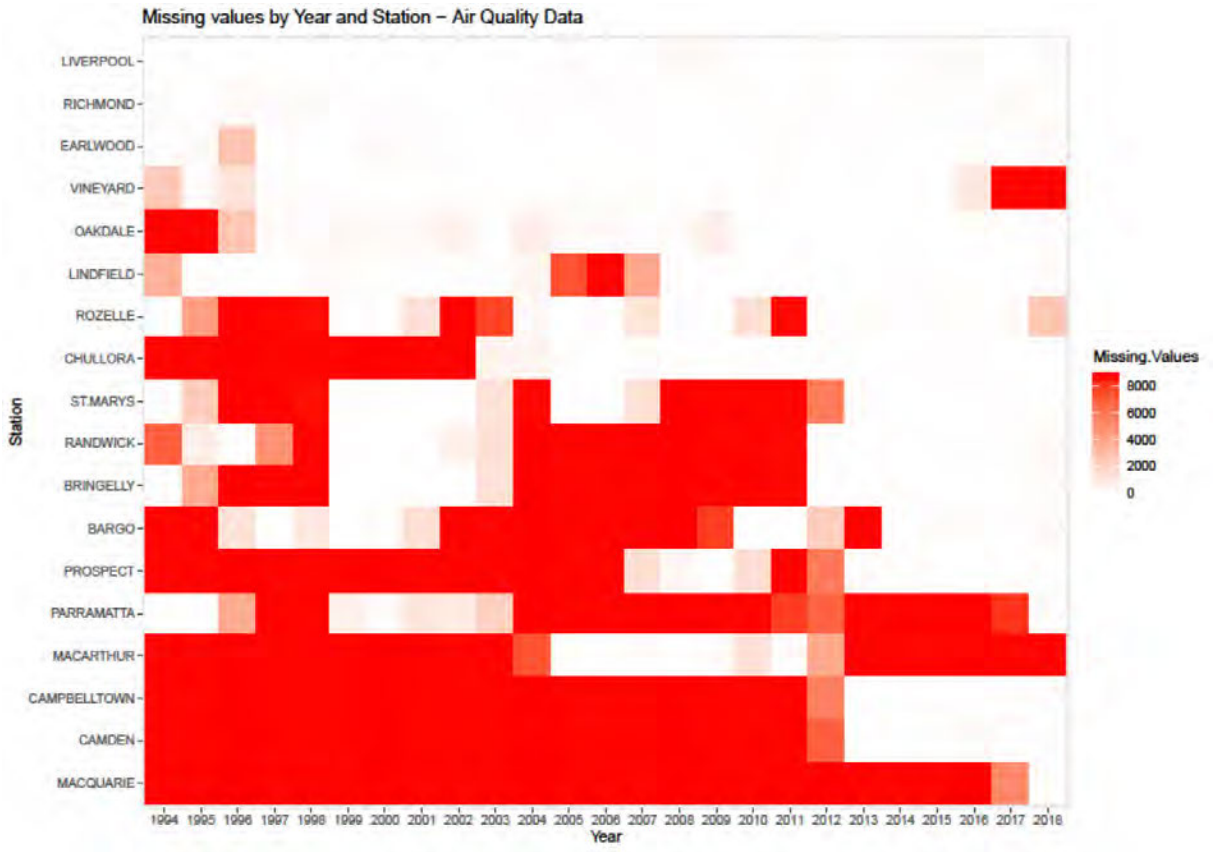


FIGURE 5.3: Number of missing values of the air quality monitoring stations in Sydney from 1994 to 2018

As can be seen in Figure 5.3, the dataset suffers from the problem of missing values. Liverpool, Richmond and Earlwood stations appeared to have less number of missing values. Therefore, these three stations were further analysed, and a subset of Earlwood hourly air quality indices for a two-year period starting from 2014.01.01 01:00:00 AEST to 2015.12.31 24:00:00 AEST with no missing values (Figure 5.4) was selected as the reference series.

Missing values for this series were created by artificially deleting observations under the Missing Completely at Random (MCAR) mechanism. Four scenarios were created where the percentages of missing values were 5%, 10%, 15% and 20%. The missing

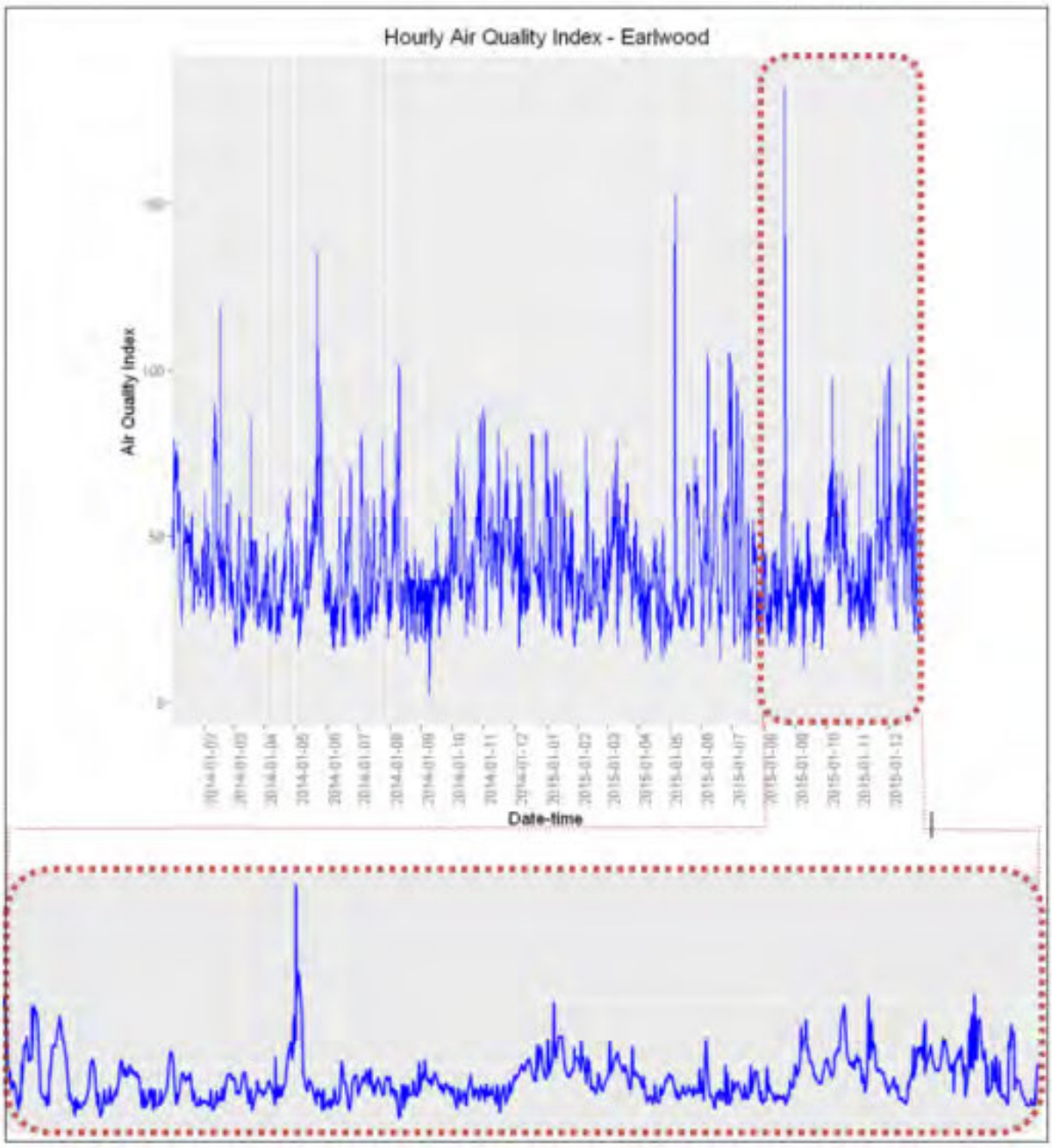


FIGURE 5.4: Distribution of hourly air quality data from 2014.01.01 01:00:00 AEST to 2015.12.31 24:00:00 AEST in Earlwood

values of each scenario were imputed by using the six methods; Mean Imputation, Spline Interpolation, Simple Moving Average, Exponentially Weighted Moving Average, Kalman Smoothing on Structural Time Series Models and Kalman Smoothing on ARIMA models. Then the performance of each method in each scenario was assessed by using the three performance measures; Mean Squared Error (MSE), Coefficient of Determination (R2) and Index of Agreement (d).

5.1.4 Results and Discussion

Figure 5.5 shows the position missing values in the time series in each of the four scenarios as mentioned in the section 5.1.3. The vertical red lines indicate the positions of missing values. Only the first 1,000 observations out of 17,517 observations are displayed for ease of viewing.

The six methods as stated in the section 5.1.3 were used to impute missing values which were artificially deleted under simulations. The performance of each method was evaluated with three measures MSE, R^2, and d.

Table 5.1 shows the performance of the six methods measured by MSE. Since the performance of the Mean Imputation method was poor compared to other methods, Figure 5.6 compares the performance of the other five methods except for Mean Imputation.

Table 5.1: MSE measures

Method	**5%**	**10%**	**15%**	**20%**
Spline Interpolation	0.406	0.954	1.128	2.020
Kalman Smoothing Structural TS	0.250	0.600	0.790	0.980
Kalman Smoothing ARIMA	0.273	0.648	0.789	0.979
Simple MA	0.479	0.926	1.321	1.777
Exp MA	0.297	0.664	0.931	1.199
Mean	12.915	25.936	37.624	48.378

The Kalman Smoothing on Structural Time Series method appeared to be the best while Mean Imputation appeared to be the worst. When the percentage of missing values increases, the performance of all the methods decreases. Kalman Smoothing on ARIMA models and Exponentially Weighted Moving Averages perform well for small percentages of missing values.

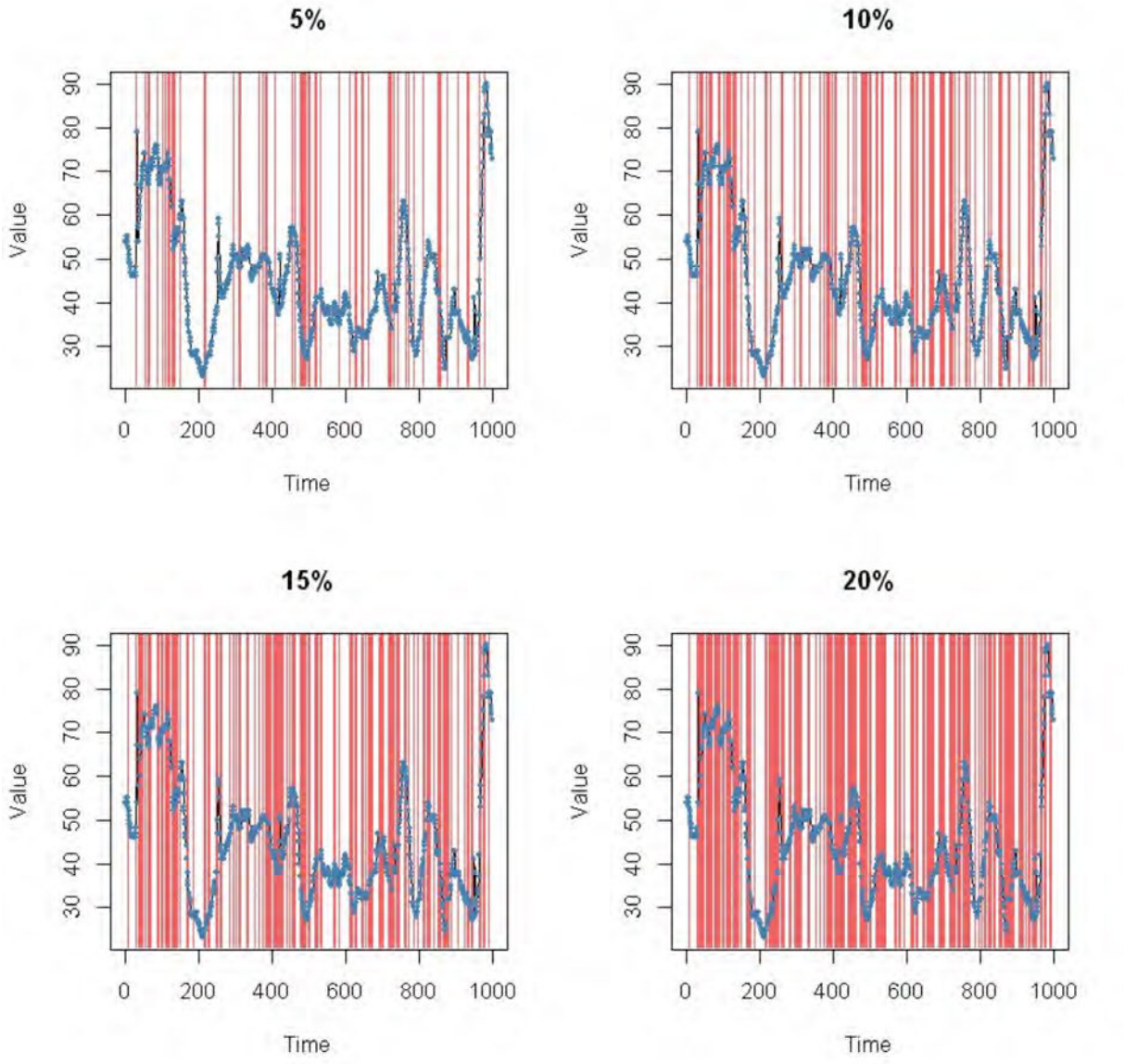


FIGURE 5.5: Distribution of missing values in the simulations for 5%, 10%, 15% and 20% of missing values in the dataset

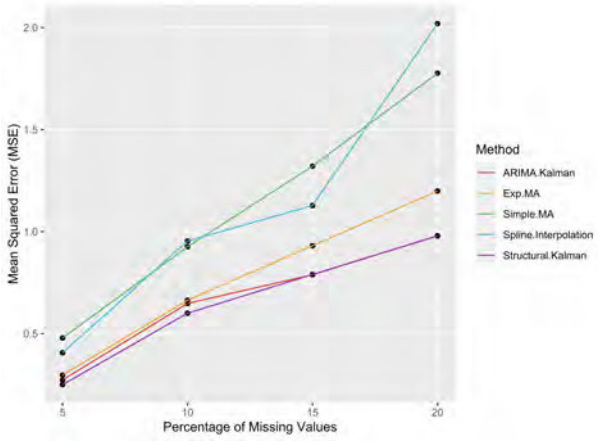


FIGURE 5.6: Comparison of MSE measures of each method for the 5%, 10%, 15% and 20% missing value scenarios

Table 5.2 shows the performance of the six methods measured by R^2. Figure 5.7 displays the performance of methods except for Mean Imputation.

TABLE 5.2: R^2 measures

Method	**5%**	**10%**	**15%**	**20%**
Spline Interpolation	0.999	0.998	0.998	0.996
Kalman Smoothing Structural TS	0.999	0.999	0.998	0.998
Kalman Smoothing ARIMA	0.999	0.999	0.998	0.998
Simple MA	0.999	0.998	0.997	0.996
Exp MA	0.999	0.999	0.998	0.997
Mean	0.973	0.944	0.918	0.893

Again the Kalman Smoothing on Structural Time Series method appears to be the best among the considered methods. The Kalman Smoothing on ARIMA and Exponentially Weighted Moving Averages methods also perform well. However, once again the performance of all the methods decreases as the percentage of missing values increases.

Table 5.3 gives the performance of methods measured by the Index of Agreement (d). Figure 5.8 shows the performance of methods excluding Mean Imputation in order to compare the other models clearly.

The Kalman Smoothing on Structural Time Series models and on ARIMA models perform equally well. Although Spline interpolation performed well with a smaller percentage of missing values, its performance drastically decreases with increasing missing values. Except for the Mean Imputation, all other methods show approximately equal

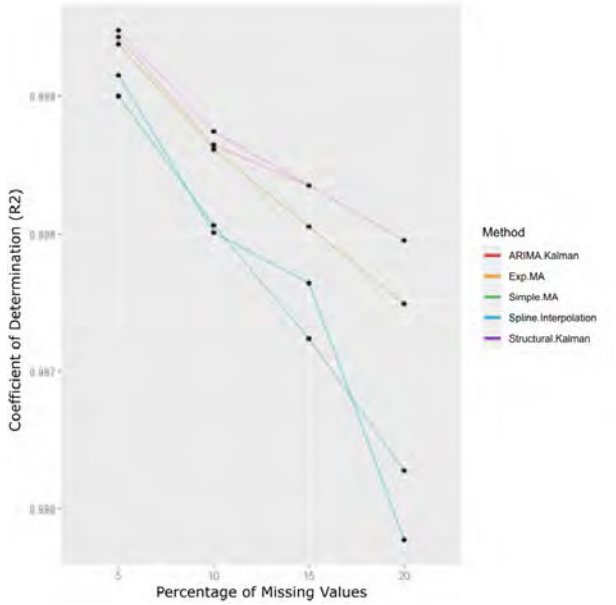


FIGURE 5.7: Comparison of R2 measures of each method for the 5%, 10%, 15% and 20% missing value scenarios

TABLE 5.3: Index of Agreement measures

Method	**5%**	**10%**	**15%**	**20%**
Spline Interpolation	0.999	0.999	0.998	0.997
Kalman Smoothing Structural TS	0.999	0.999	0.999	0.998
Kalman Smoothing ARIMA	0.999	0.999	0.999	0.998
Simple MA	0.999	0.999	0.998	0.998
Exp MA	0.999	0.999	0.999	0.998
Mean	0.985	0.970	0.955	0.940

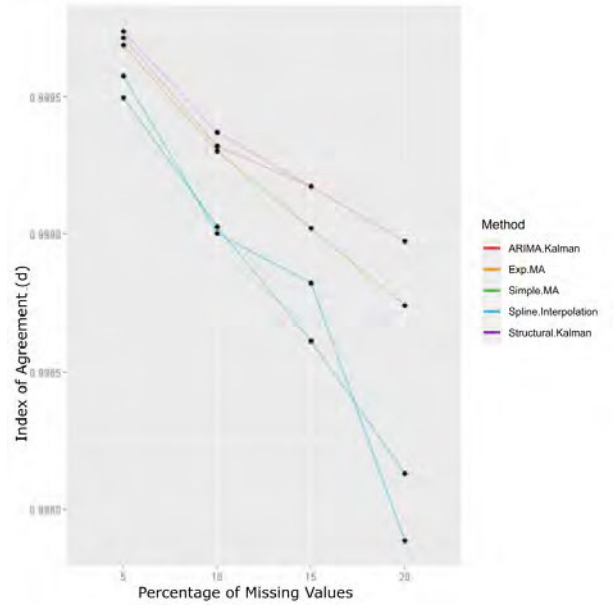


FIGURE 5.8: Comparison of Index of Agreement measures of each method for the 5%, 10%, 15% and 20% missing value scenarios

performances. It is clear that the performance of all the methods decreases when the percentage of missing values increases.

Figure 5.9 exhibits the imputed values from the Kalman Smoothing on Structural time series model for the four simulated scenarios. Red, green and blue represents the imputed values, actual values and known values respectively. Again, only the first 1,000 observations out of 17,517 observations are presented for ease of viewing.

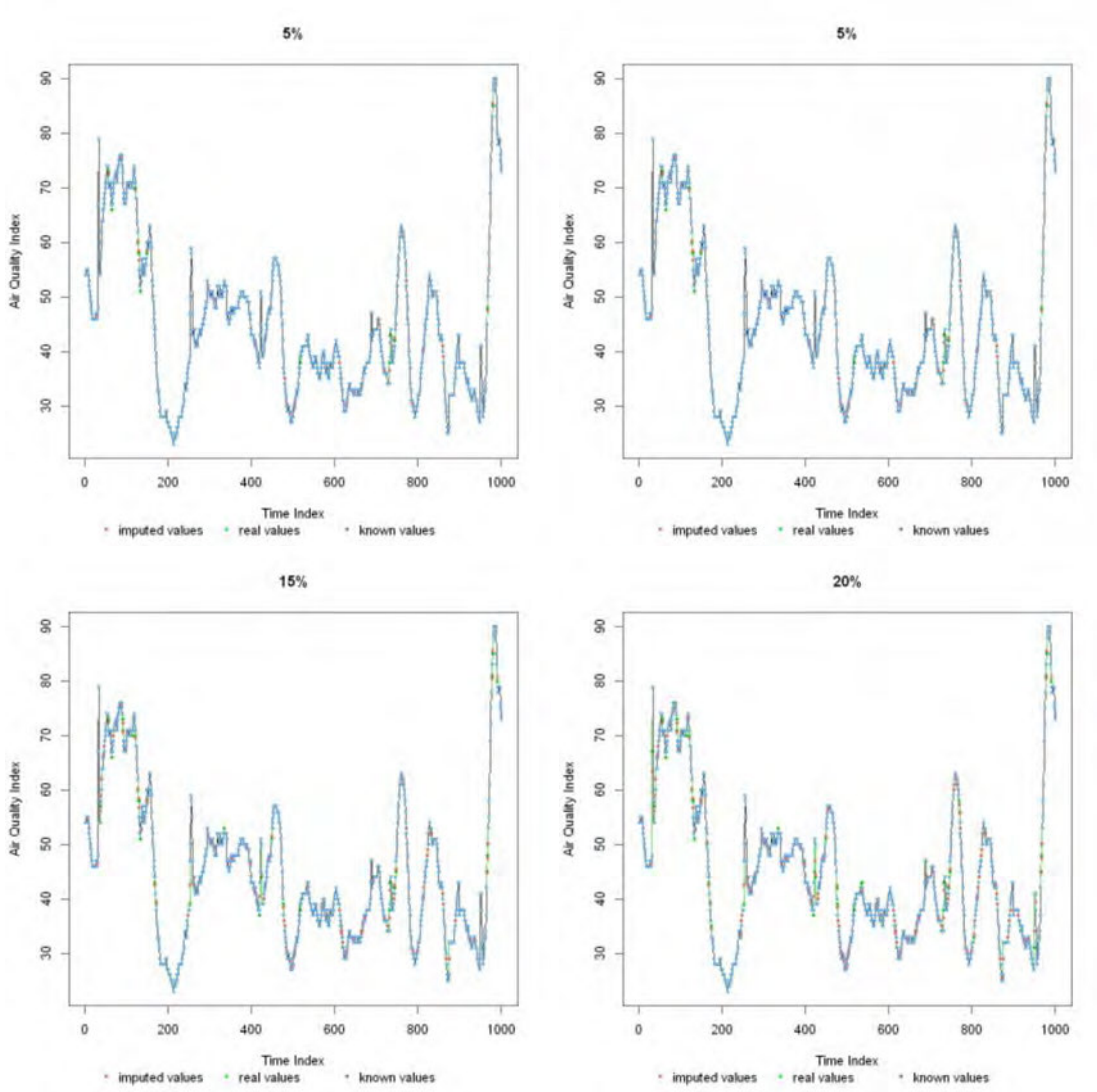


FIGURE 5.9: Comparison of imputed data using Kalman Smoothing on Structural Time Series models against the actual data for the for the 5%, 10%, 15% and 20% missing value scenarios

It can be seen that this method has performed extremely well for MCAR missing mechanism in the air quality data. However, further studies must be carried out to compare

the performance of these methods under MAR and MNAR missing mechanisms. Also, here we have considered a subset of observed series to artificially create missing values. When there are large numbers of missing values, these methods may result in sub-optimal results.

5.1.5 Conclusion and Recommendations

Among the six methods considered, Kalman Smoothing Method on Structural Time Series is the best method for imputing missing values in the context of air quality data where the missing mechanism is MCAR. Kalman Smoothing on ARIMA, and Exponentially Weighted Moving Average methods also perform considerably well. The performance of Spline Interpolation decreases drastically with an increased percentage of missing values. Even though Mean Imputation performs reasonably well for smaller percentages of missing data, all the other five methods outperform this method regardless of the number of missing values. The six methods can be ranked from best to worst as; Kalman Smoothing on Structural Time Series Models, Kalman Smoothing on ARIMA models, Exponentially Weighted Moving Average, Simple Moving Average, Spline Interpolation and Mean Imputation. However, the need of developing an improved method of imputation to deal with a higher percentage of missing values persists. These methods also need to be studied against other missing data mechanisms such as Missing at Random (MAR) and Missing Not at Random (MNAR).

5.2 Algorithm 1: Air quality data pre-processing: Novel algorithm to impute missing values in univariate time series

5.2.1 Introduction

The objective of section is to develop a robust algorithm to impute missing values of air pollution time series in the Sydney region. The imputed data will be used in analyses which aim to measure the impact of air pollution on population health. The motivation for this study is that even though there are well-established methods in univariate time series imputations, there is a trade-off between the simplicity and the accuracy

of these methods. The aim is to propose a relatively simple but efficient methodology to deal with missing values in time series, particularly air pollution data. This method could be used by the researchers to pre-process air pollution data and then carry out downstream analyses efficiently.

Algorithm 1 : Forward-Backward Imputation **(FBImp)**

Input: Univariate time series with gaps ($T = \{t_1, t_2, ..., t_n\}$) Step size ($N$)

1: N=step size ▷ this will be passed to Regimp
2: $T_F \leftarrow T$ ▷ forward series
3: $T_B \leftarrow reverse(T)$ ▷ backward series
4: $\hat{T}_F \leftarrow$ **RegImp**(T_F, N)
5: $\hat{T}_B \leftarrow$ **RegImp**(T_B, N)
6: $\hat{T_B}' \leftarrow reverse(\hat{T}_B)$
7: $\hat{T} \leftarrow average(\hat{T}_F, \hat{T_B}')$

Output: Univariate complete time series ($\hat{T}$)

The proposed methodology consists of two parts, Forward-backward imputation (FBImp) and Regularized Regression based imputation (RegImp). This algorithm basically produces a value for a missing observation based on a set of immediate predecessors. The difference of this method from the auto-regressive time series models is that it does the predictions in both forward and backward directions as well as it uses regularised regression models. First, the time series with missing values (NAs) $t_1, t_2, ..., t_n$ is processed through the RegImp algorithm and then apply the same for the reversed series $t_n, t_n - 1, ..., t_1$. Finally, the imputed values for both the forward and reversed series are averaged and then form a complete time series as shown in the Fig. 5.10

A schematic representation of the RegImp algorithm is given in Figure 5.11. It is needed to decide on a step size (N), before running this algorithm. This is done by investigating the partial autocorrelation plots after applying some approximations and selecting the autoregressive term accordingly. In the RegImp algorithm, first, the time series is converted into multivariate time series with N predictors (TX) and a response (TY).

Algorithm 2 : Regularized Regression-based Imputation **(RegImp)**

Input: Univariate time series with gaps ($T = \{t_1, t_2, ..., t_n\}$)

$T_1 \leftarrow$ Time series from the beginning of $T$to the point where first consecutive N steps occurs
$T_2 \leftarrow$ Time series from the point where first consecutive N steps occur to the end of T
Prepare multivariate dataset ▷ (TX and TY)
$X \leftarrow$ Remove missing rows of TX
$Y \leftarrow$ Remove corresponding rows of TY
procedure RegularizedRegression(X, Y)
    Split data into two: TrainData and TestData
    Fit the Lasso Regression model to TrainData ▷ Tune the model using 10-fold cross validation
    Fit the Ridge Regression model to TrainData
    Fit the Elastic-Net Regression model to TrainData
    $MSE_L \leftarrow$ MSE for TestData using Lasso model
    $MSE_R \leftarrow$ MSE for TestData using Lasso model
    $MSE_E \leftarrow$ MSE for TestData using Lasso model
    BestModel $\leftarrow$ model with lowest MSE
return BestModel
while No missing values in T2 **do**
    predicted missing values based on previous N values using BestModel
$\hat{T}_2 \leftarrow$ imputed/predicted T_2
$\hat{T} \leftarrow$ Concatinate $T_1, \hat{T}_2$

Output: Univariate time series ($\hat{T}$)

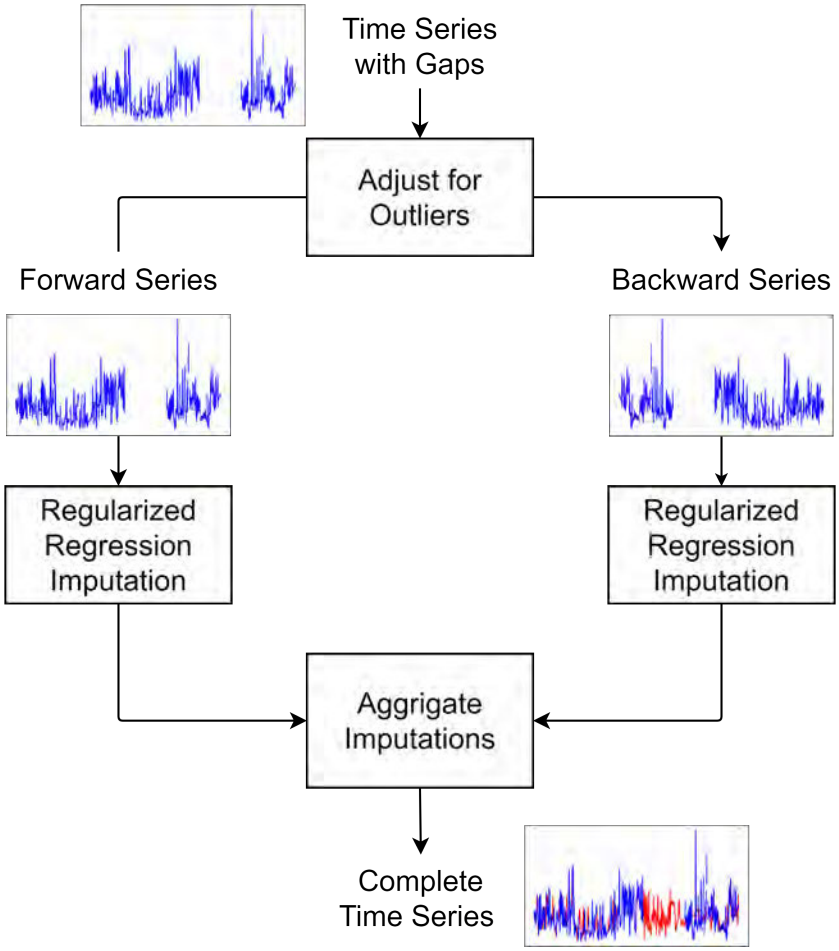


FIGURE 5.10: Forward-Backward Imputation

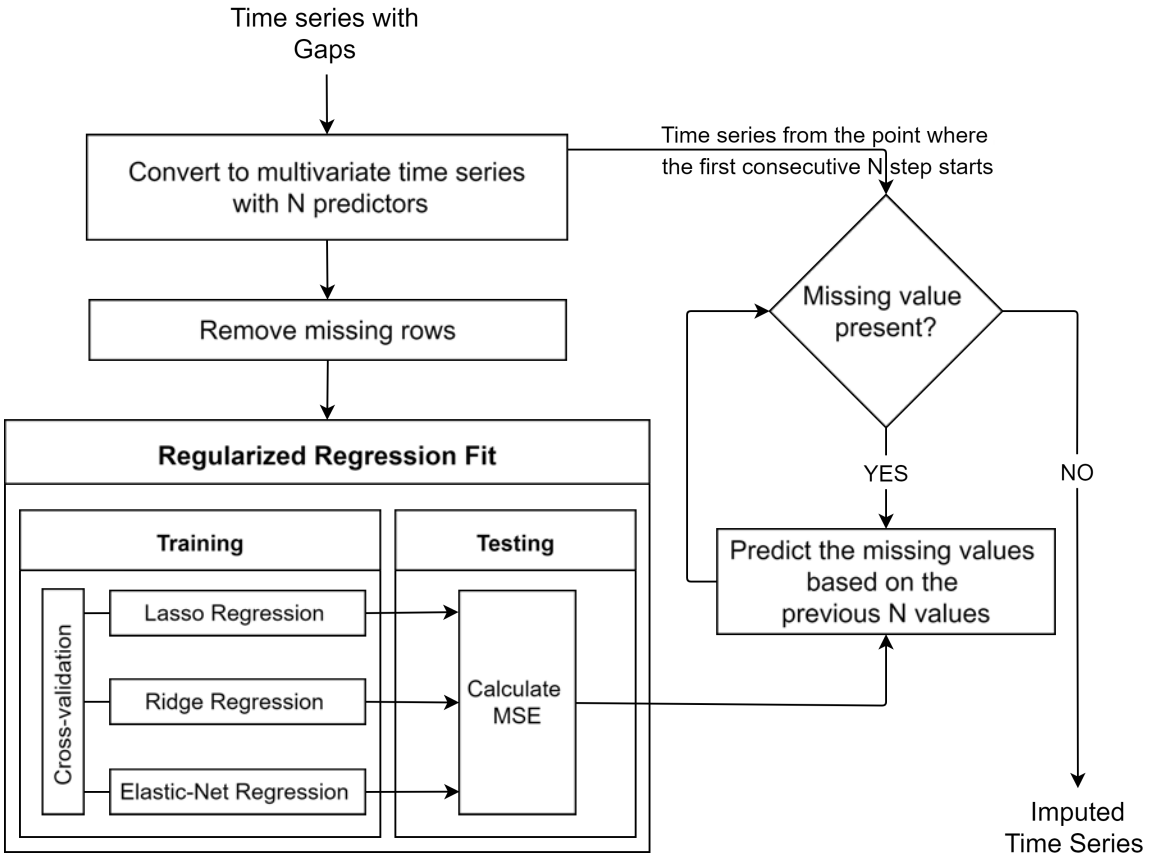


FIGURE 5.11: Regularized Regression based Imputation

$$TX = \begin{bmatrix} t_1 & t_2 & t_3 & \cdots & t_N \\ t_2 & t_3 & t_4 & \cdots & t_{N+1} \\ t_3 & t_4 & t_5 & \cdots & t_{N+2} \\ \vdots & \vdots & \vdots & & \vdots \\ t_{n-N} & t_{n-(N-1)} & t_{n-(N-2)} & \cdots & t_{n-1} \end{bmatrix}_{(n-N)\times N}$$

$$TY = \begin{bmatrix} t_{N+1} \\ t_{N+2} \\ t_{N+3} \\ \vdots \\ t_n \end{bmatrix}_{(n-N)\times 1}$$

After removing the rows with NAs in TX and appending a column with 1s at the beginning, the design matrix (X) can be created as below:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1N} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mN} \end{bmatrix}_{m\times(N+1),}$$

where m is the number of complete rows in TX. Also, the corresponding rows in TY are removed. Let $\beta_j s$ be the regression parameters of the regression model with N number of predictors and $\epsilon_i s$ be the errors.

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix}_{m\times 1} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_N \end{bmatrix}_{(N+1)\times 1} \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_m \end{bmatrix}_{m\times 1}$$

$$Y = X\beta + \epsilon \tag{5.7}$$

$$y_i = \beta_o + \sum_{j=1}^{N} \beta_j x_{ij} + \epsilon_i \ ; \ i = 1, 2, 3, \ldots, m \tag{5.8}$$

$$\hat{y}_i = \hat{\beta}_o + \sum_{j=1}^{N} \hat{\beta}_j x_{ij} \tag{5.9}$$

Let $||Y - X\hat{\beta}||^2 = \sum_{i=1}^{m} \left(y_i - \beta_o - \sum_{j=1}^{N} \beta_j x_{ij} \right)^2$, $|\beta| = \sum_{j=1}^{N} |\beta_j|$ and $||\beta||^2 = \sum_{j=1}^{N} \beta_j^2$.

The best parameters for the three types of regularisation models can be given below:

Lasso regression model,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||Y - X\beta||^2 + \lambda|\beta|, \tag{5.10}$$

Ridge regression model,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||Y - X\beta||^2 + \lambda||\beta||^2, \tag{5.11}$$

and

Elastic-Net regression model,

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} ||Y - X\beta||^2 + \lambda_1|\beta| + \lambda_2||\beta||^2, \tag{5.12}$$

where λ, λ_1 and λ_2 are penalty parameters/hyper-parameters.

After splitting the complete dataset into training and testing, all three regularisation models are applied to the training dataset. The best model under each regularised model is selected by tuning the hyper-parameters using 10-fold cross-validation. Then the prediction accuracy is measured using the testing dataset and the model with the lowest mean square error (MSE) is selected for the next step.

Then the time series from the point where the first consecutive N-steps occurs is selected and the next value will be predicted based on the previous consecutive N steps. After repeating this process until the end of the series, the imputed series is

created. However, NAs may still persist from the beginning of the series to the point where the first consecutive N-steps occur. To overcome this issue, the same procedure is repeated for the reversed series and then aggregated.

The reasons for applying regression models are because of their low complexity compared to models like Neural Networks and Deep learning. Further, the complex methods may need extensive training to be able to perform better and since this is a data pre-processing activity, such training may be expensive. Moreover, the regularised methods are proven to be efficient to reduce the variance and avoid issues of multicollinearity in the literature. (Ying, 2019)

5.2.2 Datasets

In order to evaluate the performance of the proposed algorithm, it is necessary to have a complete dataset to consider as the ground truth. However, it is almost impossible to have a complete dataset of air quality data. After examining pollution variables for a 20-year period from 2000 to 2020 at 18 selected sites in the Sydney region, three datasets were selected with a low percentage of missing values. Air Quality Index daily data from 2014-01-01 to 2015-12-31, daily PM10 data (particles less than 10 microns in $\mu g/m^3$) from 2016-01-01 to 2019-01-01 collected at Earlwood site and daily CO data from 2010-01-01 to 2012-12-31 collected at Liverpool site was selected. Data is available on the web site of the New South Wales government Planning, Industry and Environment department (https://www.dpie.nsw.gov.au/air-quality/air-quality-data-services). Since the missing percentages were less than 3% in these datasets, Kalman smoothing was applied for these series and treated as the ground truth. The reason for not selecting other available complete time series data for this study is that the focus of this study was to assess the imputation accuracy of the proposed method for air pollution data and not the time series data in general.

5.2.3 Simulation Procedure

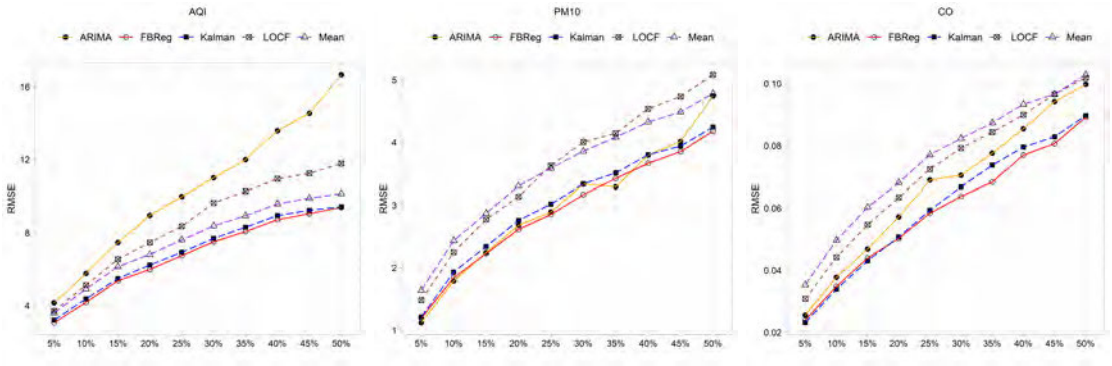
It is required to artificially create missing values to compare the performance of the proposed algorithm. A simulation study based on different missing data mechanisms

was conducted for this. To emulate the MCAR mechanism, a simulation procedure suggested by Moritz et al. (Moritz et al., 2015) for their comparison of imputation methods in univariate time series was applied. Missing data were generated by considering the exponential distribution. By using values ranging from 0.05 to 0.5 for the rate parameter of the exponential distribution, 10 scenarios were created with missing percentages 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45% and 50%. For each scenario, the missing values were imputed using a set of imputation methods: Mean imputation, Last observation carried forward (LOCF), AutoARIMA (Auto Regressive Integrated Moving Average based imputation) and Kalman smoothing (Kalman smoothing using the state space representation of an ARIMA model for imputation) (Moritz & Bartz-Beielstein, 2017b). Mean and the LOCF were chosen as the baseline models. Moreover, AutoARIMA and Kalman smoothing were chosen as they are two well-established methods in the literature (Moritz et al., 2015; Wijesekara & Liyanage, 2020c). This process is repeated 30 times with different random seeds for the missing data generation to avoid any misleading performances due to chance. The means and the standard deviations of the 30 simulations were then investigated.

Even though it is logical to think that the missing patterns can be MCAR for most situations, it could be other mechanisms as well. Therefore, it is important to consider other situations as well. Air pollution data can be sometimes missing for a large period due to failures of the devices. To emulate different types of situations, another simulation was done by considering 10 scenarios where the gap size ranges from 10 to 100. Again, the procedure was repeated 30 times with different gap positions to obtain fair results.

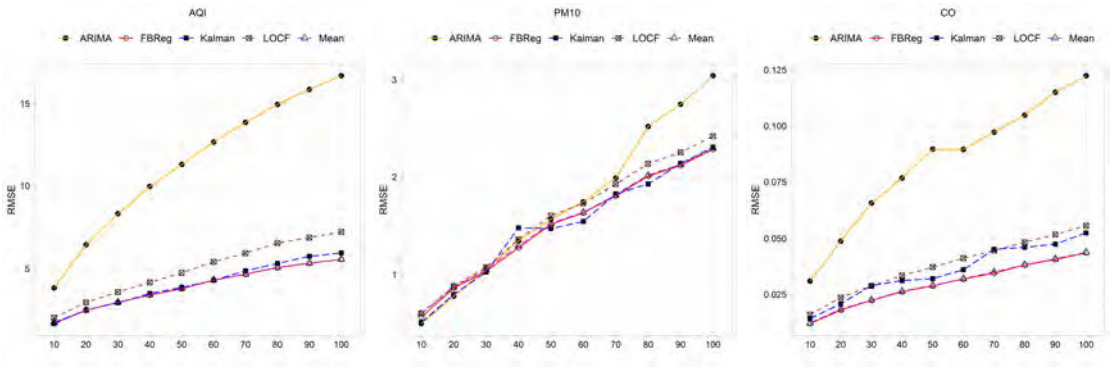
5.2.4 Results and Discussion

The plots in Figure 5.12 show the RMSE values for all the methods for different levels of missing percentages. The RMSE values were averaged for the 30 repeated simulations for each scenario. Plot (a) gives the values of the RMSE for the AQI dataset. As can be seen, FBReg (proposed method) appeared to be the best while ARIMA is the worst in dealing with AQI data regardless of the percentage of missing values. Mean imputation



(a) RMSEs for AQI imputations (b) RMSEs for PM10 imputations (c) RMSEs for CO imputations

FIGURE 5.12: Performance under MCAR



(a) RMSEs for AQI imputations (b) RMSEs for PM10 imputations (c) RMSEs for CO imputations

FIGURE 5.13: Performance for Large Gaps

also performs as closely as with FBReg. The next better model is the Kalman smoothing. Plot (b) gives the RMSE values for the PM10 dataset. Again, FBReg seemed to be the best while both Kalman smoothing and LOCF seemed to be the worst. ARIMA and Kalman smoothing also give approximately equal performance to FBReg. According to plot (c) for the CO dataset, FBReg persists in its performance over other selected methods. It appears that a similar trend follows in the performances for both PM10 and CO datasets. The performances of all the methods decline as the missing percentage increases.

Table 5.4 shows the average RMSE values along with their standard deviations under the MCAR situation. It can be seen that the standard deviations (Sd) FBReg method does not drastically change as the missing percentage increases. However, in ARIMA

Sds seems to be increasing as the percentage increases. Kalman smoothing and Mean imputation also appeared to maintain the deviation regardless of the level of missing values. LOCF also does not show drastic changes in the deviation however, its deviances are generally higher than that of the others. For the AQI dataset, FBReg gives the lowest deviation for almost all the situations.

Figure 5.13 depicts the performances of the methods for different gap sizes ranging from 10 to 100 consecutive NAs for the variables. It can be seen that the FBReg and Mean imputation perform equally well and they outperform the other methods for all the scenarios as well as for all the considered variables. Mean imputation is the next best model, while Kalman smoothing also performs reasonably well. AutoARIMA method performs poorly with large gaps.

Table 5.5 shows the average RMSE values along with their standard deviations under the situation of large consecutive gaps. It can be seen that the standard deviation of almost all the methods increases as the size of the gap increases.

5.2.5 Conclusion and Future work

Section 5.2 proposes a novel algorithm (FBReg) to impute univariate air pollution datasets using a bidirectional regularised regression model. The proposed method was evaluated against two base-line models namely Mean imputation and LOCF as well as two well-established imputation methods namely Kalman smoothing and AutoARIMA. The evaluation was done under the MCAR mechanism with exponentially distributed missing values and with varying consecutive gap sizes. FBreg outperforms all the other selected methods regardless of the percentage of missing values and the gap sizes in the series. Also, its standard deviations for the calculated RMSEs show a reasonable consistency compared to other methods. Among the considered methods the next best method is Kalman Smoothing. It is difficult to rank the other methods as their performances are not consistent for different levels of missing percentages and for different pollutants. Further analysis is to be carried out under different missing mechanisms (MAR and MNAR) to see if the proposed method displays consistent performance.

TABLE 5.4: Summary of the performance under MCAR

			Missing %									
			5%	**10%**	**15%**	**20%**	**25%**	**30%**	**35%**	**40%**	**45%**	**50%**
	AQI	Mean	3.06	4.18	5.37	5.99	6.75	7.51	8.07	8.72	9.05	9.39
		Sd	0.88	0.98	1.14	1.02	0.93	1.05	0.86	0.87	0.74	0.89
FBReg	PM10	Mean	1.19	1.84	2.23	2.62	2.85	3.17	3.42	3.67	3.85	4.18
		Sd	0.2	0.39	0.34	0.37	0.34	0.38	0.35	0.29	0.32	0.33
	CO	Mean	0.024	0.035	0.044	0.050	0.058	0.064	0.068	0.077	0.081	0.089
		Sd	0.004	0.004	0.004	0.004	0.004	0.004	0.005	0.006	0.007	0.008
	AQI	Mean	4.15	5.77	7.47	8.96	9.98	11	12	13.58	14.54	16.66
		Sd	0.91	1.10	1.60	1.18	1.59	1.57	1.91	2.01	1.59	5.65
ARIMA	PM10	Mean	1.12	1.79	2.25	2.68	2.89	3.34	3.29	3.80	4.01	4.74
		Sd	0.2	0.35	0.42	0.55	0.6	0.62	0.44	0.61	0.97	0.98
	CO	Mean	0.026	0.038	0.047	0.057	0.069	0.071	0.078	0.085	0.094	0.100
		Sd	0.005	0.007	0.009	0.012	0.014	0.011	0.015	0.01	0.01	0.01
	AQI	Mean	3.2	4.37	5.49	6.23	6.93	7.69	8.3	8.95	9.24	9.42
		Sd	0.90	0.99	1.19	1.06	1.02	1.05	0.97	0.86	0.89	0.79
Kalman	PM10	Mean	1.21	1.93	2.34	2.76	3.02	3.34	3.51	3.80	3.94	4.24
		Sd	0.19	0.42	0.33	0.36	0.33	0.37	0.31	0.3	0.28	0.32
	CO	Mean	0.023	0.034	0.043	0.051	0.059	0.067	0.074	0.080	0.083	0.090
		Sd	0.003	0.003	0.005	0.005	0.005	0.004	0.005	0.005	0.005	0.004
	AQI	Mean	3.6	4.91	6.14	6.8	7.61	8.38	8.94	9.59	9.89	10.15
		Sd	0.88	0.92	1.10	1.01	0.87	1.03	0.82	0.84	0.83	0.81
Mean	PM10	Mean	1.64	2.44	2.87	3.30	3.59	3.86	4.08	4.33	4.49	4.78
		Sd	0.27	0.45	0.35	0.38	0.32	0.36	0.34	0.31	0.27	0.32
	CO	Mean	0.035	0.050	0.060	0.068	0.077	0.082	0.087	0.093	0.097	0.103
		Sd	0.005	0.006	0.005	0.005	0.004	0.004	0.005	0.005	0.005	0.005
	AQI	Mean	3.71	5.12	6.55	7.47	8.36	9.64	10.27	10.94	11.24	11.78
		Sd	1.14	1.14	1.18	1.34	1.15	1.27	1.42	1.07	1.11	1.25
LOCF	PM10	Mean	1.48	2.25	2.77	3.14	3.63	4.01	4.15	4.54	4.73	5.08
		Sd	0.3	0.37	0.33	0.29	0.5	0.34	0.49	0.51	0.45	0.42
	CO	Mean	0.031	0.044	0.055	0.063	0.072	0.079	0.084	0.090	0.097	0.102
		Sd	0.004	0.005	0.006	0.007	0.005	0.004	0.006	0.007	0.005	0.006

TABLE 5.5: Summary of the performance for Large Gaps

			Gap Size									
			10	**20**	**30**	**40**	**50**	**60**	**70**	**80**	**90**	**100**
	AQI	Mean	1.67	2.46	2.94	3.38	3.77	4.29	4.65	5.05	5.32	5.56
		Sd	0.87	0.92	0.86	0.91	0.91	0.97	1.06	1.05	1.04	1.02
FBReg	PM10	Mean	0.56	0.86	1.03	1.27	1.52	1.64	1.81	2.01	2.12	2.28
		Sd	0.17	0.18	0.22	0.26	0.28	0.28	0.34	0.38	0.35	0.37
	CO	Mean	0.012	0.018	0.022	0.026	0.029	0.032	0.034	0.038	0.041	0.044
		Sd	0.006	0.009	0.011	0.012	0.013	0.014	0.015	0.017	0.018	0.020
	AQI	Mean	3.82	6.45	8.34	9.95	11.30	12.66	13.85	14.95	15.86	16.69
		Sd	1.12	1.12	1.07	1.17	1.13	1.21	1.35	1.26	1.23	1.12
ARIMA	PM10	Mean	0.49	0.78	1.06	1.35	1.57	1.75	1.99	2.52	2.75	3.04
		Sd	0.13	0.14	0.28	0.41	0.45	0.51	0.57	0.98	1.06	1.29
	CO	Mean	0.031	0.049	0.066	0.077	0.090	0.089	0.097	0.105	0.115	0.122
		Sd	0.010	0.019	0.019	0.020	0.012	0.013	0.016	0.021	0.019	0.023
	AQI	Mean	1.64	2.48	2.91	3.48	3.85	4.29	4.86	5.31	5.74	5.95
		Sd	0.93	1.10	1.05	1.02	0.97	1.04	1.38	1.33	1.55	1.30
Kalman	PM10	Mean	0.50	0.79	1.03	1.48	1.47	1.55	1.83	1.93	2.14	2.31
		Sd	0.23	0.31	0.24	0.58	0.43	0.33	0.45	0.44	0.46	0.55
	CO	Mean	0.014	0.021	0.029	0.031	0.032	0.036	0.045	0.046	0.047	0.052
		Sd	0.006	0.010	0.015	0.014	0.014	0.014	0.023	0.020	0.022	0.023
	AQI	Mean	1.72	2.47	2.96	3.40	3.79	4.31	4.67	5.07	5.33	5.56
		Sd	0.87	0.90	0.86	0.89	0.89	0.99	1.08	1.04	1.04	1.02
Mean	PM10	Mean	0.59	0.87	1.05	1.30	1.53	1.64	1.82	2.02	2.13	2.29
		Sd	0.17	0.18	0.21	0.25	0.28	0.28	0.34	0.37	0.35	0.36
	CO	Mean	0.012	0.018	0.023	0.026	0.029	0.032	0.035	0.038	0.041	0.044
		Sd	0.006	0.010	0.012	0.012	0.013	0.014	0.015	0.017	0.018	0.020
	AQI	Mean	2.01	2.94	3.57	4.15	4.74	5.41	5.93	6.55	6.89	7.23
		Sd	1.09	1.23	1.37	1.58	1.69	1.99	2.19	2.26	2.38	2.47
LOCF	PM10	Mean	0.60	0.88	1.07	1.36	1.61	1.73	1.92	2.14	2.25	2.42
		Sd	0.31	0.40	0.45	0.48	0.52	0.57	0.62	0.65	0.63	0.62
	CO	Mean	0.016	0.024	0.029	0.033	0.037	0.041	0.045	0.048	0.052	0.056
		Sd	0.009	0.012	0.015	0.015	0.016	0.018	0.019	0.020	0.022	0.024

5.3 Algorithm 2 : Imputing Large Gaps of High-resolution Environment Temperature

5.3.1 Introduction

Weather temperature data often consist of large missing gaps; for example, there may be missing values for a period of few days in an hourly temperature dataset. The existing methods are still not appealing in imputing large gaps. There are two main objectives of this section.

1. To examine the performance of a set of well-established methods in imputing large gaps in hourly temperature data.
2. To develop an improved methodology to impute large gaps.

5.3.2 Proposed Methodology

Since the focus is to impute missing values of temperature data, the proposed method is applicable for time series data which shows seasonal patterns. Temperature data usually exhibit seasonal patterns. Moreover, high-resolution temperature data may contain multiseasonal patterns as well. Therefore the proposed method takes into account multi-seasonality in the imputation process.

A schematic representation of the proposed algorithm is given in Figure 5.14. This algorithm produces a value for a missing observation based on a set of previous values (N steps). First, the time series with missing values is converted into a complete series using an approximation method. Linear approximation was used in this situation. This is done mainly due to two reasons. The first one is to identify the N steps using a partial autocorrelation plot. Typically this plot cannot be obtained with missing data. Secondly, to deseasonalise the series by estimating the trend component. Once the autocorrelation plot is created using the approximated complete dataset, the number of steps (N) is decided by examining the significantly correlated lags.

Then the series is decomposed and deseasonalised. Since the hourly temperature data typically contain multiseasonal patterns (annual and daily patterns), all the seasonal

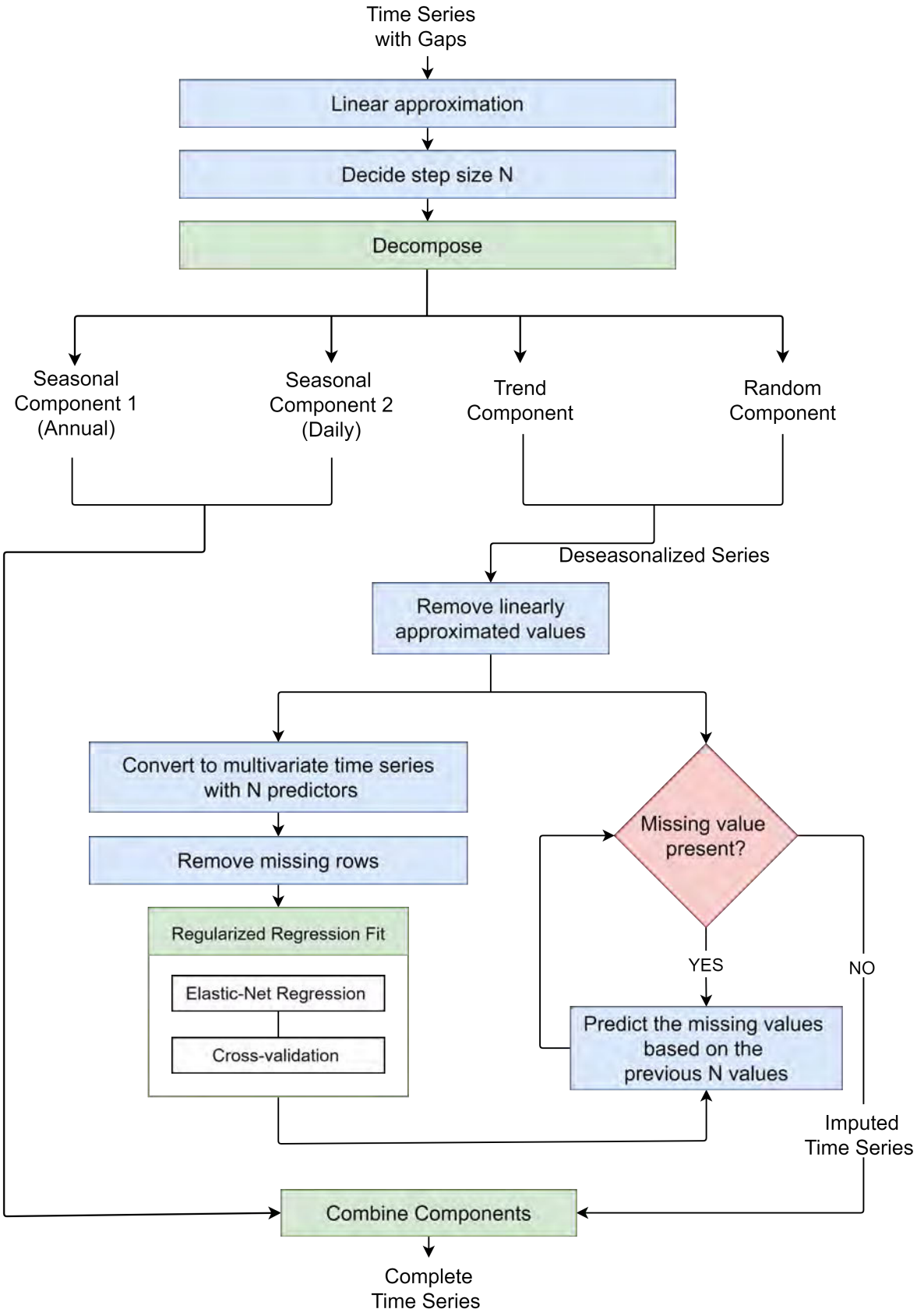


FIGURE 5.14: Proposed algorithm (DesReg)

components are removed and the deseasonalised component is used for the imputations. The procedure for the multi-seasonal decomposition in the Forecast R package was used here (Hyndman et al., 2020). In the deseasonalised series, the observations that correspond to the missing indices of the original series are then removed. This is done to remove the linearly approximated values in the beginning.

Let T, $t_1, t_2, ..., t_n$ be the deseasonalised series (with missing values). Then the series is converted into a multivariate dataset with N predictors (TX) and a response (TY).

$$TX = \begin{bmatrix} t_1 & t_2 & t_3 & \cdots & t_N \\ t_2 & t_3 & t_4 & \cdots & t_{N+1} \\ t_3 & t_4 & t_5 & \cdots & t_{N+2} \\ \vdots & \vdots & \vdots & & \vdots \\ t_{n-N} & t_{n-(N-1)} & t_{n-(N-2)} & \cdots & t_{n-1} \end{bmatrix}_{(n-N)\times N}$$

$$TY = \begin{bmatrix} t_{N+1} \\ t_{N+2} \\ t_{N+3} \\ \vdots \\ t_n \end{bmatrix}_{(n-N)\times 1}$$

After removing the rows with NAs in TX and appending a column with 1s at the beginning, the design matrix (X) can be created as below:

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{13} & \cdots & x_{1N} \\ 1 & x_{21} & x_{22} & x_{23} & \cdots & x_{2N} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{m1} & x_{m2} & x_{m3} & \cdots & x_{mN} \end{bmatrix}_{m\times(N+1),}$$

where m is the number of complete rows in TX. Then an elastic-net regression model is fitted considering the predictors as X and the dependent as Y (after removing corresponding missing rows in TY). The reason for applying regression models here is their low complexity compared to models like Neural Networks. Moreover, regularized methods such as elastic-net regression have been proven efficient in reducing the variance

(reducing overfitting) and avoiding issues of multicollinearity in the literature. The hyper-parameters in the regularized model have been obtained through 10-fold cross-validation.

Once the regularised regression model is fitted, the missing values in the original deseasonalised series are predicted using this model considering the previous N consecutive observations as predictors. Finally, the imputed deseasonalised series is combined with the seasonal components and form a complete time series.

Dataset

Hourly weather temperature data from 2016-01-01:01 to 2018-12-31:24 was selected for the study. Data is available on the website of the New South Wales government Planning, Industry and Environment department (https://www.dpie.nsw.gov.au/air-quality/air-quality-data-services). The missing values (0.14%) of this series were imputed using Kalman smoothing procedure in the imputeTS R package and treated as the ground truth/reference series. The reason for doing this is finding a reference series without any missing values is almost impossible for high-resolution temperature data. Further, this series does not contain large gaps. Any consecutive 24 missing value gap is considered a large gap throughout this study as the focus is to impute hourly temperature data.

Simulation Procedure

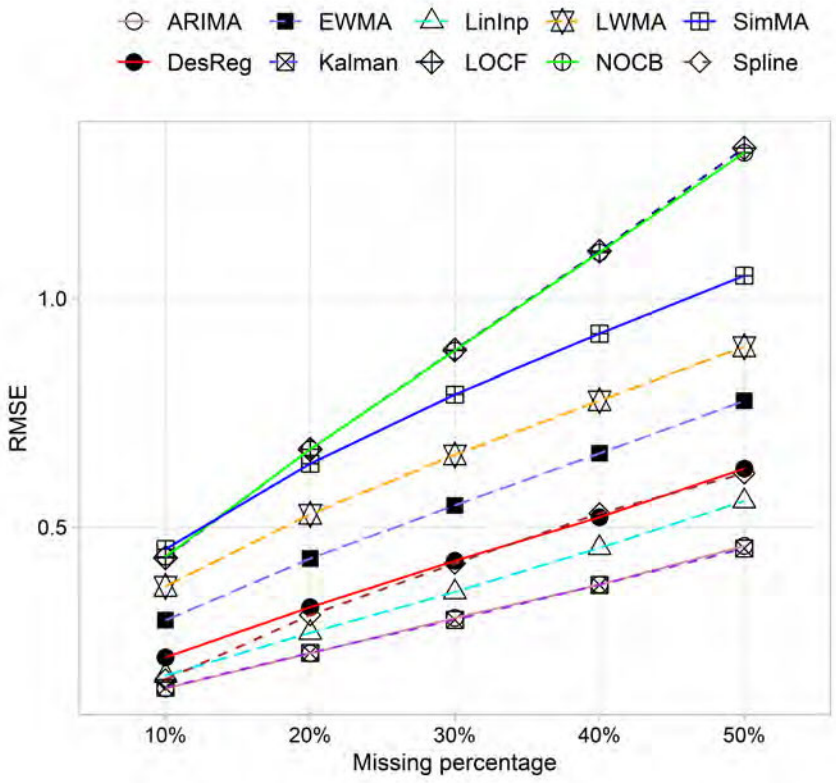
To emulate the MCAR mechanism, a simulation procedure suggested by Moritz et al. (Moritz et al., 2015) for their comparison of imputation methods in univariate time series was applied. Missing data were generated by considering the exponential distribution. By using values ranging from 0.1 to 0.5 for the rate parameter of the exponential distribution, 5 scenarios were created with missing percentages 10%, 20%, 30%,40% and 50%. For each scenario, the missing values were imputed using a set of baseline and well-established imputation methods: AutoARIMA (Auto Regressive Integrated Moving Average based imputation), Kalman smoothing (Kalman smoothing

using the state space representation of an ARIMA model for imputation), Mean imputation, Last observation carried forward (LOCF), Next observation carried backward (NOCB), Linear interpolation (Lininp), Spline, Simple moving average (SimMA), Linearly weighted moving averages (LWMA), Exponentially weighted moving averages (EWMA) (Moritz & Bartz-Beielstein, 2017b) and the proposed method (DesReg). This process is repeated 30 times with different random seeds for the missing data generation to avoid any misleading performances due to chance. The means and the standard deviations of the root mean squared errors (RMSE) of the 30 simulations were then investigated.

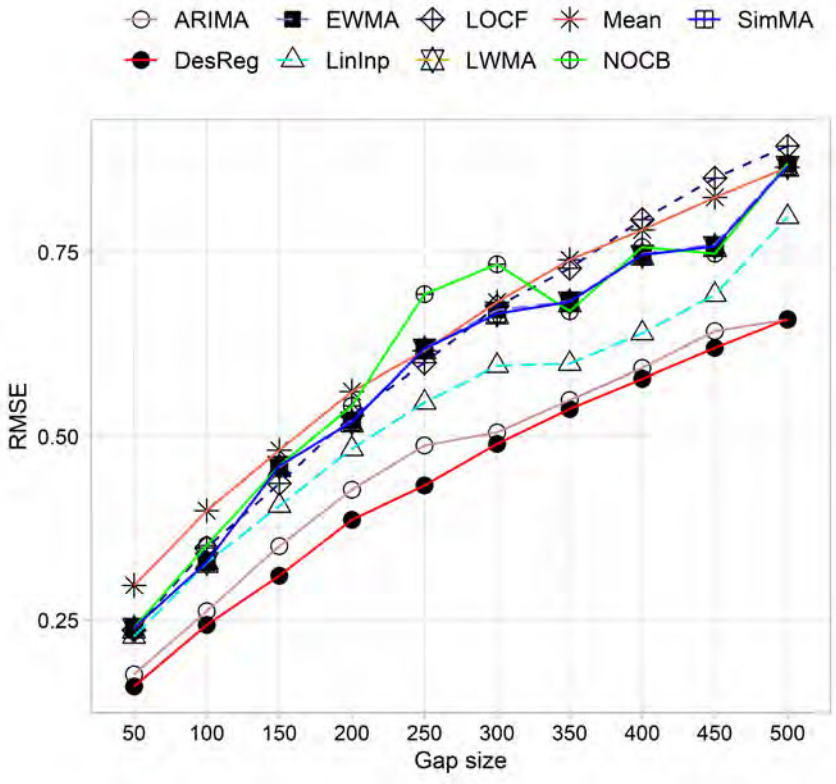
Since the main focus is to impute large gaps, another simulation was done by considering 10 scenarios where the gap size ranges from 50 to 500. Again, the procedure was repeated with different gap positions, and the means and the standard deviations of RMSE were calculated.

5.3.3 Results and Discussion

Figure 5.15a shows the RMSE values for all the methods for different levels of missing percentages under MCAR with exponentially distributed missing values. The RMSE values were averaged for the 30 repeated simulations for each scenario. It appears that ARIMA and Kalman methods perform equally well in all the levels of missing percentages and they are the best methods in these situations. Linear interpolation is the next best method while Spline and the proposed method (DesReg) are the next two good models. The performance of mean imputation was significantly poor than all the other methods and it is not included in the Figure 5.15a. Figure 5.15b exhibits the performances of methods under different situations with large gaps. It can be seen that DesReg outperform all the other methods in all the situation with large gaps. This indicates that the DesResis is more suitable for situations with large gaps than other considered MCAR situations. ARIMA method can be considered as the next two best methods while linear interpolation also performs reasonably well than all the other methods. The mean and standard deviations of the RMSE values are summarised in Table 5.6. Spline is the least performed method while Kalman method is also significantly underperformed in this situation. These two methods are not included in



(a) RMSE under MCAR (Exponential)



(b) RMSE for large gaps

FIGURE 5.15: Performance results

Figure 5.15b. It is evident that the standard deviations of the DesReg and ARIMA are comparatively lower than that of all the other methods. The standard deviation of ARIMA method tends to increase as the gap size increases while the standard deviation of DesReg does not drastically change as the gap size changes. This gives further evidence that the DesReg is a robust method, particularly in situations with large gaps.

TABLE 5.6: Summary of the performance in large gaps

		Gap size									
		50	**100**	**150**	**200**	**250**	**300**	**350**	**400**	**450**	**500**
DesReg	Mean	0.16	0.24	0.31	0.39	0.43	0.49	0.54	0.58	0.62	0.66
	Sd	0.096	0.090	0.069	0.094	0.078	0.083	0.072	0.082	0.095	0.096
ARIMA	Mean	0.18	0.26	0.35	0.43	0.49	0.51	0.55	0.59	0.64	0.66
	Sd	0.076	0.078	0.087	0.134	0.140	0.103	0.110	0.140	0.141	0.131
Kalman	Mean	0.32	1.25	2.27	2.54	4.45	6.09	5.82	8.39	7.08	7.90
	Sd	0.256	1.008	2.230	1.824	2.926	3.539	4.320	6.363	6.119	5.820
Mean	Mean	0.30	0.40	0.48	0.56	0.62	0.68	0.74	0.78	0.82	0.87
	Sd	0.127	0.131	0.127	0.138	0.142	0.156	0.168	0.174	0.176	0.182
LOCF	Mean	0.24	0.35	0.44	0.53	0.60	0.68	0.73	0.79	0.85	0.89
	Sd	0.096	0.119	0.138	0.164	0.179	0.200	0.221	0.260	0.284	0.295
NOCB	Mean	0.24	0.35	0.46	0.54	0.69	0.73	0.67	0.76	0.75	0.87
	Sd	0.126	0.168	0.144	0.191	0.249	0.524	0.184	0.306	0.139	0.247
LinInp	Mean	0.23	0.33	0.41	0.48	0.55	0.60	0.60	0.64	0.69	0.80
	Sd	0.093	0.094	0.048	0.132	0.089	0.245	0.074	0.090	0.117	0.187
Spline	Mean	0.33	1.43	2.58	3.04	4.73	6.92	6.41	9.49	8.46	8.71
	Sd	0.278	1.227	2.820	1.991	3.511	4.127	4.215	6.760	6.614	6.419
SimMA	Mean	0.24	0.33	0.46	0.52	0.62	0.67	0.68	0.75	0.76	0.87
	Sd	0.101	0.094	0.065	0.143	0.107	0.331	0.126	0.187	0.127	0.203
LWMA	Mean	0.24	0.33	0.46	0.52	0.62	0.67	0.68	0.75	0.76	0.87
	Sd	0.100	0.093	0.065	0.143	0.107	0.331	0.126	0.187	0.127	0.203
EWMA	Mean	0.24	0.33	0.46	0.52	0.62	0.67	0.68	0.75	0.76	0.87
	Sd	0.101	0.093	0.064	0.140	0.102	0.342	0.132	0.195	0.132	0.202

Figure **??** shows the imputations of DesReg, ARIMA, Kalman and Linear for a situation where the gap size is 500. It is clear that the DesReg has produced plausible results than the other methods. Table 5.7 gives the RMSE values for each of the methods. Moreover, it gives the Mean Absolute Percentage Error(MAPE), Euclidean Distance(DT) and Dynamic Time Warping (DTW) distances between the actual and imputed segment of the series. DesReg outperforms others Yeilding lowest errors and distances.

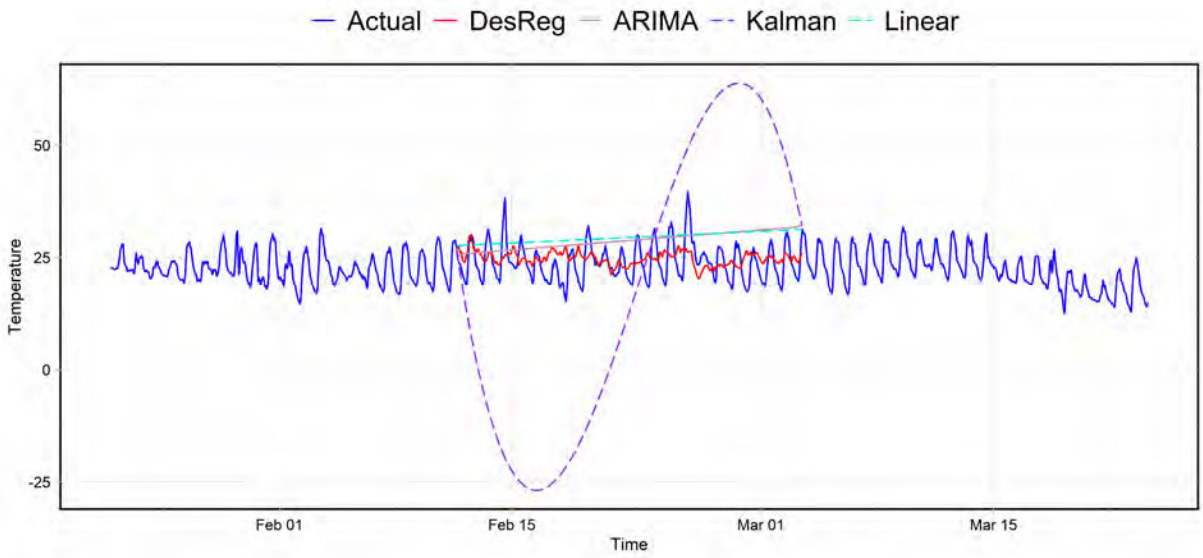


FIGURE 5.16: Imputations for a large gap

TABLE 5.7: Error measures and distances

	RMSE	MAPE	ED	DTW
DesReg	**0.593**	**0.296**	**96.24**	**1370.90**
ARIMA	0.912	0.499	148.04	2019.28
Kalman	4.587	2.453	743.98	11884.31
Linear	0.959	0.536	155.58	2524.874

5.3.4 Conclusion

Section 5.3 compares the performance of a set of univariate time series imputation methods in dealing with missing values of hourly temperature time series. It is evident that the AutoARIMA and Kalman methods are the best methods to impute missing values under the MCAR mechanism with exponential distribution. However, their performances are not consistent in other missing situations. Especially, the Kalman method performs poorly when there are large gaps. DesReg is the best method to deal with large missing gaps. AutoARIMA and Linear interpolation can be ranked as the next two best methods to deal with large gaps respectively. DesReg could be recommended to impute large gaps in univariate time series which shows seasonal patterns.

5.4 Contribution

This chapter first, discussed six well-established methods of dealing with missing values in a univariate time series context and compare their performance on imputing missing values for air quality data in the Sydney region. The methods discussed here are Mean Imputation, Spline Interpolation, Simple Moving Average, Exponentially Weighted Moving Average, Kalman Smoothing on Structural Time Series Models and Kalman Smoothing on ARIMA models. The performances of these methods were compared with three performance measures; Mean Squared Error (MSE), Coefficient of Determination (R2) and Index of Agreement (d). Among the six methods considered, Kalman Smoothing Method on Structural Time Series is the best method for imputing missing values in the context of air quality data where the missing mechanism is MCAR.

Then it proposed a novel algorithm (FBReg) to impute univariate air pollution datasets using a bidirectional regularised regression model. The proposed method was evaluated against two baseline models namely Mean imputation and LOCF as well as two well-established imputation methods namely Kalman smoothing and AutoARIMA. The

evaluation was done under the MCAR mechanism with exponentially distributed missing values and with varying consecutive gap sizes. FBreg outperforms all the other selected methods regardless of the percentage of missing values and the gap sizes in the series.

Next, it proposed another algorithm (DesReg) extending the FBReg algorithm to seasonal variables to deal with missing values. It was evaluated using hourly temperature data. AutoARIMA and Kalman methods were the best methods to impute missing values under the MCAR mechanism with exponential distribution. However, their performances are not consistent in other missing situations. Especially, the Kalman method performs poorly when there are large gaps. DesReg was the best method to deal with large missing gaps. DesReg could be recommended to impute large gaps in univariate time series which show seasonal patterns.

Chapter 6

Missing Value Imputation: Multivariate Approach

This chapter proposes an algorithm by incorporating both the time series characteristics of the series itself and the information available on other highly correlated variables to impute large intervals of missing data. This chapter is based on the following publication.

- Wijesekara, L., and Liyanage, L. (2023). Mind the Large Gap: Novel Algorithm Using Seasonal Decomposition and Elastic Net Regression to Impute Large Intervals of Missing Data in Air Quality Data. Atmosphere, 14(2), 355. https://doi.org/10.3390/atmos14020355

6.1 Introduction

Data pre-processing is a critical part of any data mining project. It includes but is not limited to, identifying and removing noise and outliers, handling data inconsistencies and dealing with missing values. The latter plays a significant role in some studies, and it primarily affects the model performance. Unless carefully dealt with, it may introduce a substantial bias for the analysis. Inaccurate data may lead to an inaccurate division of training, testing, and validating datasets and produce misleading conclusions (Chandrasekaran et al., 2016). Therefore, taking as many accurate measures as possible to deal with missing values is pivotal. The most desired thing is to have

a complete dataset for the required analysis. However, in reality, this is often not the case.

In the context of meteorological variables such as air pollutants, temperature, ozone, etc., it is very common to expect missing data. Measurements of these variables are mostly recorded using sensors. However, the data collected through sensors often exhibit missing data due to failures of the devices and human errors in recording. Most of the data processing or analysis techniques rely on complete datasets. For example, in time series, some of the methods for decomposing a time series would not work with missing data. Therefore, it is inevitable to deal with missing data. Some methods to approximate missing values highly rely on the nature of the data. However, there is no universally accepted approach.

Even though there are well-established techniques to recover missing data in the time series context, recovering a large gap in a time series is still challenging. Large gaps are ubiquitous in real-world scenarios due to data transmission failures (Chandrasekaran et al., 2016). This is the same for meteorological data as well. For example, in a daily temperature measurement series, there may be missing data for several weeks or months due to a failure of a sensor. Most of the existing methods fail to recover large gaps. In this chapter, a novel algorithm based on seasonal decomposition and elastic net regression is proposed to recover large gaps of time series data when there exist correlated variables.

6.2 Methods

6.2.1 Seasonal-Trend Decomposition procedure based on Loess(STL)

STL is a filtering procedure for decomposing a time series (Y_t) into three additive components of trend (Y^T), seasonality (Y^S), and remainder/random component (Y^R) (Cleveland et al., 1990).

$$Y_t = Y^T + Y^S + Y^R \tag{6.1}$$

Figure 6.1 shows an example of a decomposed series using STL. The data in the first(top) panel represents daily average measurements of PM2.5 measured at the Richmond site in Sydney from 2000 to 2020. The bottom panel shows the trend component, while the third panel shows the seasonal component: variation in the data at or near the seasonal frequency, which in this case is one cycle per year. The second panel represents the remainder which is the remaining variation in the data beyond that in seasonal and trend components.

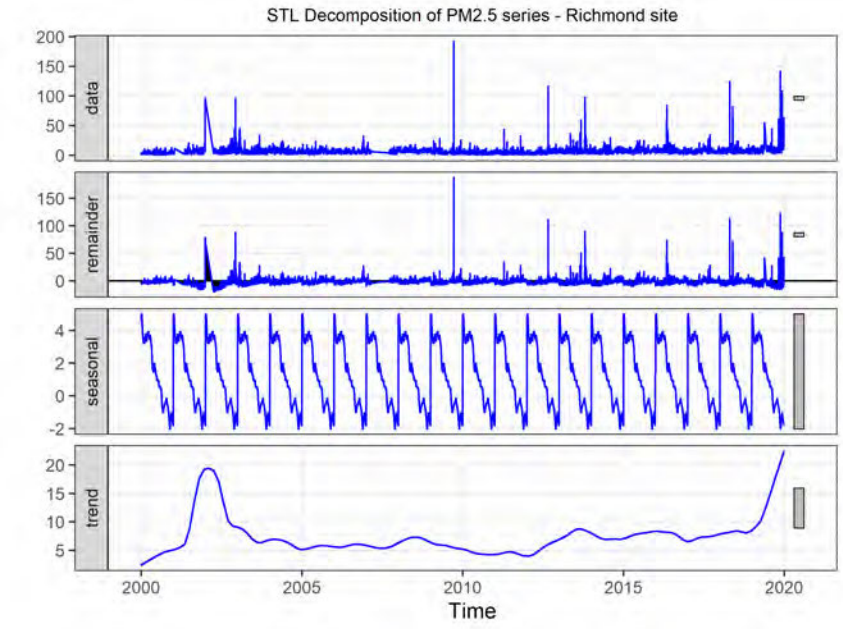


FIGURE 6.1: Decomposition

STL uses applications of the Loess smoother for the decomposition. Loess smoother applies locally weighted polynomial regressions at each point in the dataset using the values closest to the point whose response is being estimated. Description of the procedure is described by Cleveland et al. in their paper (Cleveland et al., 1990). STL procedure for decomposition is selected for this algorithm due to its applicability for different types of time series, fast computability and strong resilience to outliers.

6.2.2 Elastic-Net Regression

Let y be the dependant, $x_j s$ be the predictors, and $\beta_j s$ be the regression coefficients of a linear regression model with p predictors. The least squares fitting procedure

estimates $\beta_j s$ by minimising the residual sum of squares (RSS).

$$\hat{y} = \hat{\beta_o} + \sum_{j=1}^{p} \hat{\beta_j} x_{ij} \tag{6.2}$$

$$RSS = \sum_{i=1}^{n} \left(y_i - \beta_o - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \tag{6.3}$$

$$RSS + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=1}^{p} \beta_j^2, \tag{6.4}$$

where λ_1 and λ_2 are hyper-parameters. When $\lambda_1 = 0$, it becomes ridge regression and when $\lambda_2 = 0$, it becomes Lasso regression. Elastic-Net Regression estimates the coefficients by minimising the equation 6.4. The predictor variables are highly correlated with each other in this situation, and therefore the elastic-net regression gives the best model for this situation by overcoming the multicollinearity issues.

6.2.3 Artificially generating Missing Data

Performance evaluation of a missing value imputation technique can only be done for simulated missing data. Little and Rubin (2002) present the data structure for simulating missing values in a univariate dataset, providing the MCAR mechanism.

Let $Y = (y_1, ..., y_n)^T$ where y_i denotes the value of a random variable for observation i, and let $M = (M_1, ..., M_n)^T$ where $M_i = 0$ for units that are observed and $M_i = 1$ for units that are missing. Suppose the joint distribution of $f(y_i, M_i)$ is independent across units $i = 1, 2, ..., n$ so that the probability a value is observed does not depend on the values of Y or M for the other units. Then,

$$f(Y, M|\theta, \phi) = f(Y|\theta) f(M|Y, \phi) = \prod_{i=0}^{n} f(y_i|\theta) \prod_{i=0}^{n} f(M_i|y_i, \phi). \tag{6.5}$$

(Rantou, 2017)

Artificial missing data were simulated under the MCAR mechanism using exponential distribution to evaluate the performance of the proposed algorithm. The exponential distribution with rate λ has the density

$$f(x) = \lambda e^{-\lambda x} \; for \; x \geq 0. \tag{6.6}$$

The simulation procedure is further described in section 6.4.1.

6.2.4 Performance Measure

The Root Mean Squared Error (RMSE) is used as a measure to compare the performance of the proposed algorithm with other existing methods. RMSE between imputed value (y^{imp}) and the respective observed value (y^{obs}) of a time series ($y_1, y_2, ..., y_n$) is given by,

$$RMSE(y^{imp}, y^{obs}) = \sqrt{\frac{\sum_{t=1}^{n} (y_t^{imp} - y_t^{obs})^2}{n}}. \tag{6.7}$$

6.3 Proposed Algorithm

6.3.1 Characteristics of the Data

The proposed algorithm can be applied to a dataset with some specific characteristics. Mainly, there should be one or more correlated time series variables with the time series to be imputed, which has large gaps. This is ubiquitous, especially in meteorological datasets. Figure 6.2 shows a PCA (Principal Component Analysis) plot created after conducting PCA for a meteorological dataset which includes hourly data recorded at the Liverpool monitoring site in the Sydney region, Australia.

The variables are shown in arrows while the contribution is represented by colours. The angle between the two arrows reflects the correlation between those two variables. It can be seen that the following variable sets are highly correlated.

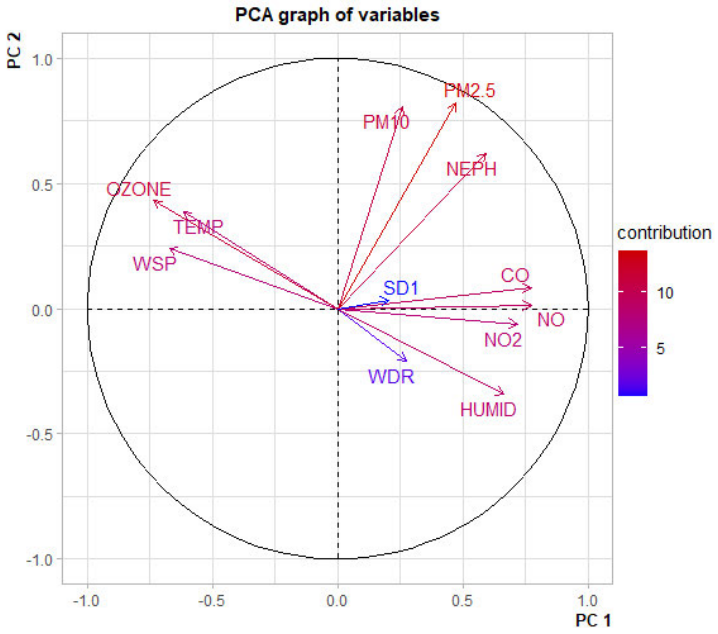


FIGURE 6.2: Biplot of PCA

- CO (Carbon monoxide in parts per million), NO (Nitrogen monoxide in parts per hundred million) and NO2 (Nitrogen dioxide in parts per hundred million) are positively-correlated
- O3 (Ozone in parts per hundred million) and Temperature (in Celsius degrees) are positively correlated
- PM2.5 (Particulate matter less than 2.5 micrograms per cubic meter), PM10 (Particulate matter less than 10 micrograms per cubic meter) and NEPH (a measure of visibility) are positively correlated
- Humidity is negatively correlated with O3 and Temperature

Similar results can be seen in other monitoring sites in the Sydney region. Therefore this algorithm is particularly applicable to meteorological datasets like this.

In this setting, we denote the time series variable to be imputed (with large gaps) as the dependent variable and the other correlated variables as predictors. The main purpose of this algorithm is to recover large missing gaps. When there is a time series with

TABLE 6.1: Missing Statistics

Site	Variable								
	PM2.5			**PM10**			**NEPH**		
	Longest Gap	Frequent Gap	Missing Percent	Longest Gap	Frequent Gap	Missing Percent	Longest Gap	Frequent Gap	Missing Percent
RICHMOND	197	1 NA 179 times	**12.4**	18	1 NA 63 times	**3.85**	5	2 NA 28 times	**2.07**
BRINGELLY	6025	1 NA 11 times	**83.1**	24	2 NA 15 times	**3.48**	14	2 NA 31 times	**3.01**
EARLWOOD	83	1 NA 48 times	**5**	15	1 NA 41 times	**3.29**	5	1 NA 14 times	**1.77**

large gaps (dependant), and it has one or more other correlated variables (predictors). Ideally, the predictors should be complete (no missing values) to apply this algorithm. This is highly unlikely in real datasets, and there may be missing values in almost all the predictors. However, since there are well-established imputation methods are available for univariate time series with a percentage of missing values as low as 5%, those methods could be used to recover small missing gaps in the predictors.

Table 6.1 shows statistics of missing values of daily PM2.5, PM10 and NEPH observations at three selected monitoring sites in the Sydney region from 2000 to 2020.

In PM10 and NEPH, the percentage of missing values is lower than 5%, whereas, in PM2.5, the percentage of missing values is relatively very high. Moreover, the longest gap size is also relatively larger for PM2.5 than that of the other two variables. Therefore, well-established univariate imputation methods could be used to recover missing values in PM10 and NEPH, which could then be used as predictors to recover significant PM2.5 (dependent) gaps. To summarise, the time series with large gaps (dependent) should have one or more correlated variables (predictors), and those variables should be complete or at least be able to recover from existing methods.

6.3.2 Algorithm

Algorithm 3 represents the main imputation algorithm proposed in this chapter.

The input for algorithm 3 is a data frame with a time index. The variable to be imputed, which has large intervals of missing values, is considered as the dependent variable (Y), and the set of correlated variables are considered as the predictors ($X_1, ..., X_p$). First,

Algorithm 3 Imputation Algorithm

Input: Data frame with time index t, Dependant Y, Predictors $X_1, ..., X_p$

$I_{Missing}$:← *indices of missing observations of* Y
$I_{Observed}$:← *indices of available observations of* Y
Separate the variables $(Y, X_1, ..., X_p)$ into time series objects
Decompose each of the time series objects into Seasonal, Trend and Random components
Y^S :← *Seosonal Component of* Y
Y^T :← *Trend Component of* Y
Y^R :← *Random Component of* Y
X^S :← *Seosonal Component of Predictors*$(X_1^S, ..., X_p^S)$
X^T :← *Trend Component of Predictors*$(X_1^T, ..., X_p^T)$
X^R :← *Random Component of Predictors*$(X_1^R, ..., X_p^R)$
Predict the missing values of Y^T using X^T (Algorithm 4)
Predict the missing values of Y^R using X^R (Algorithm 4)
$Y_{Imputed} \leftarrow Y^S + Y^T_{completed} + Y^R_{completed}$

Output: Complete data frame $(Y_{Imputed}, X_1, ..., X_p)$

Algorithm 4 Imputing Trend/Random component

Input: Data frame with time index t, Dependant Y, Predictors $X_1, ..., X_p$

$D_M \leftarrow$ *Data frame corresponding to* $I_{Missing}$
$D_{NM} \leftarrow$ *Data frame corresponding to* $I_{Observed}$
procedure PredictiveModel(D_{NM})
    TrainData ← *random split of* 70% *data*
    TestData ← *rest of the* 30% *data*
    for each model 1 to N **do**
        Fit the model using TrainData
        Predict the output for TestData
        Calculate RMSE for TestData
    BestModel ← *model with least RMSE*
    return *BestModel*
Predict missing Y values in D_M using *BestModel*

Output: Complete Y series ($Y^T_{completed}$ *or* $Y^R_{completed}$)

identify the indices of missing observations of Y. Then, all the variables are converted into separate time series objects which are then decomposed into a seasonal, a trend and a random component. Seasonal-Trend decomposition procedure based on Loess (STL) (Cleveland et al., 1990) can be used for the decomposition of the time series. As it requires a complete series to do the STL decomposition, a simple approximation such as linear approximation can be used to approximate missing values of the Y series. The purpose of this approximation is only to decompose the series and to identify the seasonal component. These approximated values are replaced later with imputed values. Chandrasekaran et al. have also used this procedure to decompose a series with missing values in their Seasonal Moving Window Algorithm (SMWA) for imputing missing values of univariate time series (Chandrasekaran et al., 2016). Next, trend and random components are fed into algorithm 2 while leaving out the seasonal component.

In algorithm 4, either trend or random components of the dependant(Y) and predictors ($X_1, ..., X_p$) with time index are inputted as a data frame. Then the data frame is divided into two parts based on the time index; one part with complete Y observations and the other part with missing Y observations. The complete dataset is further divided into train and test datasets. The training dataset is used to build models to predict Y based on predictors $X_1, ..., X_p$ and the test dataset is used to evaluate the models using RMSE. The number of candidate models (N) can be decided according to the nature of predictors and the number of predictors (p). However, since the predictors are correlated, the elastic net regression model is recommended to avoid multicollinearity issues. Figure 6.3 shows a schematic representation of the proposed algorithm.

6.4 Simulations and Performance Evaluation

In this section, the proposed algorithm is demonstrated using a dataset, and its performance is compared against several well-established imputation methods. To evaluate the performance, missing data are artificially created using a simulation procedure. R programming language (R Core Team, 2013) is used to implement the algorithm.

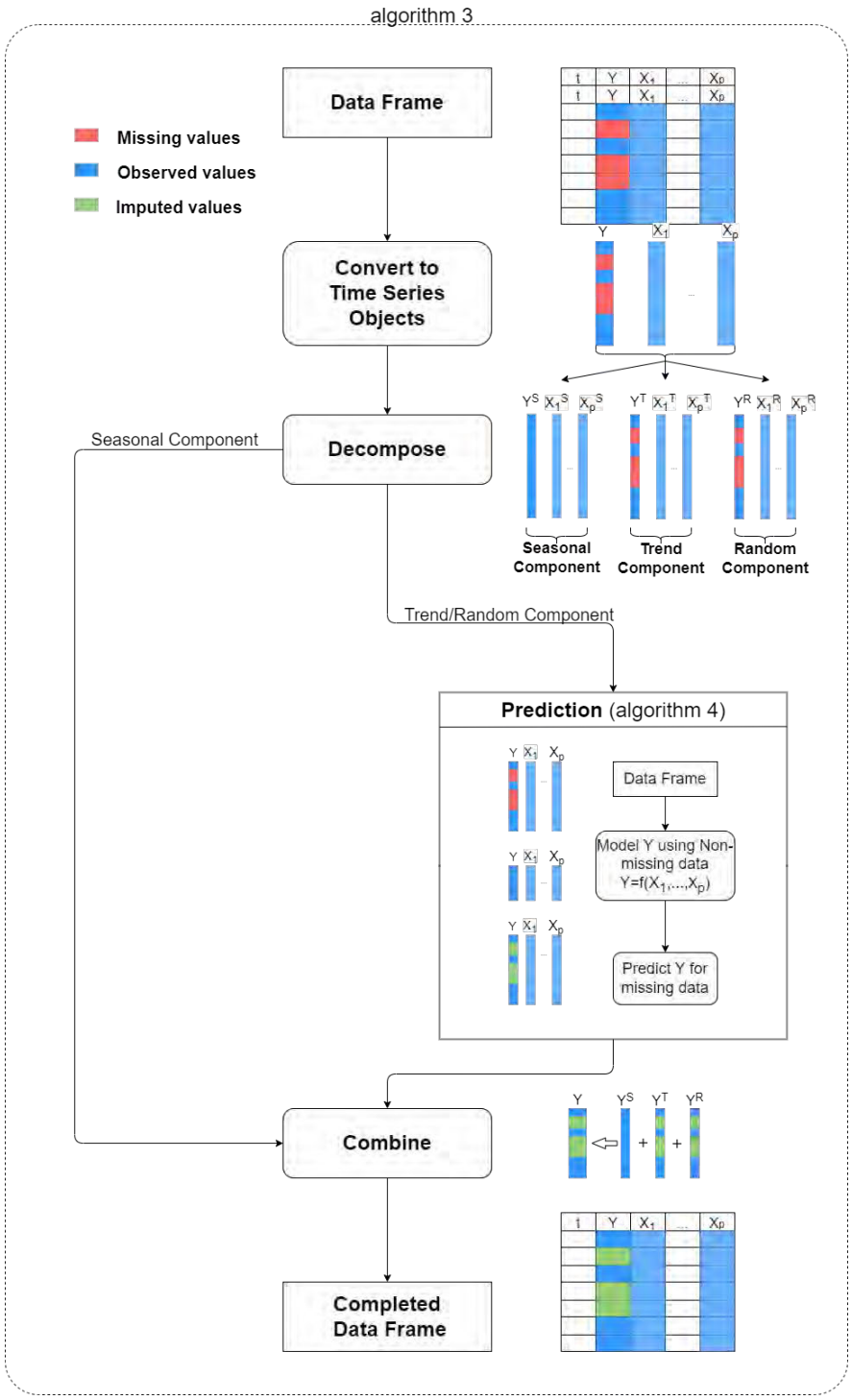


FIGURE 6.3: Schematic representation of the proposed algorithm

6.4.1 Simulation Procedure

It is necessary to have a complete dataset to evaluate the proposed algorithm. However, it is almost impossible to have a complete dataset of this nature. It is important to use a real dataset with similar nature to demonstrate the algorithm. After examining daily PM2.5, PM 10 and NEPH values for a 20-year period from 2000 to 2020 at 18 selected sites in the Sydney region, a dataset from the 2014-2020 period at the Earlwood site was selected where the percentages of missing values were equal or below 5%. Kalman smoothing on structural time series models in *ImpueTS* R package (Moritz & Bartz-Beielstein, 2017b) is used to recover the missing values, and then the dataset is considered as the ground truth. PM2.5 series is selected as the Y series, and PM10 and NEPH are used as predictors. The reason to select PM2.5 as Y is that it consists of large gaps in real data, as investigated in section 6.3.1. It is required to artificially create missing values to compare the performance of the proposed algorithm. The simulations are done under the MCAR mechanism using the exponential distribution. Moritz et al.(2015) also used this simulation procedure for their comparison of imputation methods in univariate time series. They have considered the exponential distribution to create missing values as it is applicable in modelling many real-life situations, such as the time gap between two incoming phone calls etc. (Moritz et al., 2015). Five scenarios were created using five different values for the rate parameter of the exponential distribution.

- Scenario1: Missing percentage 10% with rate 0.1
- Scenario2: Missing percentage 20% with rate 0.2
- Scenario3: Missing percentage 30% with rate 0.3
- Scenario4: Missing percentage 40% with rate 0.4
- Scenario5: Missing percentage 50% with rate 0.5

Seasonal-Trend decomposition procedure based on Loess (STL) (Cleveland et al., 1990) is used for the decomposition of the time series with a linear approximation as suggested in section 6.3. The following six models are used in algorithm 2 to predict the

trend/random component of the dependant using predictors. The purpose of using these six models is to select the best predictive model which is appropriate for the situation. However, using only the last three models or even using only the sixth model can be recommended.

1. Simple linear regression model with only NEPH as the predictor
2. Multiple linear regression model with PM10 and NEPH as predictors
3. Multiple linear regression model with PM10, NEPH and the interaction term as predictors
4. Ridge regression (best parameter was chosen using 10-fold cross-validation)
5. Lasso regression (best parameter was chosen using 10-fold cross-validation)
6. Elastic-Net Regression (best parameters were chosen using 10-fold cross-validation)

6.4.2 Performance Evaluation

The proposed algorithm was used to impute the missing values for the five different scenarios mentioned above, and the RMSE was calculated. In order to compare the performance with other existing methods, four well-established methods, namely Kalman smoothing on ARIMA models, Kalman smoothing on structural time series models, linear interpolation and mean imputation, were used. Missing values were imputed for the five scenarios using the considered methods. The RMSE values for all the methods in five scenarios are given in Figure 6.4.

The proposed algorithm has yielded the lowest RMSE for all five scenarios, whereas the mean imputation has yielded the largest RMSE values. Linear interpolation and the Kalman smoothing methods perform equally. It is clear that the proposed algorithm outperforms all the other considered methods. The reason is that the proposed algorithm incorporates the information of other correlated variables for the imputation, while all the other methods consider only the time series characteristics.

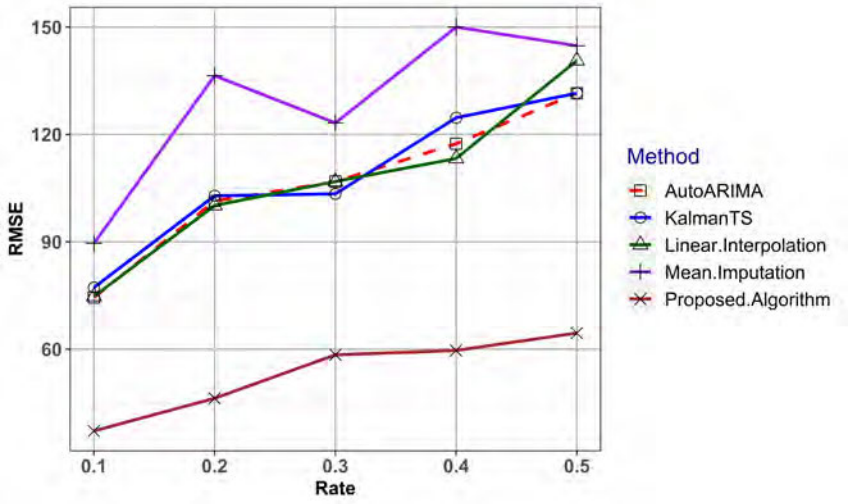
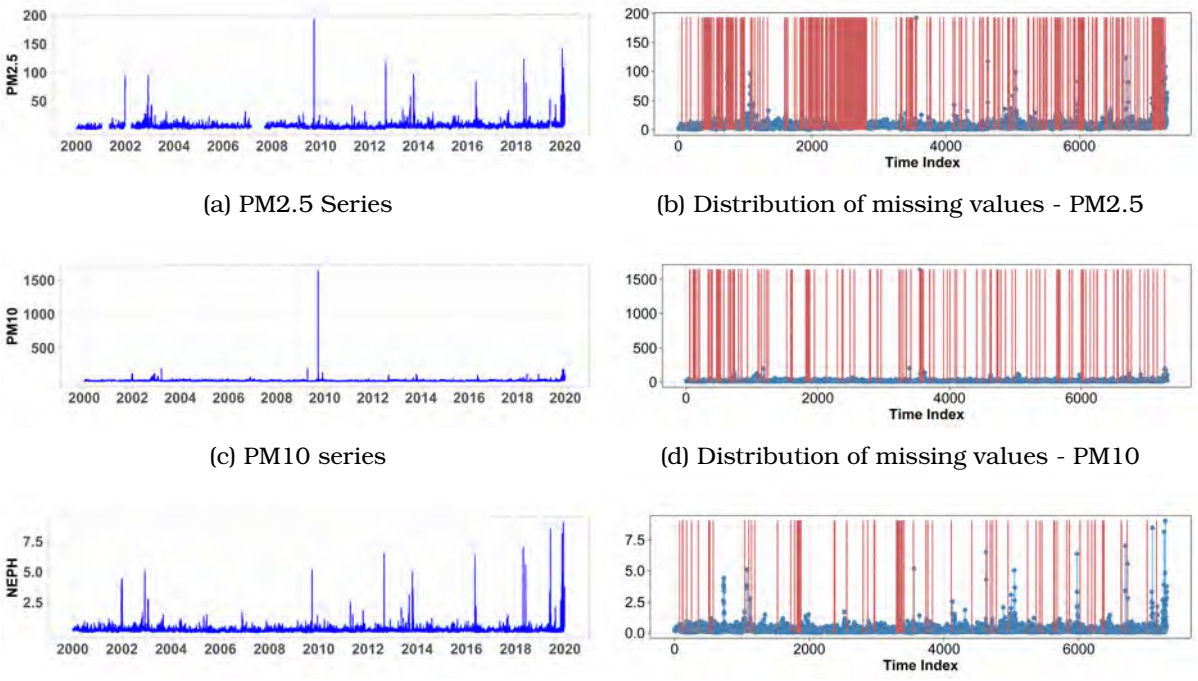


FIGURE 6.4: Performance Comparison

6.5 Real application

In this section, the applicability of the algorithm is further demonstrated using a real dataset with large gaps. In the dataset at the Richmond site, the percentage of missing values in PM2.5 is relatively large (12%), and the longest gap size is 197 days, as shown in Table 6.1. In contrast, PM10 and NEPH have relatively low missing percentages and small gaps. The left-hand side three plots in Figure 6.5 (a,c and e) show the PM2.5, PM10 and NEPH series from 01-01-2000 to 31-12-2019, and the right-hand side three plots (b, d and f) shows the positions of the missing values in each series, highlighting missing regions by red vertical lines. As can be seen, the PM2.5 consists of large gaps, while the other two series consist of small gaps.

The proposed algorithm was applied to recover the missing values in PM.2.5 series. There is no way to compare the performance in this particular situation, as the missing values are actually unknown. However, for a visual comparison, three other methods, namely linear interpolation, Kalman smoothing on structural models and Kalman smoothing on ARIMA models, were used to recover the same series, and the results were examined. Figure 6.6 shows the original PM2.5 series with missing values, imputation results of the three other methods and imputation results of the proposed algorithm. It can be seen that the proposed method has produced smoother values than the other methods.



(a) PM2.5 Series

(b) Distribution of missing values - PM2.5

(c) PM10 series

(d) Distribution of missing values - PM10

(e) NEPH series

(f) Distribution of missing values - NEPH

FIGURE 6.5: Distribution of missing values

6.6 Conclusion

Missing values are ubiquitous in most real-world scenarios, especially in meteorological datasets. Recovery of missing values is an essential part of most of the analyses. Even though there exists a plethora of methods, the performances of those methods depend upon the nature of the data and the missing mechanism. There is no universal approach to recovering missing values. When the percentage of missing values is higher and the gap size is larger, the existing methods perform poorly. The proposed method performs reasonably well in recovering large gaps as it incorporates both the time series characteristics and the information on other correlated variables to recover data. This algorithm can be recommended especially for meteorological datasets, as large gaps and correlated variables are ubiquitous in these data. This method is not applicable in univariate situations.

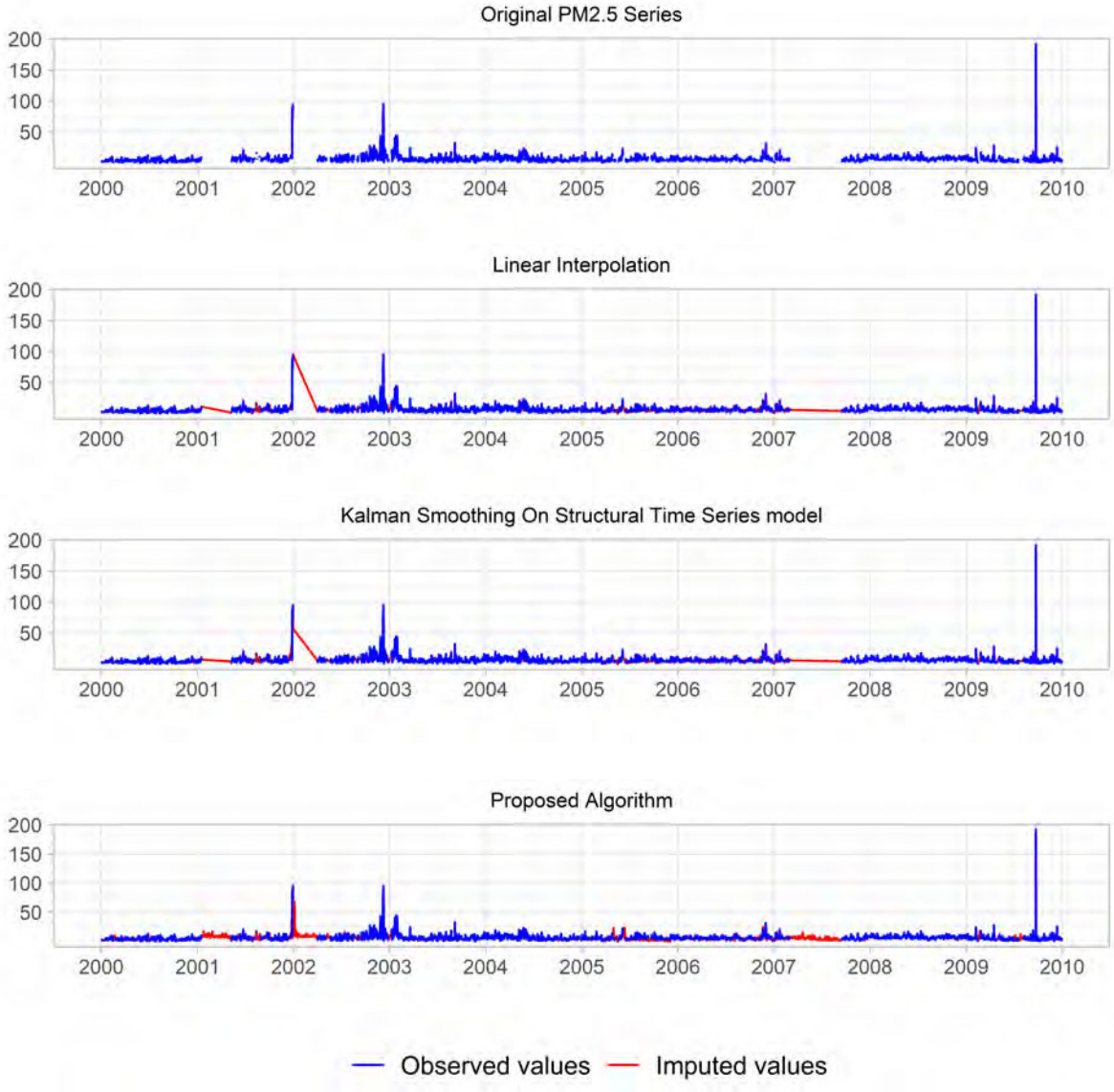


FIGURE 6.6: Visual Comparison

6.7 Contribution

The main contribution of this chapter is that it incorporates both the time series characteristics of the series itself and the information available on other highly correlated variables to impute large intervals of missing data. The strength of using elastic net regression which is a regularized regression method is that it minimizes the multicollinearity issues. This is important as the correlated predictors generally cause multicollinearity violating assumptions of basic regression models. To the best of the authors' knowledge, no such methodology has been proposed in the literature.

Chapter 7

Air Quality Data Analysis

This chapter is based on the following publication. The author's contribution to this paper is the concept and design of the work, data collection and cleaning, clustering of monitoring sites and interpretation.

- Wijesekara, L., Nanthakumaran, P., and Liyanage, L. (2023). Space and Time Data Exploration of Air Quality Based on PM10 Sensor Data in Greater Sydney 2015-2021. In International Conference on Sensing Technology (pp. 295-308). Cham: Springer Nature Switzerland.
  https://doi.org/10.1007/978-3-031-29871-4_30

7.1 Introduction

Exposure to air pollution creates numerous adverse effects on population health. Particulate matter in the air with a diameter of 10 micrometres or less (PM10) is one of the measurements which indicates outdoor air pollution. The common sources of PM10 include dust particles, smoke from fires, sea salt, car and truck exhausts and industry (Environment Protection Authority Victoria, 2021). Studies report that there is a link between exposure to PM10 with asthma (Donaldson et al., 2000; Pope III et al., 1991; Scibor & Malinowska-Cieslik, 2020), cataract (Shin et al., 2020), cardiovascular diseases (Lee et al., 2014; Polichetti et al., 2009), lung cancer (Zhou et al., 2017) and diabetes (Orioli et al., 2018; Yang et al., 2020b) . Moreover, long-term exposure to PM10 is reported to be associated with premature death (Kihal-Talantikite

et al., 2019). Therefore, it is essential to monitor air pollution and take necessary actions to reduce these health burdens.

National Environment Protection Council (NEPC), Australia has established air quality standards for key air pollutants including PM10 as part of the National Environment Protection Measure for Ambient Air Quality (Air NEPM). According to Air NEPM, the daily PM10 concentrations are recommended to be below $50\mu g/m^3$ and the goal is set not to allow for any exceedances, excluding exceptional event days (National Environment Protection Council (Australia), 2021). Exceptional events include fire or dust occurrence that is directly related to bushfire, jurisdiction-authorised hazard reduction burning or continental scale windblown dust (NSW Department of Planning and Environment, 2021). Therefore, analysis of exceedances is important to monitor whether the NEPC goal is achieved and to plan actions accordingly.

Sensors are used to monitor ambient air quality. Implementing cost-effective air quality monitoring systems is a growing research area (Abraham & Li, 2014; Liu et al., 2020; Morawska et al., 2018; Zheng et al., 2016). Therefore, it is important to make the most use of the collected data. Identifying space-time characteristics of the pollutants is useful to make precautionary arrangements for the adverse effects of poor air quality. To the best of the authors' knowledge, no published study can be found by analysing Greater Sydney's PM10 during the period 2015-2021 which includes black summer (late 2019-early 2020) and COVID-19 lockdown periods (first lockdown March-April 2020 and second lockdown June-August 2021) (Duc et al., 2021). Duc et al.(Duc et al., 2021) have studied the effect of the lockdown period on air quality considering a set of air pollutants which does not include PM10. Also, grouping locations based on the behaviour of PM10 over time will be useful in arranging similar precautionary measures in similar locations.

The objectives of this chapter are to explore space-time characteristics of daily PM10 concentrations and exceedances (greater than $50\mu g/m^3$) in Greater Sydney from 1 January 2015 to 31 December 2021 and to identify clusters of pollution sites. The findings of this chapter could be useful for the authorities to make evidence-based air policies

and recommend strategies for air quality improvement.

Section 7.2 of the chapter describes the data and the methodology used in the analysis. Section 7.3 presents the results of the analysis and finally section 7.4 presents the conclusions.

7.2 Data and Methodology

New South Wales air quality monitoring network continuously measures air pollutants through a network of NATA (National Association of Testing Authorities)- accredited stations. PM10 is measured using Tapered Element Oscillating Microbalance (TEOM). This conforms to Australian and International Standards (NSW Department of Planning and Environment, 2022). For this study, air quality data were accessed via an application programming interface (API) available on the website of the Department of Planning and Environment, NSW Government. PM10 daily measurements were downloaded from 1 January 2015 to 31 December 2021 at 18 monitoring sites in the Sydney region. A schematic representation of the methodology used in this study is given in Figure 7.1.

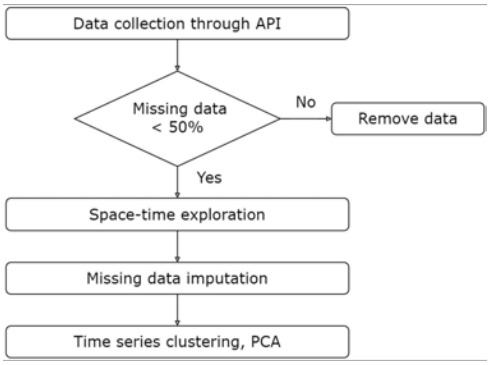


FIGURE 7.1: Methodology

Missing value percentages, longest gap sizes and the distribution of the degree of missingness over time (refer Appendix A) were analysed in each of the PM10 series at different sites. Sites, where the missing percentage is greater than 50%, were removed from the analysis. Space-time characteristics of the rest of the sites were analysed.

For the sites where the missing value percentage is less than 20%, missing values were imputed using Kalman Smoothing on Structural Time Series method in the imputeTS R package (Moritz & Bartz-Beielstein, 2017b) as it has been proven efficient for relatively small missing values (Moritz et al., 2015; Wijesekara & Liyanage, 2021a). Missing values with large gaps were imputed using a bi-directional method based on regularised regression models (Wijesekara & Liyanage, 2021a). Sites with more than 30-consecutive-day gaps and missing percentages greater than 50% were removed from the analysis. Time series clustering and principal component analysis (PCA) were carried out using the remaining cleaned dataset. Sites in this analysis, are shown in Figure 7.2 which includes 11 pollution monitoring sites.

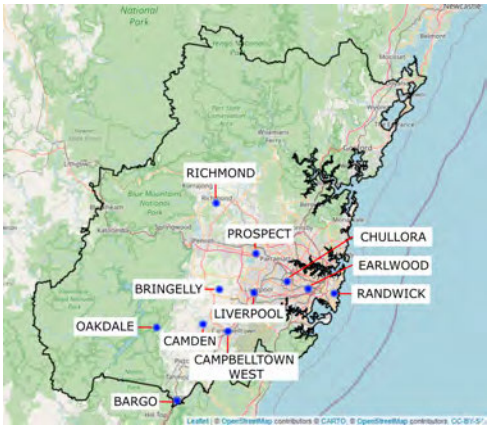


FIGURE 7.2: Air quality monitoring sites in the Sydney Region

A variety of methods are available for time series clustering in the literature. Some of them are partitioning, hierarchical, grid-based, model-based, density based and multi-step clustering (Aghabozorgi et al., 2015). In this analysis, agglomerative hierarchical clustering(James et al., 2013b) was applied for the sites (PM10 series). This method was chosen to understand the similarities of the sites using a dendrogram. Dynamic Time Warping (DTW) distance was used as the measure of dissimilarity. DTW is a shape-based dynamic programming algorithm that compares two series by finding the optimum warping path between them. For time series variables DTW distance make more sense than the Euclidean distance as it is less sensitive to noise, scale, and time shifts (Sardá-Espinosa, 2017).

Let two series be Q and C where $Q = q_1, q_2, ..., q_n$, $C = c_1, c_2, ..., c_m$ with lengths n and m

respectively. To calculate DTW distance, an n by m matrix is formed where element (i, j) contains the distance $d(q_i, c_j) =\mid q_i - c_j \mid$ between the two points q_i and c_j. Dynamic programming is employed to compute the cumulative distance matrix by applying the reiteration as follows:

$$\omega(i, j) = d(q_i, c_j) + min\{\omega(i-1, j=j-1), \omega(i-1, j), \omega(i, j-1)\}, \tag{7.1}$$

where $\omega(i, j)$ is the cumulative distance and *min* represents the minimum of the cumulative distances of the adjacent elements.

The best match between two time series sequences is the one with the lowest distance path $W = \{w_1, w_2, ..., w_k, ..., w_c\}$ after aligning one time series to the other.

DTW distance is given by the equation

$$DTW(Q, C) = \min \sum_{k=1}^{c} w_k \tag{7.2}$$

Warping paths are made by satisfying the following conditions.

(i) Boundary conditions: $w_1 = d(1, 1)$ and $w_k = d(n, m)$,

(ii) Continuity: $w_k = d\left(i_k, j_k\right)$ and $w_{k+1} = d\left(i_{k+1}, j_{k+1}\right), i_{k+1} - i_k \leq 1$ and $j_{k+1} - j_k \leq 1$,

(iii) Monotonicity: $i_{k+1} - i_k \geq 0$ and $j_{k+1} - j_k \geq 0$.

After calculating the DTW distances among the selected PM10 time series, agglomerative hierarchical clustering (James et al., 2013b) was applied and the dendrograms were analysed. Average linkage was chosen to measure the inter-cluster distances. DTW distance-based clustering is rarely used to cluster pollution monitoring sites (Suris et al., 2022). Recently, Suris et al.(Suris et al., 2022) have used DTW-based clustering to air pollution sites in Malaysia. They have used k-Means, partitioning, Fuzzy k-Means (FKM) algorithms and agglomerative hierarchical clustering based on complete linkage to cluster the air quality time series data. In contrast, this study used agglomerative hierarchical clustering based on average linkage and the clusters

were further justified using geographical locations, elevations and principal component analysis of the time series.

Considering the sites as observations and the time points as variables, principal component analysis(James et al., 2013b) was carried out and the data were visualised using the first two principal components.

7.3 Results and Discussion

The distribution of PM10 concentrations over space as well as time was explored to identify any interesting patterns. The exceedances were further analysed to identify the locations with frequent exceedances.

7.3.1 Space-time exploration based on daily PM10 concentrations

Figure 7.3 shows the distribution of daily PM10 concentrations at each site from 2015 to 2021 highlighting the days when PM10 concentration exceeded the Air NEPM threshold.

It can be seen that the daily PM10 concentrations at different monitoring sites each year follow a positively skewed distribution with spatially and annually varying patterns. On most days the PM10 concentrations had reasonably good air quality. However, the number of daily PM10 exceedances has increased considerably from 2018 to 2020 followed by a drop in 2021. The increase in 2019-2020 could be due to bushfires that happened during the black summer period (July 2019 - March 2020). However, it is worth noting that even in 2018, there was a considerable increase in the exceedances possibly indicating any future extreme events. The drop in 2021 is possibly due to less traffic and other effects of the COVID-19 lockdown.

The spatial variation is further explored as presented in Figure 7.4 which shows the mode of the daily PM10 level distributions at each monitoring site each year (i.e. the daily PM10 concentrations recorded most of the time at each monitoring site in each year).

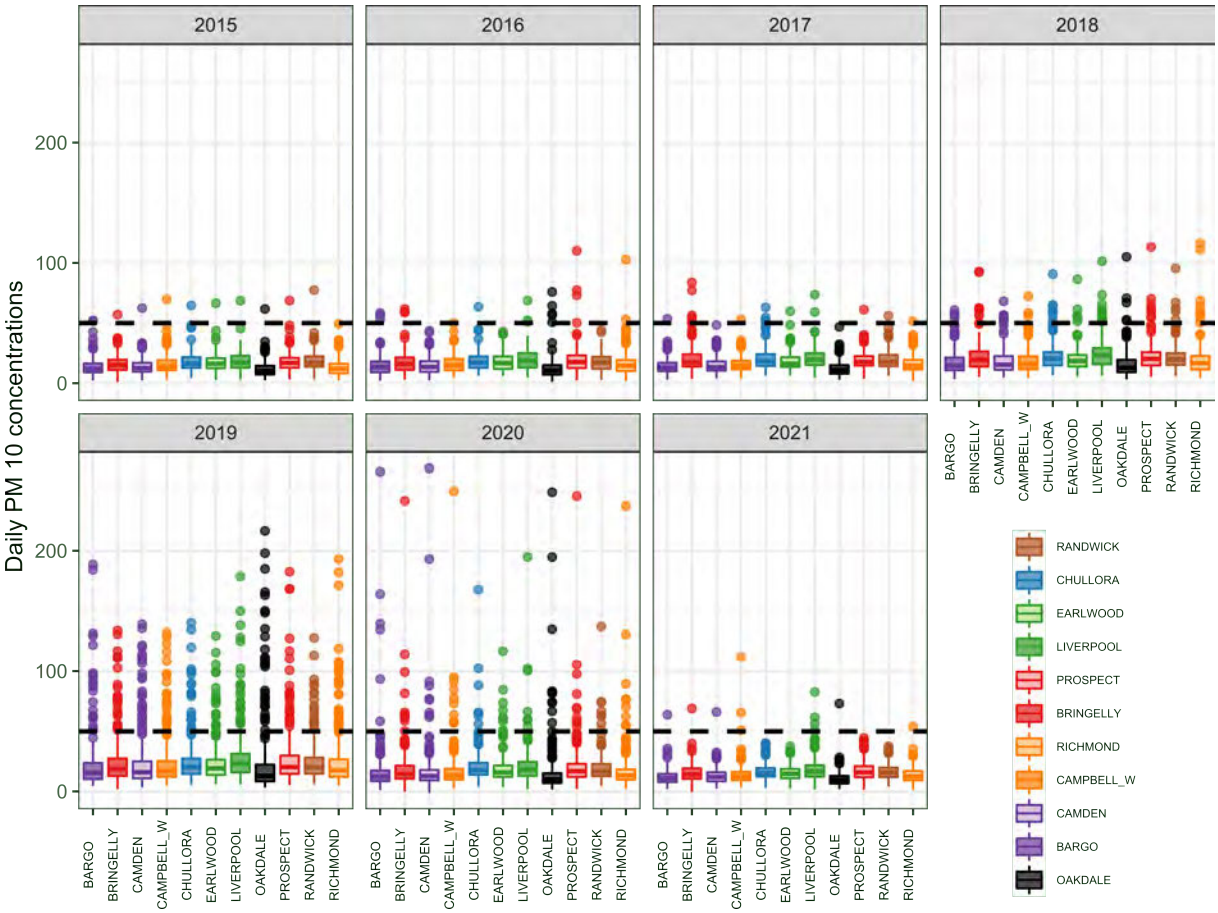


FIGURE 7.3: Box-plots showing the distributions of daily PM10 concentrations each year at different monitoring sites. The horizontal black dashed line indicates the Air NEPM threshold which is $50\mu g/m^3$.

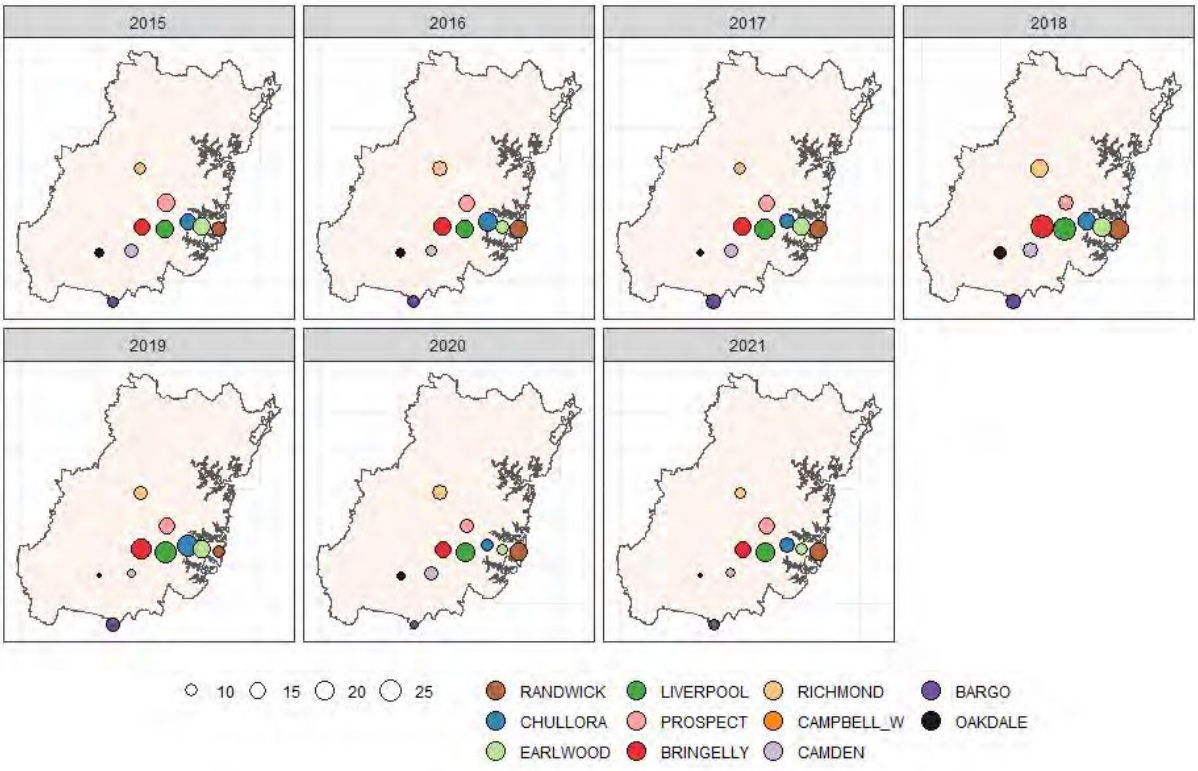


FIGURE 7.4: Space-time variation of the mode of daily PM10 concentrations each year at different monitoring sites.

It can be seen that the mode of the daily PM10 level distributions is high in metropolitan Sydney. It is worth noting that Oakdale recorded the lowest mode of PM10 concentration compared to other locations during the study period.

7.3.2 Space-time exploration based on daily PM10 exceedance

Figure 7.5 and 7.6 depict the number of exceedances at each site from 2015 to 2021.

As can be seen in Figure 7.6, there is a gradual increase in the number of exceedances from 2015 to 2019 followed by a sharp drop from 2019 to 2021. Figure 7.5 also reveals that the number of exceedances from 2015 to 2019 is increasing at all the 11 monitoring sites considered. However, the increasing trend is varying among different monitoring sites. It is worth noting that Liverpool recorded the highest number of exceedances (13 days) in 2018 which was a significant increase from 2017 (2 days). After 2020 (i.e. the first COVID lockdown period), the monitoring sites Chullora, Earlwood, Prospect and Randwick have not recorded any exceedance whereas all the other sites

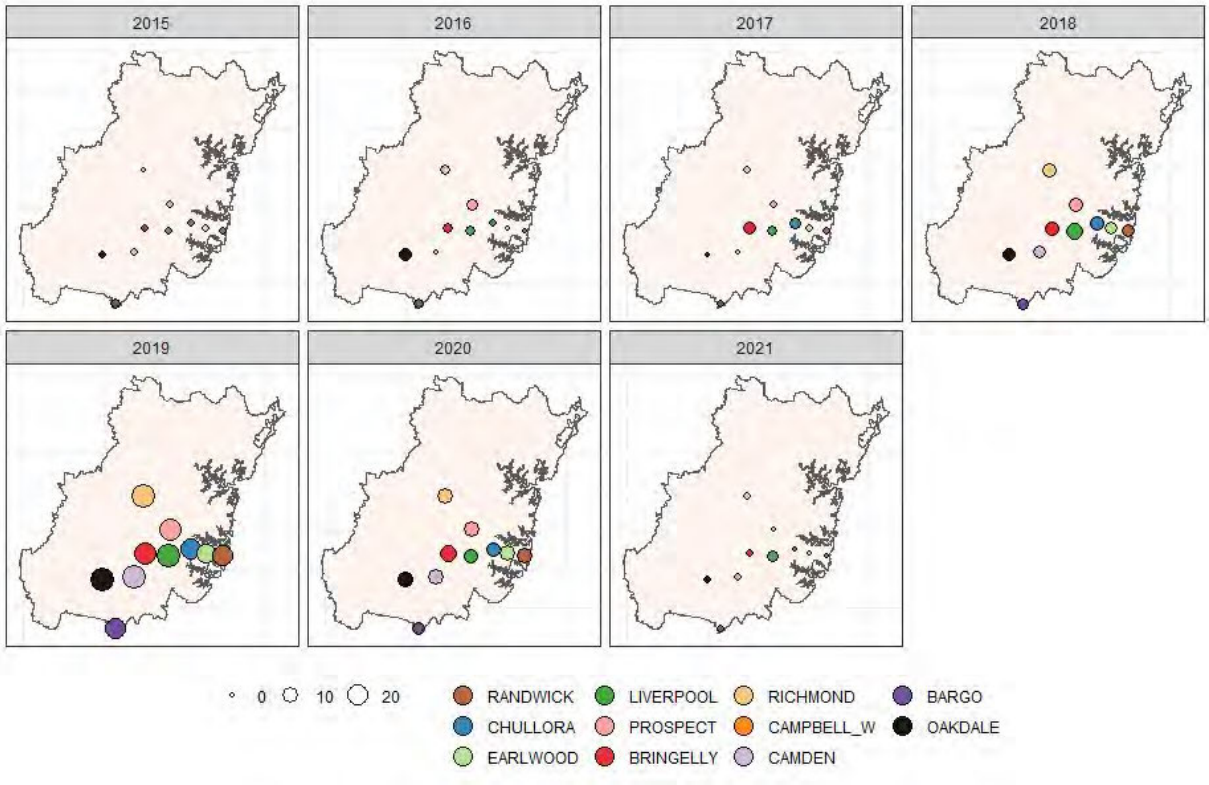


FIGURE 7.5: Space-time variation of the daily PM10 exceedances each year during 2015 - 2021 at different monitoring sites.

have recorded at least one exceedance. Again, it is worth noting that Liverpool has recorded the highest number of exceedances (4 exceedances) in 2021. Campbeltown West also has recorded 3 exceedances in 2021 and one of which is the highest PM10 concentration (above 100 µg/m3) in 2021. (refer Figure 7.3). The air quality in terms of PM10 exceedance has improved at all the monitoring sites except Campbelltown West and Liverpool after the black summer and the COVID-19 first lockdown period. Further examination is needed to identify the reasons behind the occurrences of high exceedance at Campbelltown West and Liverpool. Additional care should be given to improve the air quality near these places.

Figure 7.7 further displays the exceedances from 2015 to 2021 with four different seasons of the year.

According to Figure 7.7, there is a slight increase in the daily PM10 exceedances during 2017-2018 compared to 2015-2016. Spring in 2019 and Summer in 2019-2020 show

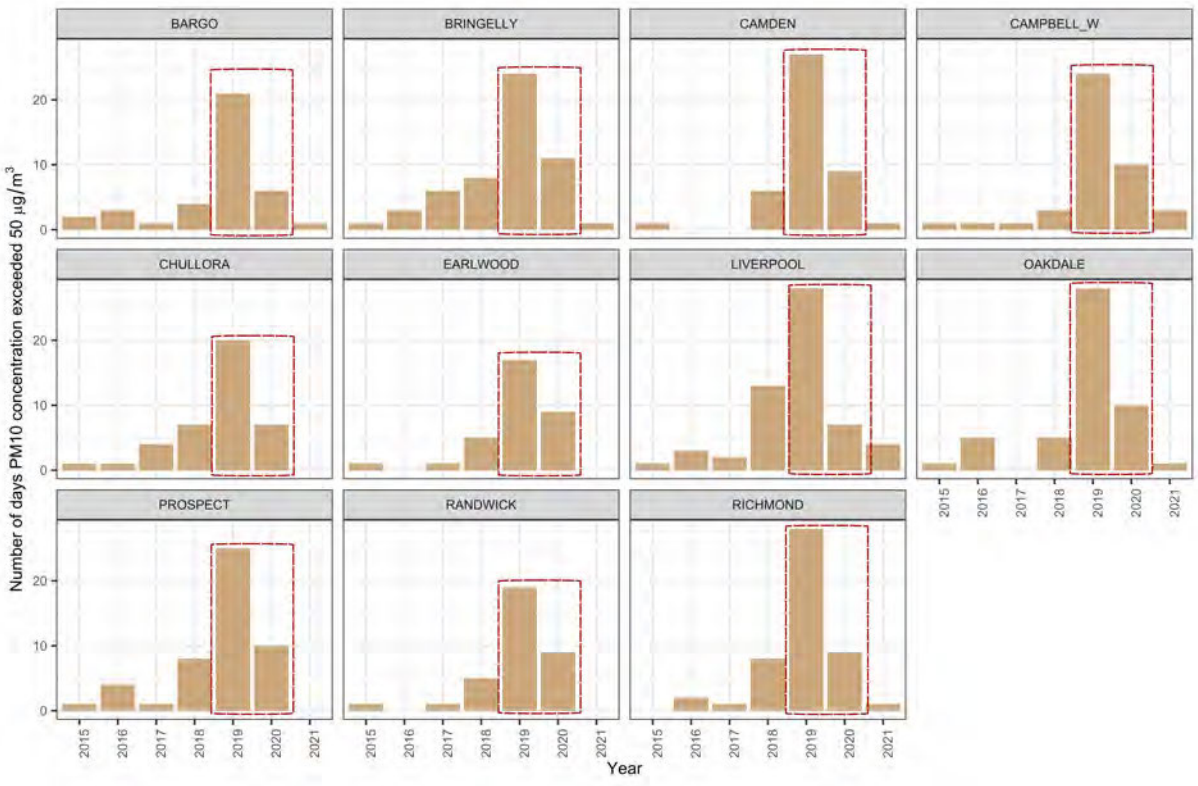


FIGURE 7.6: The daily PM10 exceedances at different monitoring sites in each year during 2015 - 2021. Red dash-lined boxes highlight the period 2019-2020 which includes Black Summer and COVID-19 first lockdown.

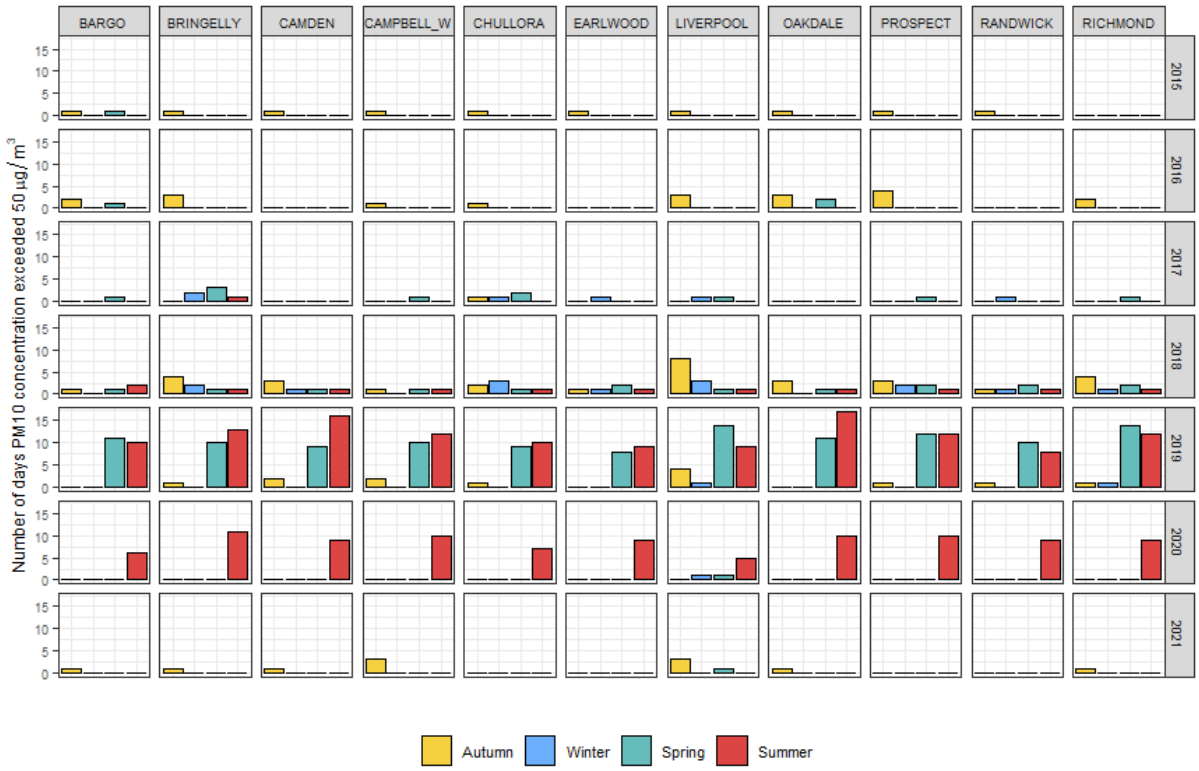


FIGURE 7.7: Number of daily PM10 exceedances each year during 2015 - 2021 with four seasons at different monitoring sites

remarkable exceedances as can be expected due to bushfires during the black summer. Exceedances have rarely occurred after 2020 (after the first COVID-19 lockdown) compared to the period 2015 - 2018 (before black summer and COVID-19 lockdown) revealing an improvement in air quality in terms of exceedances. PM10 exceedances have occurred mostly during the autumn season in all the years considered except in 2017, 2019 and 2020. In 2017, most of the monitoring sites recorded PM10 exceedance during the Spring season followed by winter.

7.3.3 Clustering of monitoring sites

To identify any groupings of sites with similar characteristics, clustering has been carried out. As an initial step, the distances among the sites have been analysed. Figure 7.8 shows a heat map of the distances between each pair of stations. The upper triangle of the heat map shows distances in meters while the lower triangle shows DTW distances between the PM10 series of each pair.

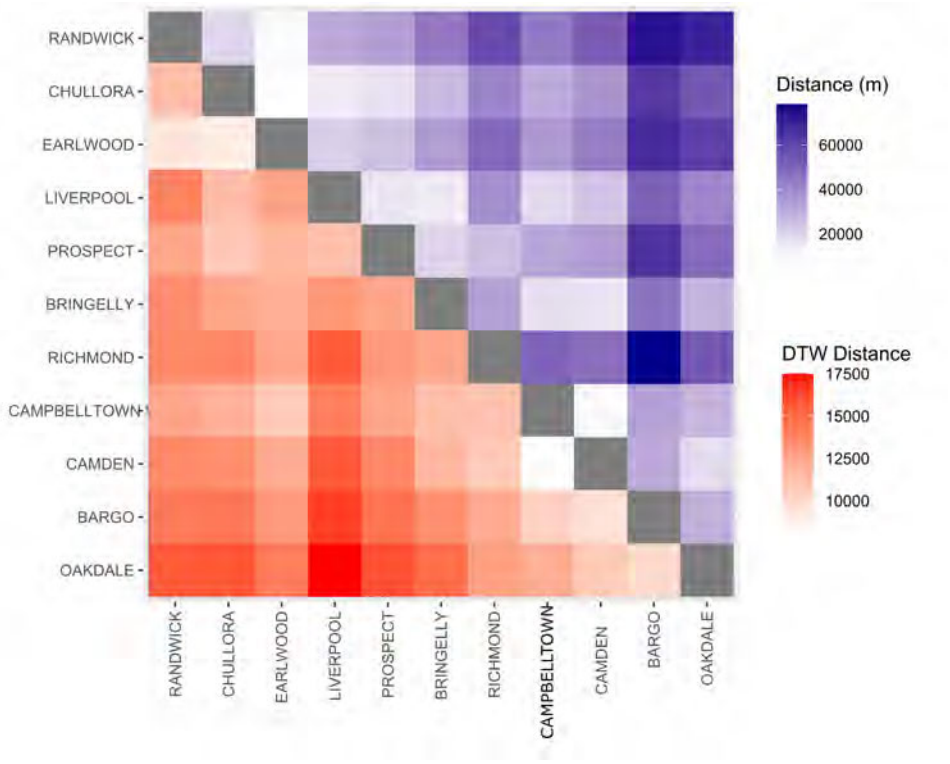


FIGURE 7.8: Distances among the sites

Figure 7.8 reveals some symmetrical patterns around the diagonal of the heat map. The colour densities of the top-right and bottom-left tiles are high indicating that, when the actual distance between the two sites is high, the dissimilarity of the corresponding two PM10 series is also high.

After applying agglomerative hierarchical clustering using DTW distance as the dissimilarity measure, the dendrogram was analysed to identify possible clusters. The cluster dendrogram is presented in Figure 7.9 (a) which displays four interesting clusters. Principal component analysis was also carried out to visualise the clusters as in Figure 7.9 (b). Elevations of each site are given in Figure 7.9 (c) and the locations are given in Figure 7.9 (d) to compare cluster locations.

There are some clusters according to the dendrogram presented in Figure 7.9 (a). Camden and Campbelltown West PM10 series are closely related. Bargo and Oakdale series are also closely related even though they are far apart than Camden and Campbelltown West. All these four sites can be considered as one cluster comparatively. Richmond is closer to this cluster than all the other sites. Chullora-Earlwood

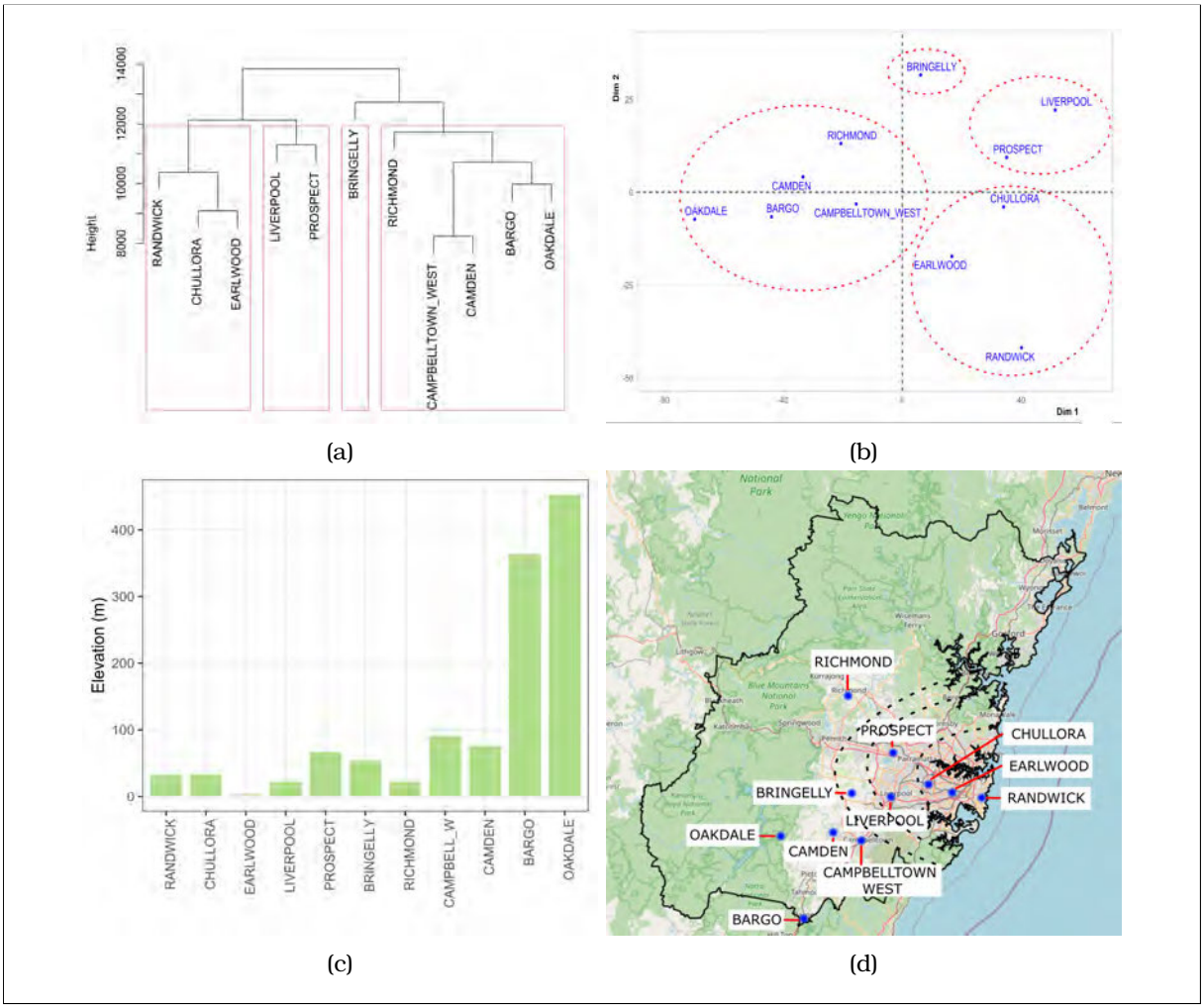


FIGURE 7.9: Cluster Analysis of the sites. (a) Cluster dendrogram with four clusters boxed (b) PCA plot with first two principal components (Dim1 and Dim2). Dashed ellipses roughly separate the four clusters in Figure (a). (c) Bar chart showing the elevations of the sites (d) Spatial distribution of each site with black dashed lines roughly separating the four clusters in Figure (a).

and Liverpool-Prospect are two pairs which are closely related. Randwick shows the closest behaviour to the Chullora-Earlwood pair of sites. Bringelly displays a different behaviour compared to other sites. If the interested number of clusters is four, it can be represented by the four red boxes.

PCA graph of the sites shown in Figure 7.9 (b) also provides further support for the identified clusters. Four ellipses represent the four clusters shown in the dendrogram. It can be seen that the sites within a cluster are located relatively closer in the 2-dimensional space of the first two principal components.

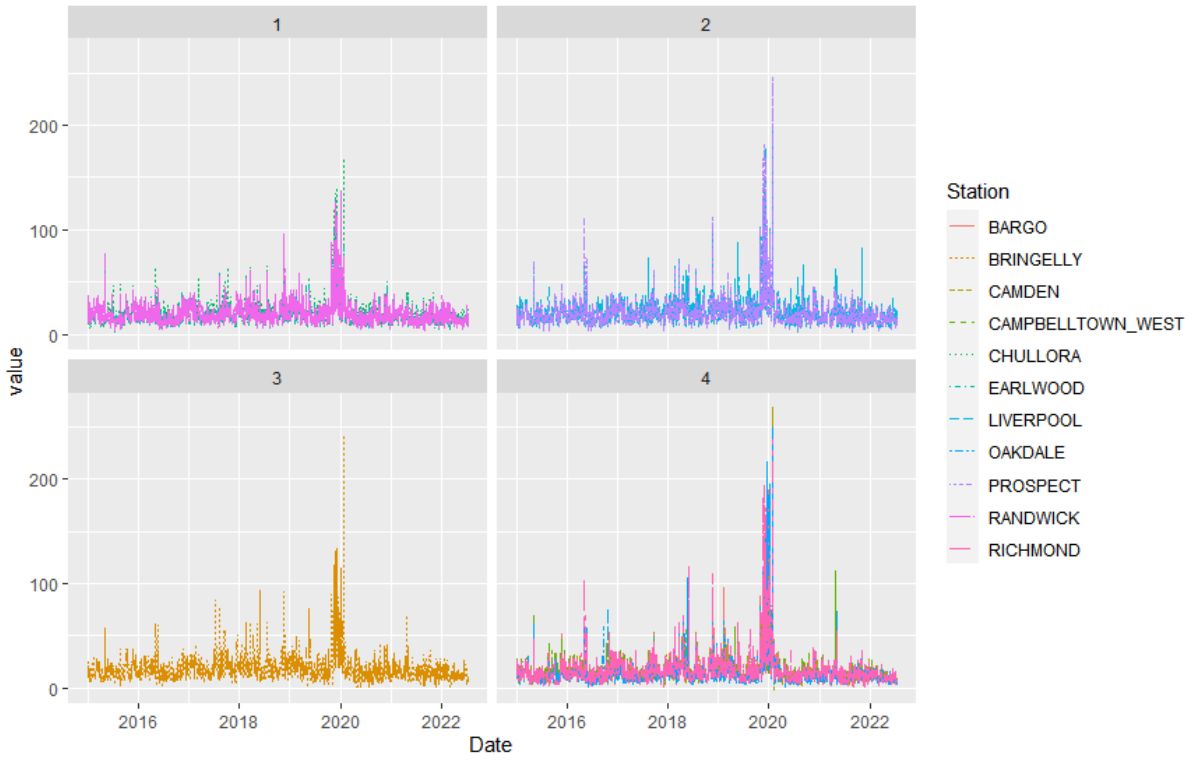


FIGURE 7.10: PM10 time series clusters

Considering the elevations presented in Figure 7.9 (c), Bargo and Oakdale have relatively higher elevations which may be a reason for their PM10 similarity. Campbelltown West and Camden show approximately similar elevations and their PM10 behaviour is also closely related. However, elevation itself does not explain the other clusters.

It is also interesting that the four clusters display a pattern with the geographical locations as can be seen in Figure 7.9 (d). Dashed lines approximately separate the clusters. However, note that they are only an indication to separate the clusters and do not represent any geographical boundaries or any other groupings. Geographical locations as well as the PM10 behaviour of Chullora, Earlwood and Randwick are relatively close, supporting Tobler's first law of geography that states "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). Liverpool and Prospect also display the same pattern (i.e. relatively closer in location as well as PM10 behaviour). Even though Bringelly site is relatively closer to Liverpool site, it is far apart considering the behaviour of the PM10 series. This is an interesting fact for further investigation. However, only the geographical location itself

cannot explain the behaviour of the PM10 concentration as there may be other factors associated with the PM10 level. Richmond site is far from other sites. However, its PM10 series is related to the two pairs Camden-Campbelltown and Bargo-Oakdale.

Considering all this information, four reasonable clusters can be identified among the sites considered in the Greater Sydney Region as below:

1. Randwick, Chullora, Earlwood
2. Liverpool, Prospect
3. Bringelly
4. Richmond, Campbelltown West, Camden, Bargo, Oakdale

Figure 7.10 shows the time series plots of PM10 daily concentrations for each of the four clusters providing graphical evidence of some overlapping series within a cluster.

7.4 Conclusion

In this study, PM10 air pollution sensor data of the Greater Sydney Region from 2015-2022 were analysed to identify patterns. The mode of the daily PM10 levels distribution was varying spatially. Daily PM10 levels exceeded the national air quality standards in the study period during the autumn season. After the first COVID lockdown period, the number of days daily PM10 levels exceeded the national air quality standards and was reduced at all the monitoring sites except Campbelltown West and Liverpool. Further examination is needed to identify the reason behind these occurrences of high exceedance days at these sites. Additional care is to be given at Campbelltown West and Liverpool sites to improve the air quality at these places. Time series clustering indicates four possible clusters of sites according to the behaviour of the PM10 sensor data. The four clusters are Randwick-Chullora-Earlwood, Liverpool-Prospect, Bringelly and Richmond-Campbelltown West-Camden-Bargo-Oakdale.

Chapter 8

Application of Machine Learning in Health: Case Study with Asthma

The analysis of this chapter uses a readily available integrated asthma dataset pertaining to a hospital in Victoria State. The purpose of this chapter is to demonstrate the methodology for identifying environmental conditions on diseases and to highlight the value of data integration. This analysis was carried out before diabetes data was available. It presents an application of machine learning methods in predicting the risk of asthma based on air pollution and weather. The methods applied in this chapter may apply to other diseases associated with environmental factors. In future, these methods could be applied in the categorization of individuals who may be at risk of developing a disease or areas at high risk using environmental conditions. The future direction is to carry out a similar analysis for diabetes in the NSW State. This chapter is based on the following publication.

- Wijesekara, L., and Liyanage, L. (2020, November). Modelling Environmental Impact on Public Health using Machine Learning: Case Study on Asthma. In 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA) (pp. 1-7). IEEE. https://doi.org/10.1109/CITISIA50690.2020.9397488

8.1 Introduction

World Health Organization (WHO) reports reveal that 24% of the global burden of disease can be attributed to environmental factors (Prüss-Üstün et al., 2016). Connections between the environment and health have been identified ages ago. Hippocrates described the link between health and environmental characteristics such as water quality, air quality and living place in 400 B.C. (Jouanna, 2012). Monitoring the environmental impact on health can be identified in three ways as hazard surveillance, exposure surveillance and outcome surveillance. Hazard surveillance is the process of identifying, monitoring and modelling environmental hazards. Exposure surveillance involves examining the extent of exposure and biological process of unhealthy effects whereas outcome surveillance involves recording and monitoring clinical indicators of ill-health (Maantay & McLafferty, 2011). In addition to investigating causal analyses, it is always useful to study associations of environmental factors such as weather and pollution with diseases to build a comprehensive understanding of links between the environment and health. It will help to manage and mitigate not only the burden of diseases but also other issues in climate change using a holistic approach.

To better understand the impact of the environment on health, it is essential to integrate different sources of data such as demography, weather, pollution, hospital data etc. Some researchers have paid much effort to the data integration process as it is essential for any further studies (Comer et al., 2011; Liyanage & Liyanage, 2010). Liyanage et al. have developed a spatiotemporal data integration platform to evaluate the environmental impact on public health. It currently includes weather, pollution, demography and health data of some hospitals in Victoria state, Australia. The goal of this platform is to build an intelligent decision-making framework to evaluate the environmental impact on public health, especially on asthma, diabetes, obesity and cataract.

This chapter presents a case study on asthma using a dataset obtained from the aforementioned platform as an early step of the decision-making framework. The objective

of this chapter is to identify classification models which could be used to predict vulnerability and risk of getting future episodes of asthma based on weather and pollution conditions. Moreover, this will contribute to the knowledge of asthma distribution in Victoria state and also to the process of data preparation and modelling with integrated data consisting of weather, pollution and health data. Results could be used for policymaking to manage and mitigate the burden of asthma in Victoria State. More importantly, it could be expanded to a real-time application to alarm the likeliness of a disease based on patients' general characteristics and environmental conditions.

8.2 Related work

Asthma is a common chronic non-communicable disease that leads to reduced quality of life. The recent records of asthma depict that as many as 339 million people of all ages worldwide have been affected by asthma. Asthma is the 16^{th} most important disorder in the world in terms of the extent and duration of disability ("The Global Asthma Report 2018", 2018). The reports of the Global Asthma Network (GAN) suggest that research in this field is much needed as the number of hospital admissions is too high in some periods. Air pollution can be considered as one of the main causes of asthma. A plethora of studies show the effect of air pollution on public health (Bousquet et al., 2018; Breysse et al., 2010; Guarnieri & Balmes, 2014; Koenig, 1999; Liu et al., 2019). Some studies have argued that the association between air pollution and asthma prevalence is weak as the prevalence of asthma and other allergic diseases has been rising in many countries in some periods, at a time when exposure to most measured air pollutants has declined. Indoor air quality is one of the main determinants of personal exposure to pollution. This might be a reason for the weak association between outdoor air pollution and asthma prevalence (Strachan, 2000). This raises the need for analysing these associations thoroughly in a geographically specific manner. Also, it suggests that analysing air pollution and the prevalence of asthma alone may produce misleading information. Therefore, it is important to consider the other variables as much as possible to draw useful conclusions. There are

ample studies which show evidence that climatic variables have significant associations with asthma prevalence (Zanolin et al., 2004). Most of the available literature is based on clinical or experimental studies and focused on causal relationships. This chapter involves studying the associations and building a model to classify the presence of asthma episodes using the information available on environmental conditions and the patient's general characteristics.

8.3 Methods

8.3.1 Support Vector Machine (SVM)

This is a generalisation of the maximal margin classifier. The maximal margin classifier performs the classification based on a hyperplane. In a p-dimensional scale, a hyperplane is a flat affine subspace of dimension p-1. The equation of a hyperplane can be written as,

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p = 0, \tag{8.1}$$

where Xs are the features, β_0 is the intercept and other βs are coefficients/slopes of the features.

SVM with radial kernel deals with non-linear decision boundaries. This enlarges the feature space so that separation is possible in the enlarged or high-dimensional feature space. Enlarging the feature space can be done using kernels. Radial kernel function is of the form,

$$K(x, y) = exp(-\gamma \sum_{j=1}^{p} (x_{ij} - y_{ij})^2). \tag{8.2}$$

SVM with radial kernel uses hyper-parameters cost and gamma (γ). The optimal values for these parameters are estimated through cross-validation (James et al., 2013a).

8.3.2 Artificial Neural Network (ANN)

In a neural network for K-class classification, there are K units at the output layer, with the k^{th} unit modelling the probability of class k. There are K target measurements Y_k, k = 1, . . . ,K, each being coded as a 0 - 1 variable for the k^{th} class. Then,

$$Z_m = \sigma\left(\alpha_{0m} + \alpha_m^T X\right),\ m = 1, ..., M, \tag{8.3}$$

$$T_k = \beta_{0k} + \beta_k^T Z,\ k = 1, ..., K, \tag{8.4}$$

$$f_k(X) = g_k(T),\ k = 1, ..., K, \tag{8.5}$$

where $Z = (Z_1, Z_2, Z_3, ..., Z_M)$ and $T = (T_1, T_2, ..., T_K)$.

Derived features Z_m are created from linear combinations of the inputs, and then the target Y_k is modelled as a function of linear combinations of the Z_m. The symbol σ denotes the activation function which can be sigmoid, rectified linear unit (ReLU), etc (Friedman et al., 2001).

$$Sigmoid(v) = \frac{1}{1 + e^{-v}} \tag{8.6}$$

$$ReLU(v) = max(0, v) \tag{8.7}$$

8.3.3 Decision Tree and Random Forest

Decision tree classifies an observation by segmenting the feature space into several simple regions. The splitting rules used to create these segments can be summarised in a tree, hence known as decision tree. It splits nodes based on available input features by selecting the input feature resulting in the best homogenous dataset. To select the feature to be used for splitting, it uses either the Gini index or Cross-entropy.

$$Gini\ Index, G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}), \tag{8.8}$$

where K is the number of classes and p_{mk} is the proportion of training observations in the m^{th} region that are from the k^{th} class.

$$Cross\ entrophy, D = -\sum_{k=1}^{K} \hat{p}_{mk} log p_{mk} \tag{8.9}$$

In random forest, it builds several decision trees on bootstrapped training samples and considers the aggregated decision. Therefore, it reduces the problem of over-fitting compared to decision trees. Also, it works well with large datasets with better predictive accuracy. However, compared to decision trees it is less interpretable (James et al., 2013a).

8.3.4 Regularization

Regularisation shrinks the parameters by adding a penalty to the loss function and hence reduces the problem of over-fitting. Ridge regression which is also known as L2 regularisation is of the form,

$$Loss = Error + \lambda \sum_{i=1}^{p} w_i^2, \tag{8.10}$$

where w's are the weights/parameters and λ is the penalty.

8.4 Data Preparation and Preliminary Analysis

The dataset used for this study was obtained from the spatiotemporal data integration platform developed by Liyanage et al (Liyanage & Liyanage, 2010). The goal of this platform is to develop an intelligent decision-making framework to evaluate the environmental impact on public health. The dataset contains a set of variables of 108 patients admitted to a particular hospital in Victoria State in Australia from 2013 to 2015. The data have been integrated with weather and pollution variables of the nearest weather/pollution site for the patients considering the time of hospital admission. The data cleaning process was done at the time of data integration. Missing

values of the weather and pollution variables were imputed using Kalman smoothing on structural time series models (Moritz & Bartz-Beielstein, 2017b; Moritz et al., 2015; Wijesekara & Liyanage, 2020c) for univariate time series. A cleaning algorithm developed using seasonal decomposition and elastic-net regression was applied for large gaps when correlated variables exist (Wijesekara & Liyanage, 2020a). A description of the variables is given in Table 8.1.

TABLE 8.1: Variable Description

Variable	**Description**
Age	Categorical
Gender	Categorical
Triage	Categorical
Length of stay	Length of hospital stay (in minutes)
CO	Carbon Monoxide (parts per million)
O3	Ozone level (parts per billion)
NO2	Nitrogen Dioxide (parts per billion)
SO2	Sulfur Dioxide (parts per billion)
PM10	TEOM (tapered element oscillating microbalance) particles less than 10 micron in ug/m3
Air Quality Index	Standard measurement of air quality
Visibility reduction	Minimum visible distance - 20 km (equivalent to a visibility reduction index of 2.35)
Precipitation	Rainfall in millimeters
Relative humidity	Amount of moisture in the air as a percentage of the amount the air can actually hold
Vapour pressure	Partial pressure of water vapour in the atmosphere in hectopascal (hPa)
Wind speed	Speed of the wind in Km/h
Max wind speed	Maximum wind speed
Asthma	Binary indicator of the diagnosis of asthma

In addition to the variables in Table 8.1, the suburb of patients was used only for the descriptive analysis to identify the spatial distribution of asthma patients. Figure 8.1 shows the spatial distribution of patients by the variables. The colour of each plot represents the maximum value of the variable in the dataset. Dark colours represent

high values whereas light colours represent low values. The size of the black dots in each suburb is proportional to the number of individuals diagnosed with asthma in each suburb. It is interesting to see that as the maximum value of the variable in suburbs increases the number of asthma-positive individuals also increases for variables CO, O3, NO2, SO2, PM10, visibility reduction and precipitation. However, these plots are not sufficient to conclude this pattern as the number of total individuals in each suburb may not be representative. Also, the location of the hospital may have an impact as the individuals usually tend to admit to the nearest hospital.

Principal Component Analysis (PCA) was carried out to understand the pattern and relationships of the variables and the individuals in the dataset. Firstly, age and triage variables were converted into rank variables as they are ordinal. PCA plot of the variables and PCA plot of the individuals are shown in (a) and (b) in Figure 8.2 respectively. The variables are shown in arrows and the observations are shown in points. The angle between the two arrows reflects the correlation between those two variables. There can be seen that highly correlated variable sets are present such as [CO, NO2], [PM10, AQI, vapour pressure] and [visibility reduction, SO2]. The colour scale from blue to red represents the contribution of variables from lowest to highest. It is interesting to note that PM10, air quality index and visibility reduction contribute to explaining a larger proportion of variation in the dataset. Even though asthma is a categorical variable, it is included in the plot as its value is either 1 indicating asthma presence or 0 indicating otherwise. As can be expected asthma is positively correlated with relative humidity whereas negatively correlated with triage. However, these results are not sufficient to make conclusions. The labelled individuals in Plot (b) are the first 10 individuals who strongly contributed to the construction of the plane. Plots (c) and (d) are the PCA plot of variables and the PCA plot of individuals respectively, after considering age, triage and asthma as categorical variables. Plot (c) shows consistent results of the relationship among variables with Plot (a). In addition to the information present in Plot (b), Plot (d) shows asthma-positive individuals in black colour whereas others in red colour and 95% confidence ellipses corresponding to the two categories. Even though the two classes are mixed in the plane of PC1 and PC2, it is interesting

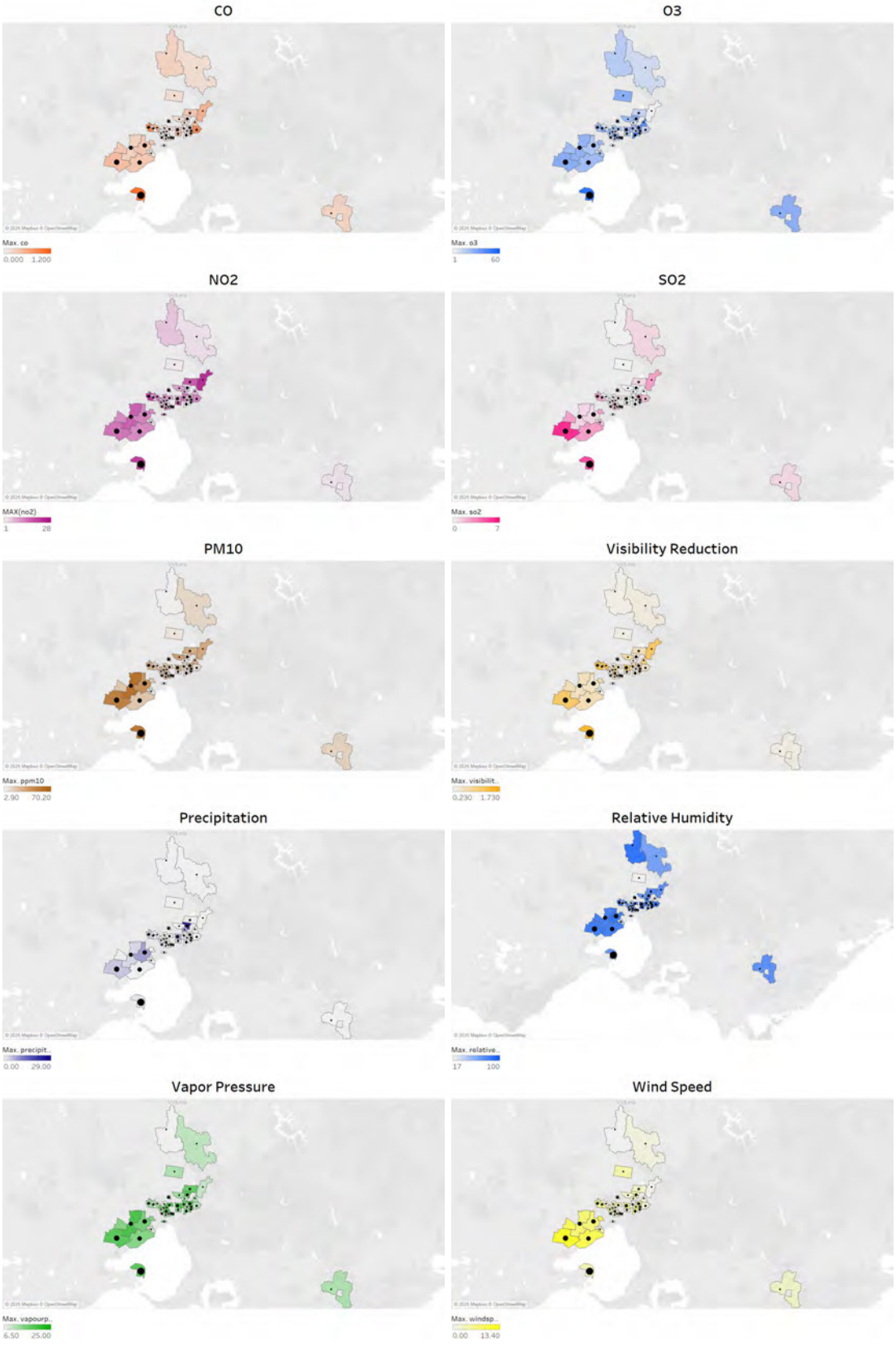
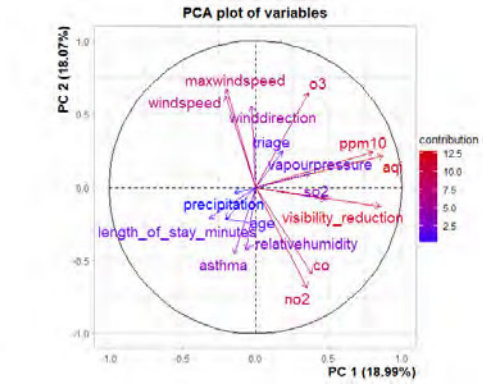


FIGURE 8.1: Spatial distribution of asthma patients by variables

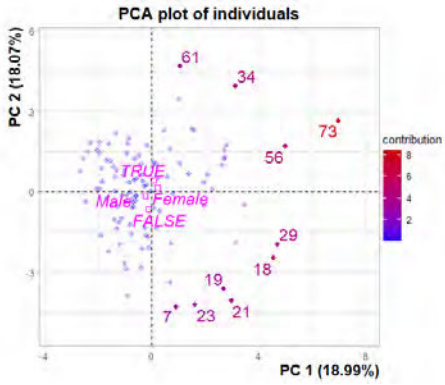
to note that the two ellipses do not overlap with each other.

According to PCA with numerical variables (excluding age and triage), the first two principal components explained only 42.95% of the variation. To explain more than 90% of the variation, it requires at least the first 9 PCs. Therefore the original variables were used for the classification as the original dimensionality is relatively low.

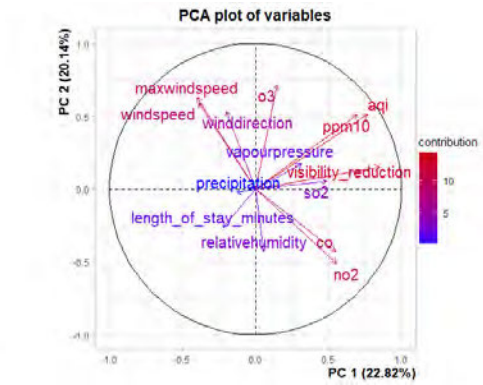
Length of stay and triage were studied only in the preliminary analysis and it is not used in classification models as the focus was on weather and pollution variables.



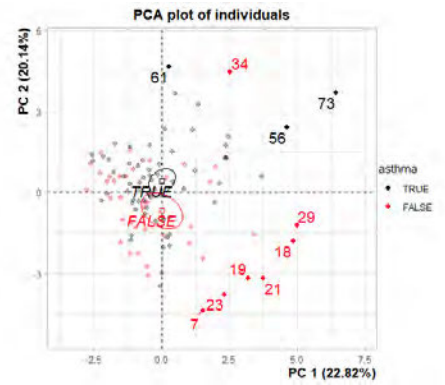
(a) PCA plot of variables including rank variables



(b) PCA plot of individuals including rank variables



(c) PCA plot of variables excluding rank variables



(d) PCA plot of individuals excluding rank variables

FIGURE 8.2: Principal Component Analysis results

8.5 Modeling and Evaluation

The dataset was divided into two parts, 80% training set to be used in model building and 20% testing set to be used in evaluating and comparing the models. Table 8.2 shows the distribution of the two classes in the datasets.

Table 8.2: Class Distribution

Class	**Full dataset**	**Train set**	**Test set**
Asthma True	39%	37%	45%
Asthma False	61%	63%	55%

It can be seen that the two classes do not show a problematic imbalance in both full and training sets. Division of the two classes in test set appears to be well balanced.

8.5.1 Support Vector Machine

After applying one-hot encoding using dummy variables for the nominal variables and treating ordinal variables as numerical variables, SVM models were fitted to classify the presence of asthma in individuals. SVM with radial kernel was used as the data appeared to have a non-linear decision boundary. The optimal hyperparameters were chosen by applying grid search cross-validation. Performance metrics were obtained for the testing dataset and Table 8.3 shows the performance of the SVM model.

Table 8.3: Performance of SVM

Class	**Recall/Sensitivity**	**Precision**	**F-score**	**Accuracy**
Asthma True	0.50	0.50	0.56	0.64
Asthma False	0.75	0.64	0.69	

This model does not perform well based on precision and recall. Out of the total asthma-positive individuals, 50% were correctly classified and among the individuals who are predicted as positive, 50% of them are actually positive. This model has yielded an accuracy of 64% which is quite well for this dataset.

8.5.2 Artificial Neural Network

Artificial Neural Network (ANN) was fitted with a deep learning architecture. After experimenting with several architectures of neural networks, an architecture was chosen by considering the low complexity, higher performance and computational power.

Even though this dataset is relatively small, it is expected to grow these models for different scenarios in the data integration platform. Therefore, consideration was given to the complexity and computational power. The details of the final architecture are presented in Table 8.4

Table 8.4: Neural Network Architecture

Layer	**Number of nodes**	**Activation function**	**Regularization**
Input	18 (original variables)		
Hidden layer 1	10	ReLU	Ridge Regression
Hidden layer 2	5	ReLU	Ridge Regression
Output	1	Sigmoid	Ridge Regression

The training dataset discussed in Table 8.4 was further divided into two sets, one for training(80%) and the other for validation(20%). The training was done with several initial random states. The optimal weights were calculated using backpropagation with Adam optimisation algorithm (Kingma & Ba, 2014). Binary cross-entropy loss function was chosen as the objective function as it appeared to perform significantly better than squared-error objective functions in neural networks (Kline & Berardi, 2005). Figure 8.3 shows the loss of training and testing sets of the model. It can be seen that ridge regression regularisation has effectively overcome the problem of over-fitting.

Table 8.5 shows the performance metrics of neural network model obtained for the test dataset.

Table 8.5: Performance of ANN

Class	**Recall/Sensitivity**	**Precision**	**F-score**	**Accuracy**
Asthma True	0.30	0.60	0.40	0.59
Asthma False	0.83	0.5	0.69	

This model also does not perform well. Sensitivity of the asthma true class is only 30% which very low. Its accuracy is lower than that of the SVM model. The reason for the poor performance of this model may be the insufficiency of training data to learn. It could be improved with more training data.

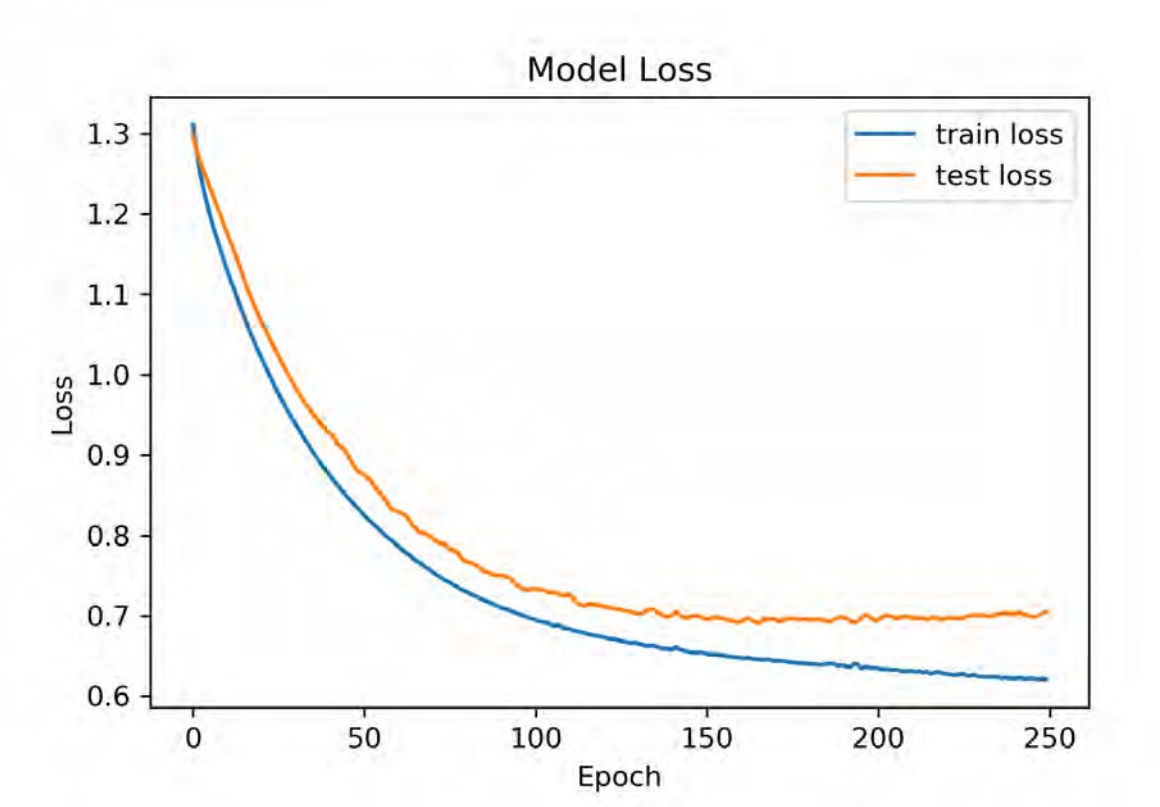


FIGURE 8.3: Loss of the model

8.5.3 Decision Tree

Decision tree was applied for the same training set used in the SVM and ANN models above and performance metrics were measured using the same test set. Decision tree pruning was done by using cross-validation and splitting rules were selected based on the Gini index. Table 8.6 shows the performance of this model.

TABLE 8.6: Performance of Decision Tree

Class	**Recall/Sensitivity**	**Precision**	**F-score**	**Accuracy**
Asthma True	0.70	0.64	0.67	0.68
Asthma False	0.67	0.73	0.70	

This model outperforms both SVM and ANN models based on sensitivity, precision and accuracy. Its sensitivity is reasonably well and it indicates that 70% out of the total asthma true individuals were correctly classified. Among the individuals who are predicted as positive, 67% of them are actually positive.

8.5.4 Random Forest

Random forest classifier is considered a robust method with high accuracy. Therefore a random forest model with 100 decision trees was built and the performance was measured with the test dataset. Table 8.7 presents the performance of the model.

TABLE 8.7: Performance of Random Forest

Class	Recall/Sensitivity	Precision	F-score	Accuracy
Asthma True	0.60	0.86	0.71	0.77
Asthma False	0.92	.73	0.81	

This model outperforms all the other methods with considerably higher performance metrics. It classifies more than three-quarters of individuals correctly. Also, it exhibits a reasonable level of sensitivity, precision and F-measures. However, decision tree performs better than random forest based on sensitivity.

Since all the models have their advantages and disadvantages, ROC (Receiver Operating Characteristic) curves for the test set were drawn for each of the classifiers. This allows for further comparison of the models. Figure 8.4 displays the ROC curves including AUCs (Area Under the Curve).

It can be seen that the AUC is highest in the random forest model whereas the lowest is in the SVM model. ANN shows a higher AUC than that of the decision tree and SVM.

8.6 Conclusion

Among the models considered for classification, the best accuracy of 77% was yielded from the random forest model. Its sensitivity, precision and F-measure are also reasonably higher compared to other models. The classifiers can be ranked from best to worst as; random forest, decision tree, SVM and ANN. Even though ANN gives the lowest accuracy of 59%, its learning curve exhibits a good fit based on the loss of training and validation. Therefore it could be improved using more training data from the data-integration platform. The decision tree gives the highest sensitivity of 70% among the methods considered. Sensitivity is more important than precision for this

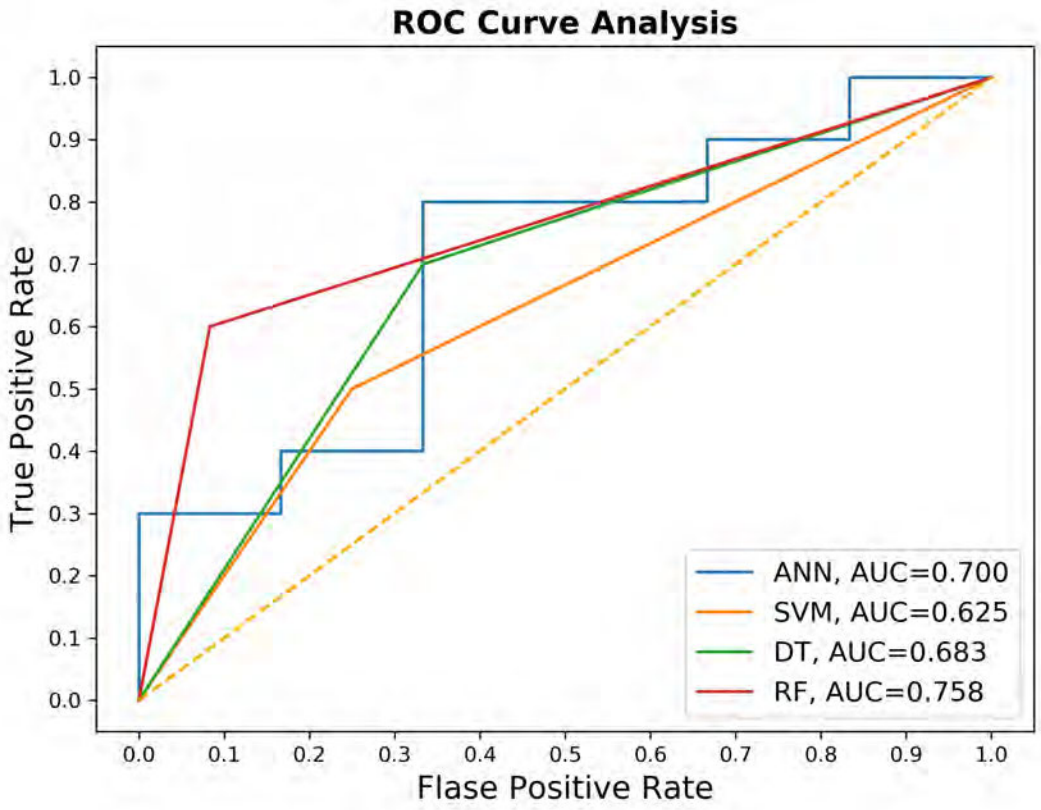


FIGURE 8.4: ROC curves for the test set

study as with a higher value of sensitivity, the model will be able to correctly predict a higher number of asthma-positive individuals out of the actual positive individuals. Even though the sensitivity of the random forest model (60%) is lower than that of the decision tree model, overall it was the best model for this study. This model could be used to produce initial alarms of future episodes of asthma based on weather and pollution predictions. This study could be further expanded to a real-time application to generate alarms for risky individuals to take precautions and thereby manage and mitigate the risk.

Chapter 9

Spatiotemporal Data Analysis on Diabetes

This chapter analyses Diabetes admissions in the NSW admitted patient dataset from 2013 to 2018.

9.1 Methods

9.1.1 Spatial autocorrelation measures

Spatial autocorrelation indices quantify the spatial dependence between values of the same variable in different places in space. The more the observation values are affected by values that are geographically close to them, the greater the spatial correlation. The following hypothesis can be tested to check for spatial autocorrelation.

Null Hypothesis: Data is randomly distributed (No spatial autocorrelation).

Alternative Hypothesis: Data is more spatially clustered than one would expect by chance alone.

Autocorrelation indices summarise the degree to which similar observations tend to occur near each other. The generic form of an index of spatial autocorrelation is of the form (Lee & Wong, 2001; Waller & Gotway, 2004),

$$\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\,\mathrm{sim}_{ij}}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}}, \tag{9.1}$$

where sim_{ij} denote the similarity between two observations (attribute similarity/distance of two observations) Y_i and Y_j, w_{ij} is the weight describing the proximity between locations i and j, for $i = 1, ..., N$ and $j = 1, ..., N$.

There are several measures to check the spatial autocorrelation. GLobal Moran's I (Moran, 1948) which is a well-established method in the spatial statistics. The Moran's I statistic is given by the following equation (Moran, 1948; Waller & Gotway, 2004).

$$\begin{aligned} I &= \left(\frac{1}{s^2}\right)\frac{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(Y_i-\bar{Y}\right)\left(Y_j-\bar{Y}\right)}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}}, \\ s^2 &= \frac{1}{N}\sum_{i=1}^{N}\left(Y_i-\bar{Y}\right)^2. \end{aligned} \tag{9.2}$$

The expected value of Moran's I under the null hypothesis of no spatial autocorrelation is given by,

$$E(I) = -\frac{1}{N-1},$$

and it approaches zero as N increases. Under the normality assumption, the variance is given by,

$$\mathrm{Var}(I) = \frac{N^2S_1 - NS_2 + 3S_0^2}{(N-1)(N+1)S_0^2} - \left(\frac{1}{N-1}\right)^2,$$

,

with $S_0 = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}, S_1 = 1/2\sum_{i=1}^{N}\sum_{j=1}^{N}\left(w_{ij}+w_{ji}\right)^2$, and $S_2 = \sum_{i=1}^{N}\left(w_{i+}+w_{+j}\right)^2$, with $w_{i+} = \sum_{j=1}^{N} w_{ij}$ and $w_{+i} = \sum_{j=1}^{N} w_{ji}($ (Cliff et al., 1981). The significance of the hypothesis can be compared using the calculated I with $z-score$, $z = [I - E(I)]/\sqrt{\mathrm{Var}(I)}$. However, the inferences based on this may not be appropriate when the normality assumption is not valid.

An alternative approach is suggested by Anselin et. al (2006) based on permutations under randomisation assumption (i.e. each value is equally likely to occur at any

location). Then, the test statistic is calculated by randomly shuffling the observations over the locations. Pseudo p-value is calculated using,

$$p = \frac{R+1}{M+1},$$

where M is the number of permutations, and R is the number of times the calculated statistic is equal to or more extreme than the observed statistics (Anselin et al., 2006).

A positive and significant I indicates clustering of similar values (High-Hgh, Low-Low, or a combination of both), whereas a negative and significant value indicates clustering of different values (checkerboard pattern or outliers).

In this study, Moran's I statistic was used to check for the spatial autocorrelation of the diabetes admissions per 1000 population in SA3 levels in NSW. In addition, the Moran scatter-plot was also used to analyse graphically (Anselin, 2019).

Spatial clusters

Global Moran I can be used only to identify clustering but not the locations of clusters. However, local indicators of spatial association (LISA) can be used to identify clusters (Anselin, 1995). Local version of the Moran's I is widely used as a LISA and is given by (Waller & Gotway, 2004),

$$I_i = (Y_i - \bar{Y}) \sum_{j=1}^{N} w_{ij} \left(Y_j - \bar{Y}\right),$$

for the i^{th} region/location.

This study used Anselin's (1995) index, which divides each deviation from the overall mean by the variance of the Y_i values. It can be represented as,

$$I_i = \frac{Y_i - \bar{Y}}{s} \sum_{j=1}^{N} w_{ij} \frac{Y_j - \bar{Y}}{s},$$

where s represents the square root of the sample variance of Y_i's.

FIGURE 9.1: Neiborhood structure

9.1.2 Spatial Proximity Matrices

Collection of weights w_{ij} is known as spatial proximity matrices.

$$w_{ij} = \begin{cases} 1 & \text{if regions } i \text{ and } j \text{ share a boundary,} \\ 0 & \text{otherwise.} \end{cases} \tag{9.3}$$

This study used the queen-neighbourhood structure as shown in Figure 9.1 for global and local autocorrelation measures.

9.1.3 Extreme value maps

Box maps and standard deviation maps were used to visualise and identify extreme values in maps. Box map represents 6 categories (Anselin, 1994). The middle point is the median; three categories are below the median (median to Q1, Q1 to minimum and lower extreme values) and three are above the median (median to Q3, Q3 to maximum and upper extreme value). In the standard deviation map, the data are standardised and the number of categories depends on the range of values. The middle point is the mean and categories depend on how many standard deviation units cover the data range.

9.1.4 Poisson and Quasi-Poisson regression for disease count modelling

If $Y \sim Poi(\lambda)$, then the probability Mass function is given by,

$$P(Y = y) = \frac{\lambda^y e^{-\lambda}}{y!}, \quad y = 0, 1, 2, \ldots \tag{9.4}$$

In this context, Y is the number of hospital admissions per 1000 population per year due to type 2 diabetes-related issues and it is assumed to follow a Poisson distribution. Observations are the different SA3 levels and they are assumed to be independent. Spatial autocorrelation among the SA3 levels was checked using Global Moran's I index. For n number of observations with p number of predictors x_j, for $j = 1, 2, ..., p$, the model is given by,

$$\ln\left(E\left[y_i \mid x_i\right]\right) = \ln\left(\lambda_i\right) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} = \mathbf{x}_i^T \boldsymbol{\beta}, \tag{9.5}$$

with $i = 1, ..., n$, with $\mathbf{x}_i^T = \left(1, x_{i1}, \ldots, x_{ip}\right)^T$ and $\boldsymbol{\beta} = \left(\beta_0, \ldots, \beta_p\right)$.

$$E\left[y_i \mid x_i\right] = \lambda_i = e^{\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}} = e^{\mathbf{x}_i^T \boldsymbol{\beta}}. \tag{9.6}$$

Parameters β_0 and β_js are estimated using maximum likelihood estimation. Predictors x_j are the annual averages of humidity, visibility, NO, NO2, PM2.5, PM10, SD1, S02, temperature and the average age of the population in each SA3. The number of hospital admissions was modelled considering the population as the offset.

Model assumptions are count responses, independent events, and constant variance (variance equal to mean). When the variance is not equal to the mean, the Quasi-Poisson model can be used by considering,

$$\text{var}(Y) = \phi\lambda, \tag{9.7}$$

where ϕ is a scale parameter of dispersion. It can be estimated as,

$$\hat{\phi} = \frac{\sum_{i=1}^{n} \frac{\left(y_i - \hat{\lambda}_i\right)^2}{\hat{\lambda}_i}}{n - (p+1)} = \frac{X^2}{n - (p+1)}. \tag{9.8}$$

9.1.5 Goodness of fit tests

Deviance goodness of fit

$$D_{\text{model}} = 2\sum_{i=1}^{n}\left(y_i \ln\left(\frac{y_i}{\hat{\lambda}_i}\right) - (y_i - \hat{\lambda}_i)\right), \tag{9.9}$$

where $\hat{\lambda}_i = e^{\mathbf{x}_i^T\hat{\beta}}$ is the fitted value of λ_i.

H_0 : The model is appropriate.

H_1 : The model is not appropriate.

Under H_0, $D_{\text{model}} \sim \chi^2_{1-\alpha,n-(p+1)}$ where $p+1$ is the number of parameters of the model and α is the significance level.

Pearson goodness of fit

The Pearson goodness-of-fit statistic is given by,

$$X^2 = \sum_{i=1}^{n}\frac{(y_i - \hat{\lambda}_i)^2}{\hat{\lambda}_i},$$

where $\hat{\lambda}_i = e^{\mathbf{x}_i^T\hat{\beta}}$ is the fitted value of λ_i.

H_0 : The model is appropriate.

H_1 : The model is not appropriate.

Under H_0, $X^2 \sim \chi^2_{1-\alpha,n-(p+1)}$ where $p+1$ is the number of parameters of the model and α is the significance level.

9.1.6 Health data preparation process

The geographical locations of the recorded admissions were also analysed. Only aggregated records were analysed due to data confidentiality. Statistical Areas Level 3 (SA3) was used for the visualisation as Statistical Areas Level 2 (SA2) had smaller counts. There were zero population counts in some of the SA3 areas, which were created by the Australian Bureau of Statistics (ABS) for special purposes. These areas were removed

from the analysis. Moreover, Lord Howe Island was also removed as it is far from the rest of the NSW and there were no monitoring sites that could be found there.

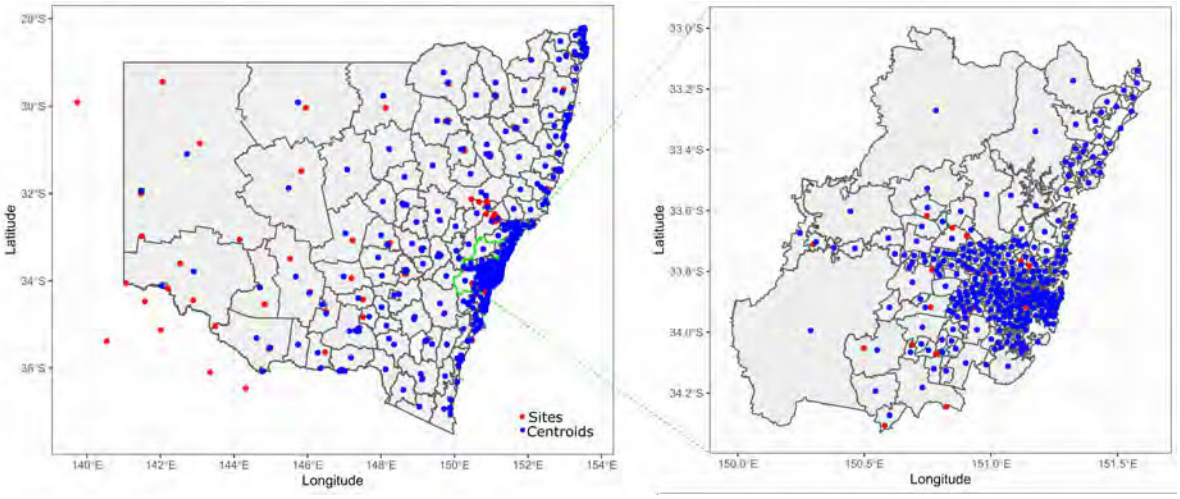


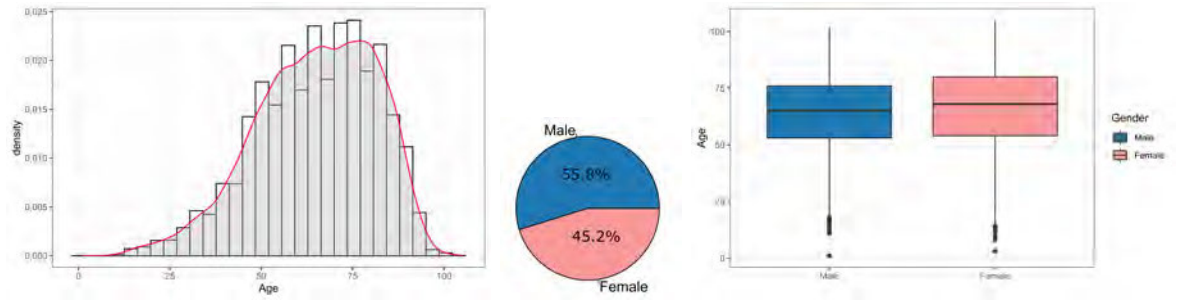
FIGURE 9.2: Locations of the monitoring sites and SA2 centroids in NSW. The enlarged area is Greater Sydney.

Figure 9.2 shows the centroids of SA2 levels and environment monitoring sites in NSW and Greater Sydney. Annual averages of the weather and pollution variables at each site were linked with the nearest SA2 centroid. An SA3 area is an aggregated form of SA2s.

9.2 Data Exploration

Figure 9.3 presents the distribution of hospital admissions of type II diabetes by age and gender during the study period. It can be seen that the age distribution of type II admissions is slightly negatively skewed (Figure 9.3a). This result is expected as type II diabetes occurs most often in middle-aged and older adults. The percentage of male admissions is slightly higher than female admissions (Figure 9.3b). There is no apparent difference between the age distribution of males and females (Figure 9.3c).

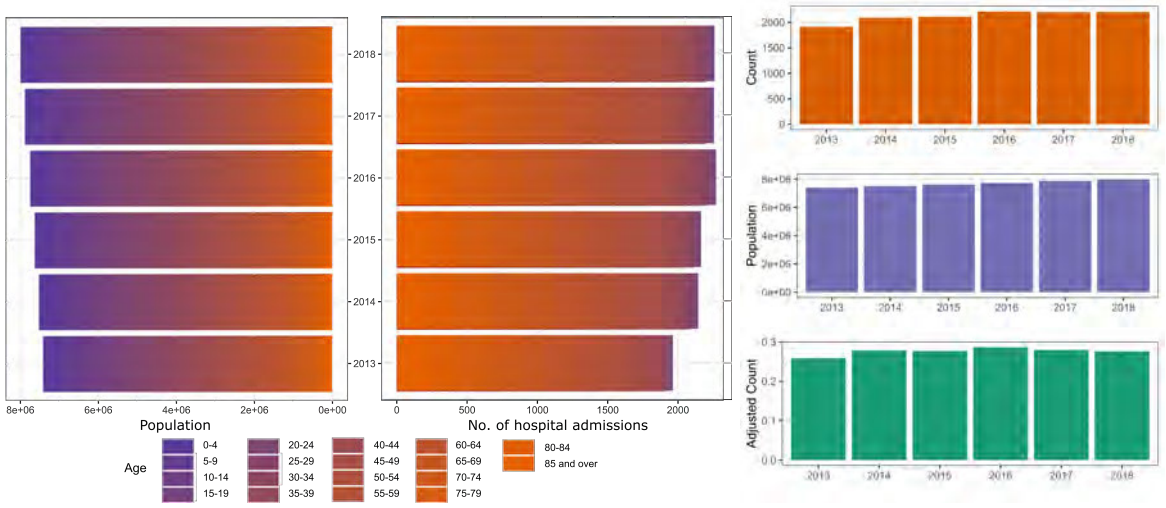
Figure 9.4a compares the number of admissions recorded each year and the estimated population of the corresponding year from 2013 to 2018. Figure 9.4b shows the number of admissions (count), population, and the number of admissions adjusted for the



(a) (b) (c)

FIGURE 9.3: (a) Age distribution of type II diabetes admissions from 2013 to 2018. (b) Age distribution of type II diabetes admissions by gender for the same period. (c) Gender distribution of type II diabetes admissions for the same period.

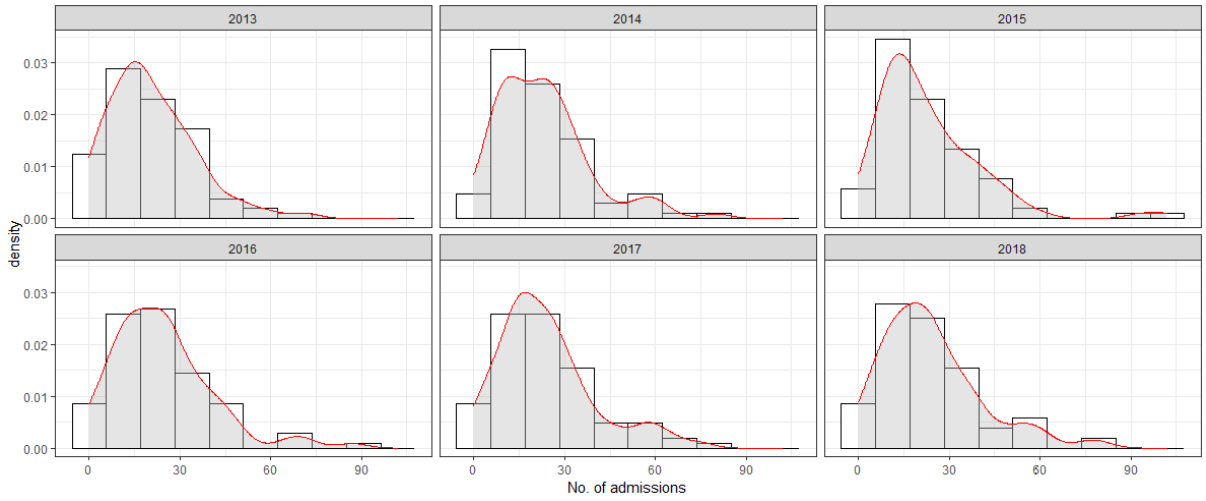
population (i.e., the number of admissions per 1000 people). There can be seen continual growth in the population and the number of admissions. Age distribution among the population shows consistency across time. The number of admissions is higher in older people than younger people in all the years. Apparently, the age distribution of the admission is also consistent across time. Population-adjusted admissions show a slight increment from 2013 to 2016, followed by a decrease from 2016 to 2018.



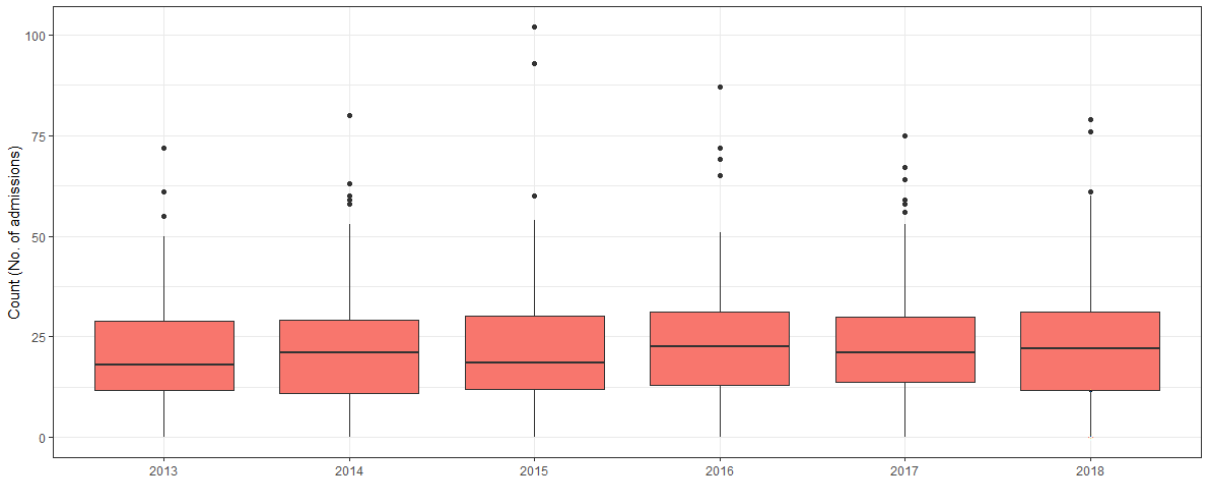
(a) (b)

FIGURE 9.4: (a) Distribution of Population and the number of Type II hospital admissions by age each year from 2013 to 2018. (b) Total number of hospital admissions (Count), Population, and the number of hospital admissions per 1000 people (Adjusted Count)

Figure 9.5a and 9.5b show the distribution of the number of admissions across the



(a)



(b)

FIGURE 9.5: (a) Distribution of the no. of admissions across the area of interest each year with an estimated density curve. (b) Distribution of the no. of admissions across the area each year displayed through box plots

area of interest (SA3 levels in NSW) in each year. The distributions are positively skewed and do not show a dramatic difference over time. Some of the years' two distinct peaks, shown in Figure 9.5a, suggest a bimodal distribution. However, further investigations are needed to verify that.

9.3 Space - Time Analysis

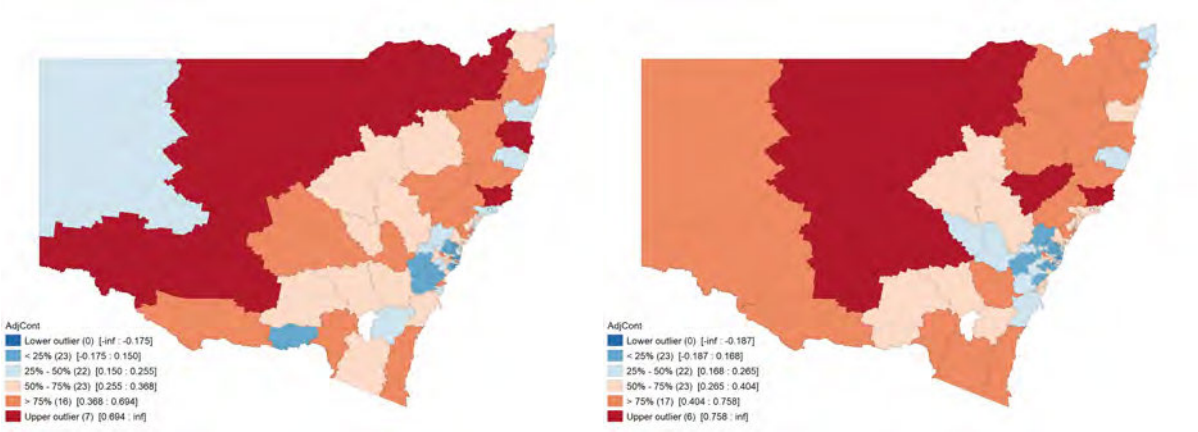
9.3.1 Analysis of diabetes rates

The distributions of the number of admissions per 1000 population were investigated using box-maps (Figure 9.6). What is interesting about all box-maps is that there are no lower outliers in any of the years. However, there are at least 3 upper outliers in all the years. In general, the diabetes rates seemed to be lower in the Greater Sydney region compared to other regions whereas diabetes rates seemed to be higher in top-middle areas of the NSW. Most occurring upper outliers were further analysed using a word cloud chart as in Figure 9.7. Bourke-Cobar-Coonamble SA3 area is the most frequently occurring outlier with high diabetes rates. Great Lakes, Inverell - Tenterfield, Lower Murray and Moree - Narrabri are also some of the areas which need attention for having higher rates.

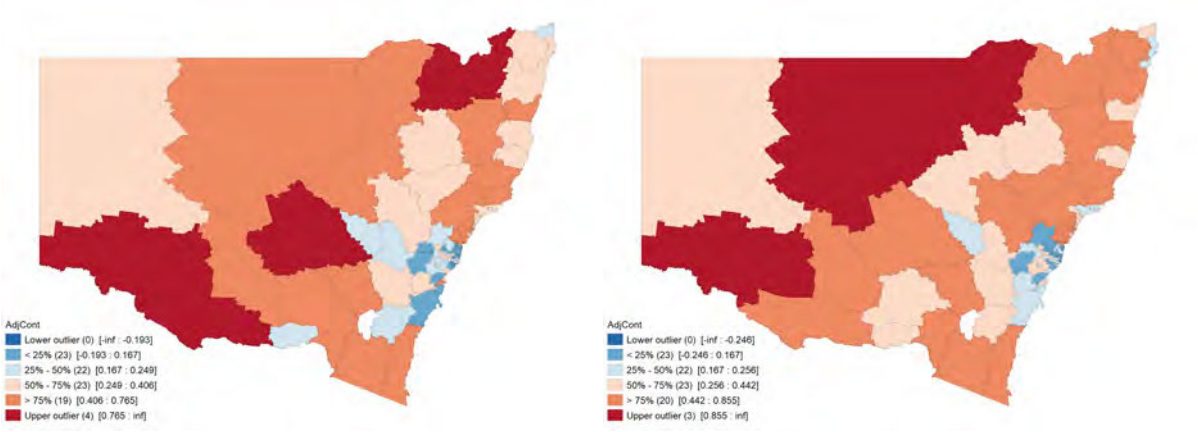
Standard deviation maps (Figure 9.8) were also created to investigate the distributions of the number of admissions per 1000 population. These maps also support the information obtained through box maps in Figure 9.6 showing only upper outliers. There are no lower outliers. The Greater Sydney region shows lower rates compared to other regions. Top-middle area of the NSW displays higher rates in the years considered. The word cloud was generated to represent the frequency of their occurrences in Figure 9.9. This also confirms that Bourke-Cobar-Coonamble SA3 area is an extreme value area which may require special attention. Further, the Inverell - Tenterfield area is also a frequently occurring outlier.

Global spatial autocorrelation of the type II diabetes rates was analysed each year and their significance was assessed through permutation approach suggested by Anselin (2006).

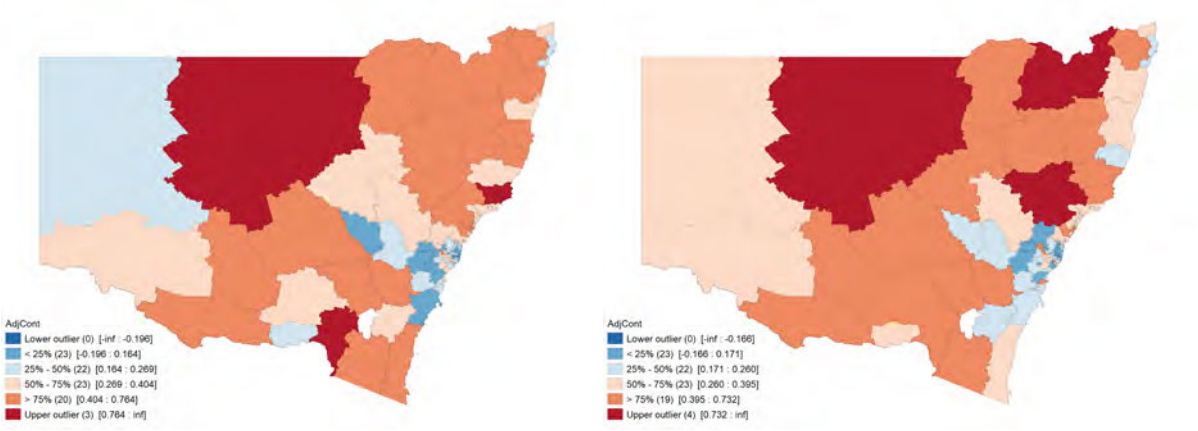
Global Moran's I statistics are positive for all the years considered as can be seen in the scatter plots in Figure 9.10 and they are significant as the pseudo p-values are very small. This indicates that there is some clustering present in the data. Local Moran's I indices were used to identify those clusters.



(a) 2013 (b) 2014



(c) 2015 (d) 2016



(e) 2017 (f) 2018

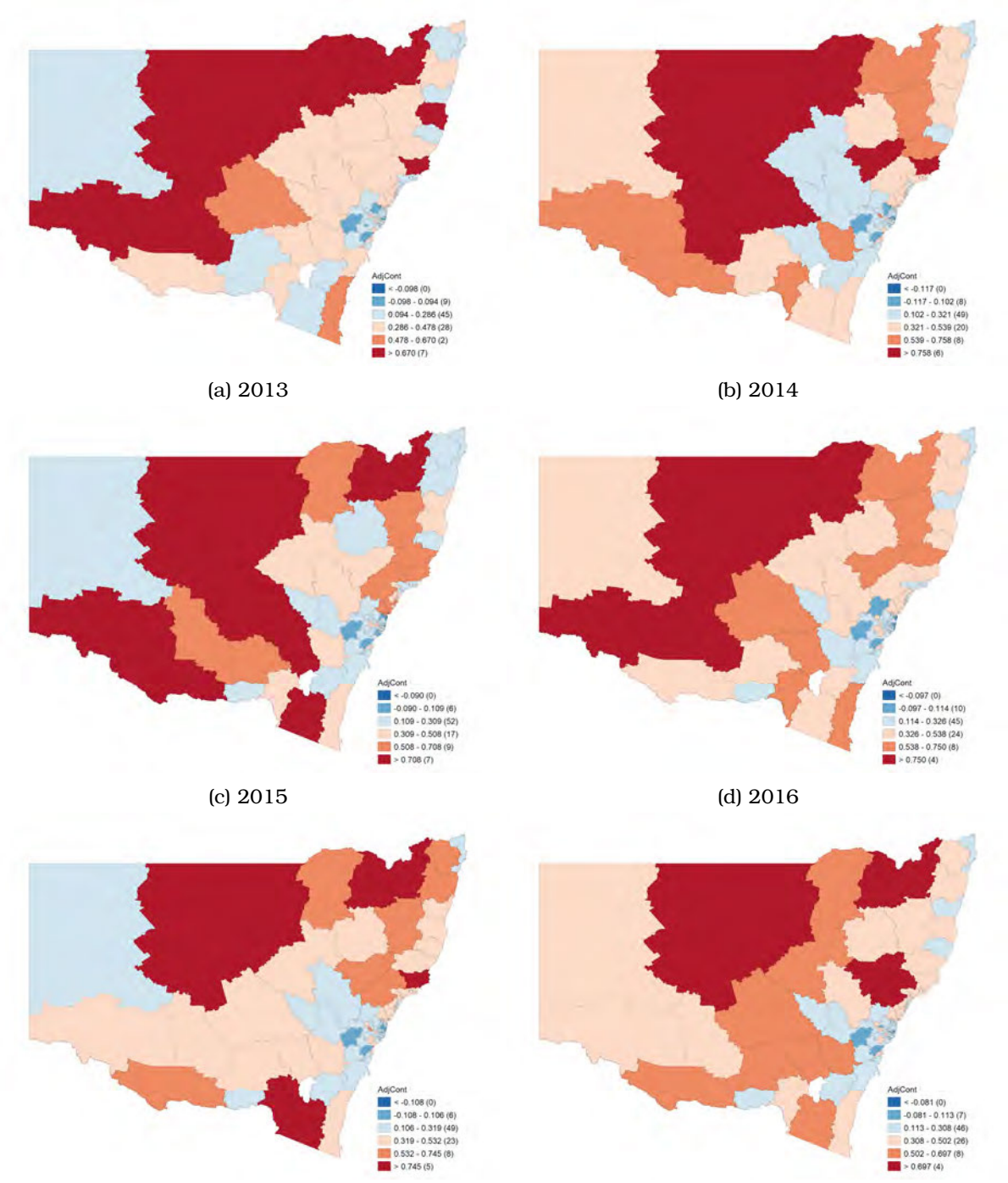
FIGURE 9.6: Box maps of the spatial distribution of population adjusted no. of admissions from 2013 - 2018 Note: Box maps group values into six fixed categories plus two outlier categories at the low and high end of the distribution

FIGURE 9.7: Word cloud representing the frequency of the SA3 upper outliers in box maps. Larger the word higher the frequency.

Figures 9.11 to 9.16 show the clusters for each year and their significance maps. It is clear that significant high-high clusters are spread in the middle and top areas of NSW except in 2017. High-high clusters indicate that their diabetes rates (number of hospital admissions per 1000) are high and their neighbour areas also indicate similar values. In contrast, low-low clusters indicate lower diabetes rates and they are scattered near the Sydney region.

9.3.2 Analysis of diabetes rates and air pollution

For each of the air pollution monitoring sites, the annual averages of the air pollution variables were calculated. Figure 9.17 shows the spatial distribution of diabetes rates along with PM2.5 annual averages. It can be seen that PM2.5 data were not available in most of the areas in NSW. According to the national standards, the maximum acceptable PM2.5 concentration for annual averages is $8\mu g/m^3$. The values less than the standard were categorised as good and the other values as bad. There is no apparent pattern between PM2.5 and diabetes. Since the available PM2.5 concentration are scattered near the Sydney region, maps of Greater Sydney were enlarged and analysed as in Figure 9.18. However, still, no clear pattern is visible. An important fact to notice here is that in most of the places with higher diabetes rates, pollution measurements were not available. This is a limitation of this study. Similarly, the spatial distribution of diabetes rates along with PM10 annual averages was also visualised for NSW and Greater Sydney as in Figure and respectively. The national standard for the annual



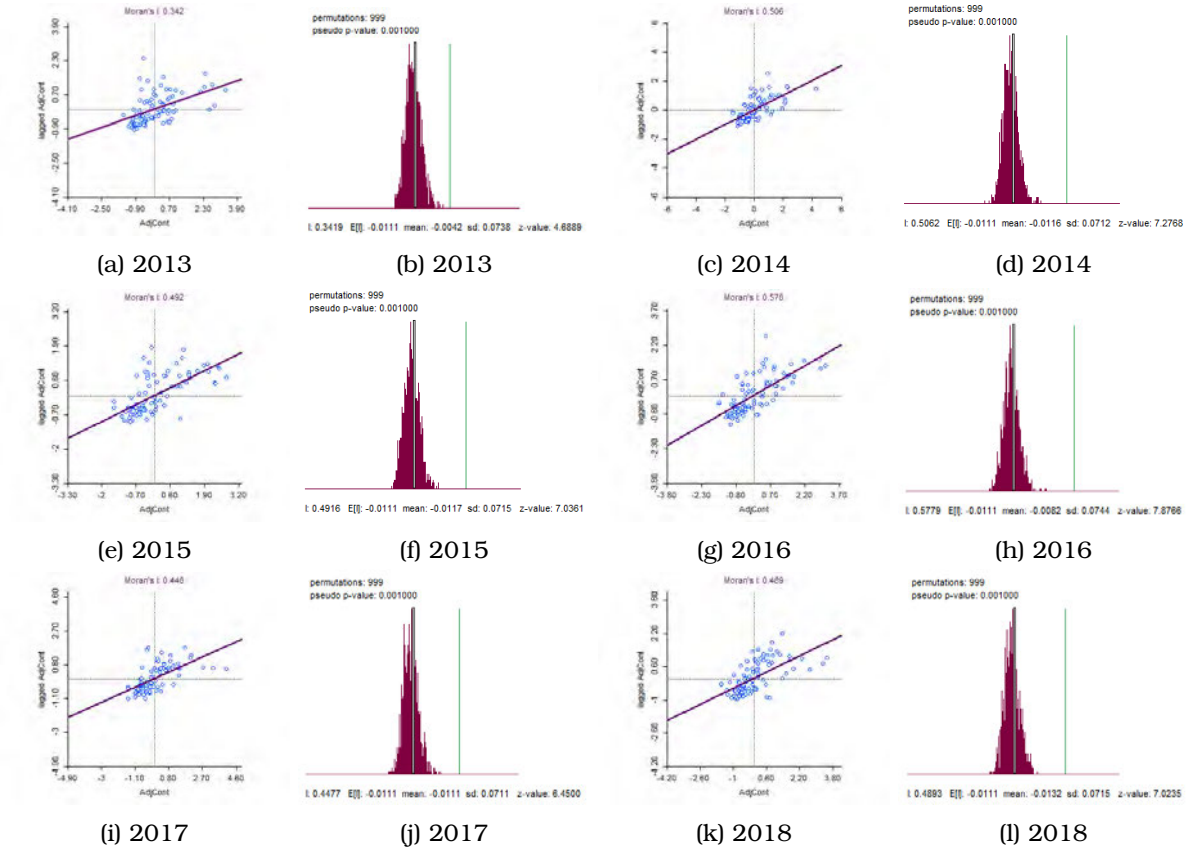
(a) 2013 (b) 2014

(c) 2015 (d) 2016

(e) 2017 (f) 2018

FIGURE 9.8: Standard deviation maps of the spatial distribution of population-adjusted no. of admissions from 2013 - 2018. Note: In Standard deviation, the variable under consideration is transformed into standard deviation units

FIGURE 9.9: Word cloud representing the frequency of the SA3 upper outliers in standard deviation maps. Larger the word higher the frequency.

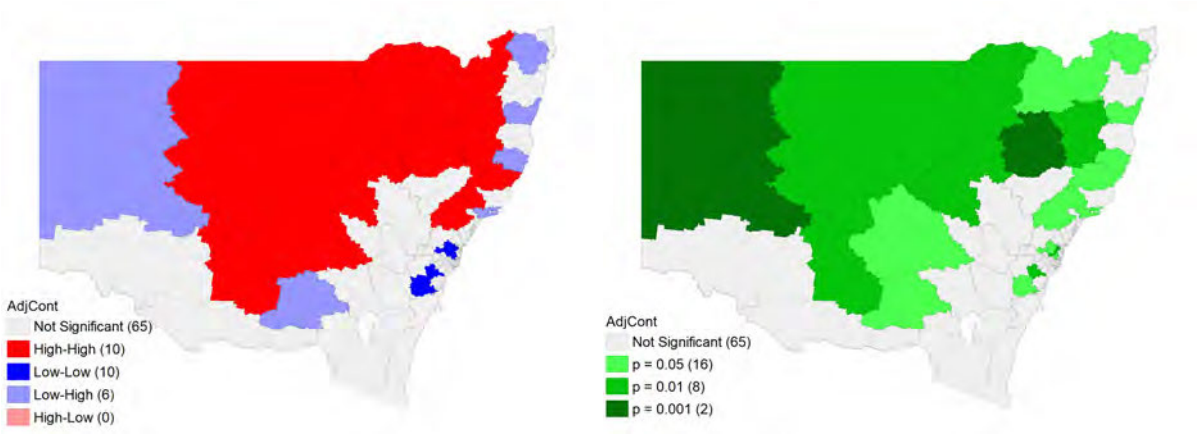


(a) 2013 (b) 2013 (c) 2014 (d) 2014

(e) 2015 (f) 2015 (g) 2016 (h) 2016

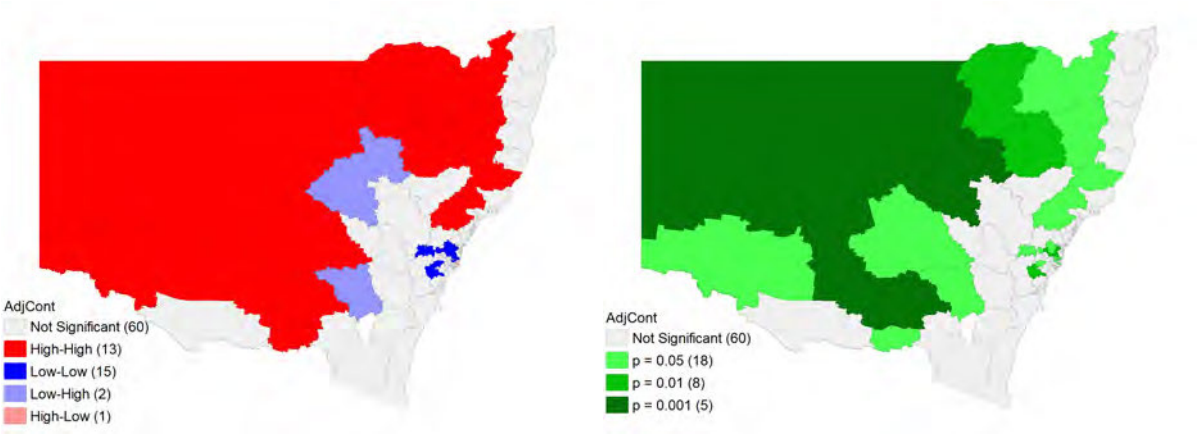
(i) 2017 (j) 2017 (k) 2018 (l) 2018

FIGURE 9.10: Moran Scatter Plots for each year and the distributions of the Moran's I statistic under the null hypothesis with green vertical line indicating the calculated test statistic



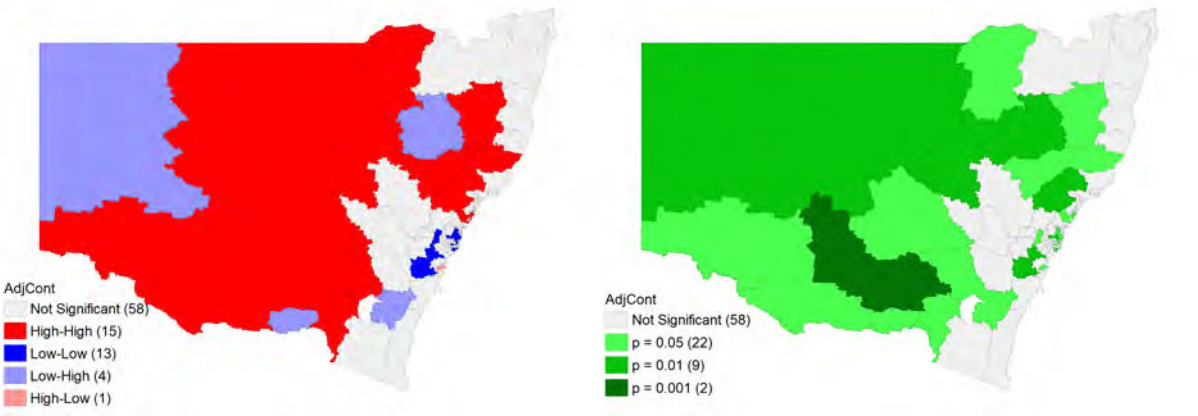
(a) Cluster map (b) Significance map

FIGURE 9.11: Spatial clusters 2013



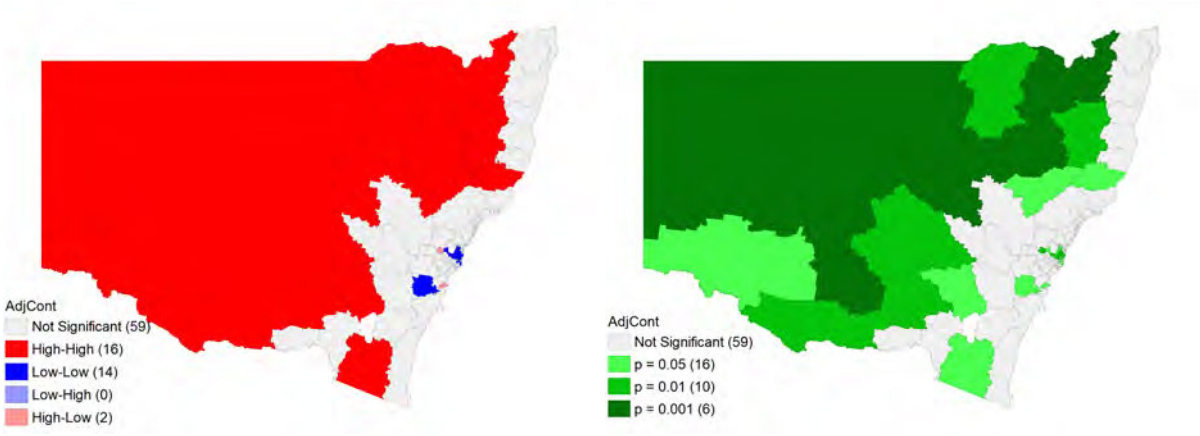
(a) Cluster map (b) Significance map

FIGURE 9.12: Spatial clusters 2014



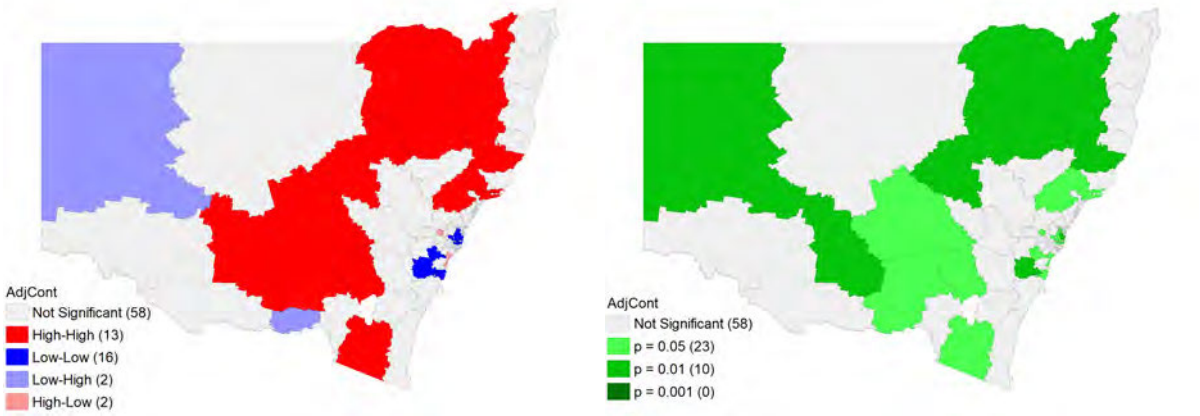
(a) Cluster map (b) Significance map

FIGURE 9.13: Spatial clusters 2015



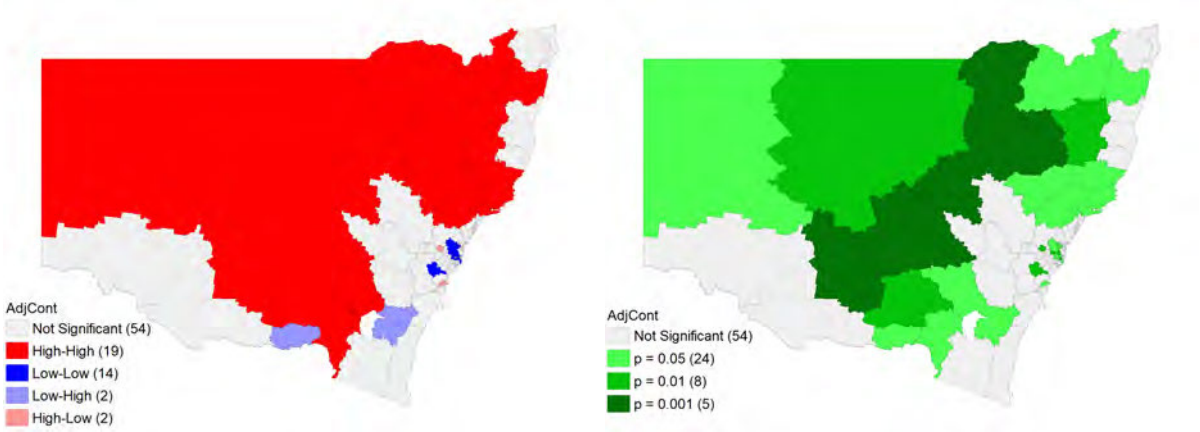
(a) Cluster map (b) Significance map

FIGURE 9.14: Spatial clusters 2016



(a) Cluster map (b) Significance map

FIGURE 9.15: Spatial clusters 2017



(a) Cluster map (b) Significance map

FIGURE 9.16: Spatial clusters 2018

PM10 averages was $25\mu g/m^3$. Again, no clear pattern between diabetes and PM10 can be found.

9.4 Poisson regression models for disease counts

Assuming that the number of hospital admissions due to type II diabetes-related issues per thousand population per year in an SA3 area is a Poisson distributed random variable and assuming that the observations are independent, the Poisson regression model was fitted to the data. Also for a Poisson distributed variable, the mean and the variance should be equal. These assumptions were checked after fitting the model. Annual averages of humidity, visibility, NO, NO2, PM2.5, PM10, SD1, S02, temperature and the average age of the population were considered as the predictors. The model was fitted to one cross-sectional dataset in 2018 considering the SA3 level areas in NSW. Due to a lack of data, some of the SA3 areas were removed from the analysis. Visualisation of the missing values of the dataset is given in Appendix B. There were 51 SA3 levels in the 2018 dataset which was the dataset with the lowest missing value proportion.

The selected SA3 levels and their spatial autocorrelation based on Moran's I were given in Figure B.1. It can be seen that the Global Moran's I (0.046) is very small and closer to zero. Also, it is not significant as the p-value is large (0.262). Correlation coefficients among the variables were calculated and given in Appendix B. None of the predictors shows a reasonable linear relationship with the disease counts or population-adjusted disease counts. However, there were some linear relationships among the predictors. Therefore, variance inflation factors (VIF) were calculated for the predictors. Poisson regression model was fitted considering the disease count as the response with the log-transformed population as an offset. The model coefficients, confidence intervals, p-values and VIF values are given in Table 9.1.

Visibility(neph), NO, NO2, Ozone, PM10, PM2.5, SO2 and temperature showed high variation inflation factors of more than 5 (Appendix B.8). Therefore, multicollinearity is present among the predictors and the estimated regression coefficients are not

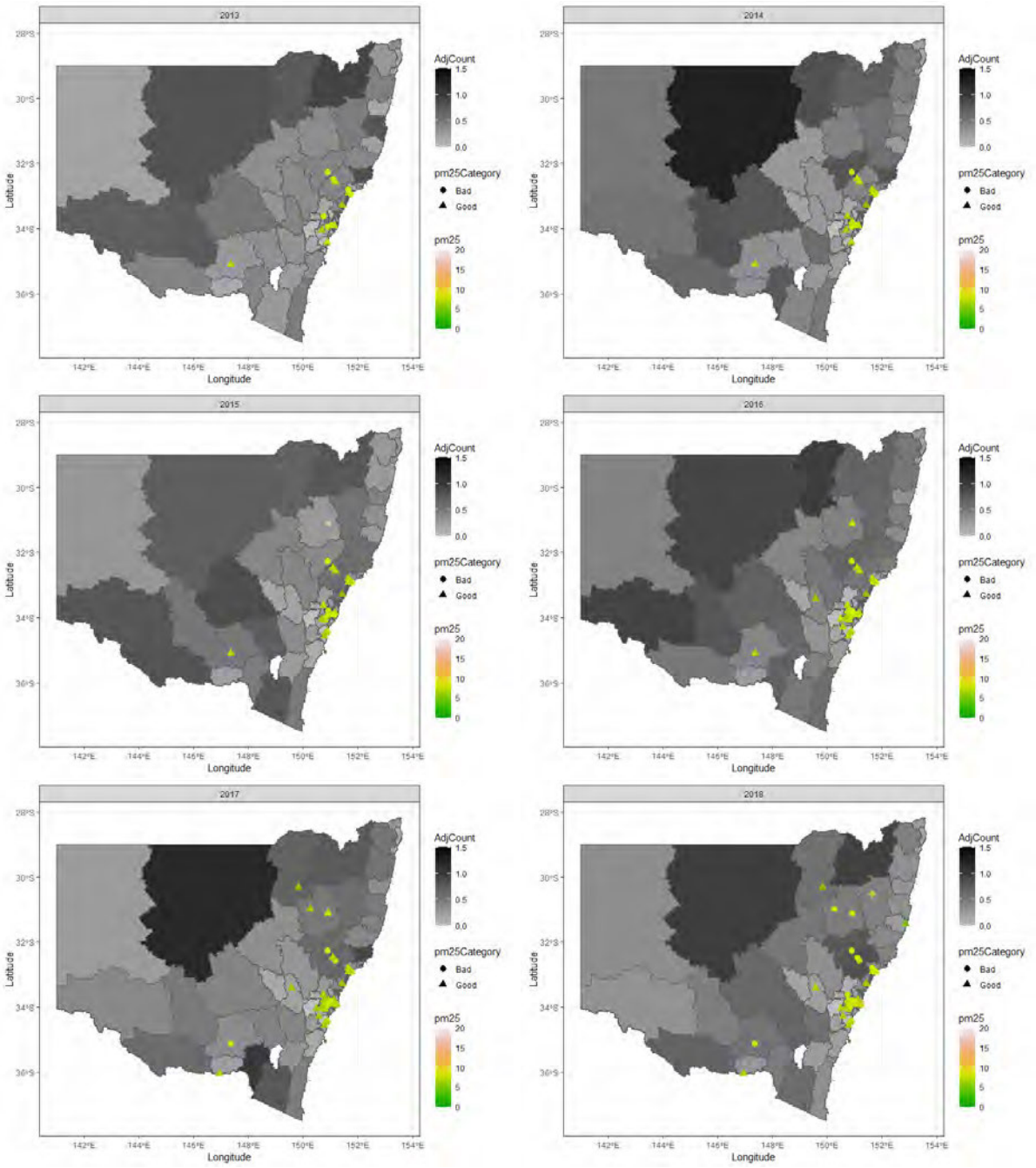


FIGURE 9.17: Annual PM 2.5 concentrations and the number of type II diabetes hospital admissions per 1000 population (Adj. count) in NSW from 2013-2018.

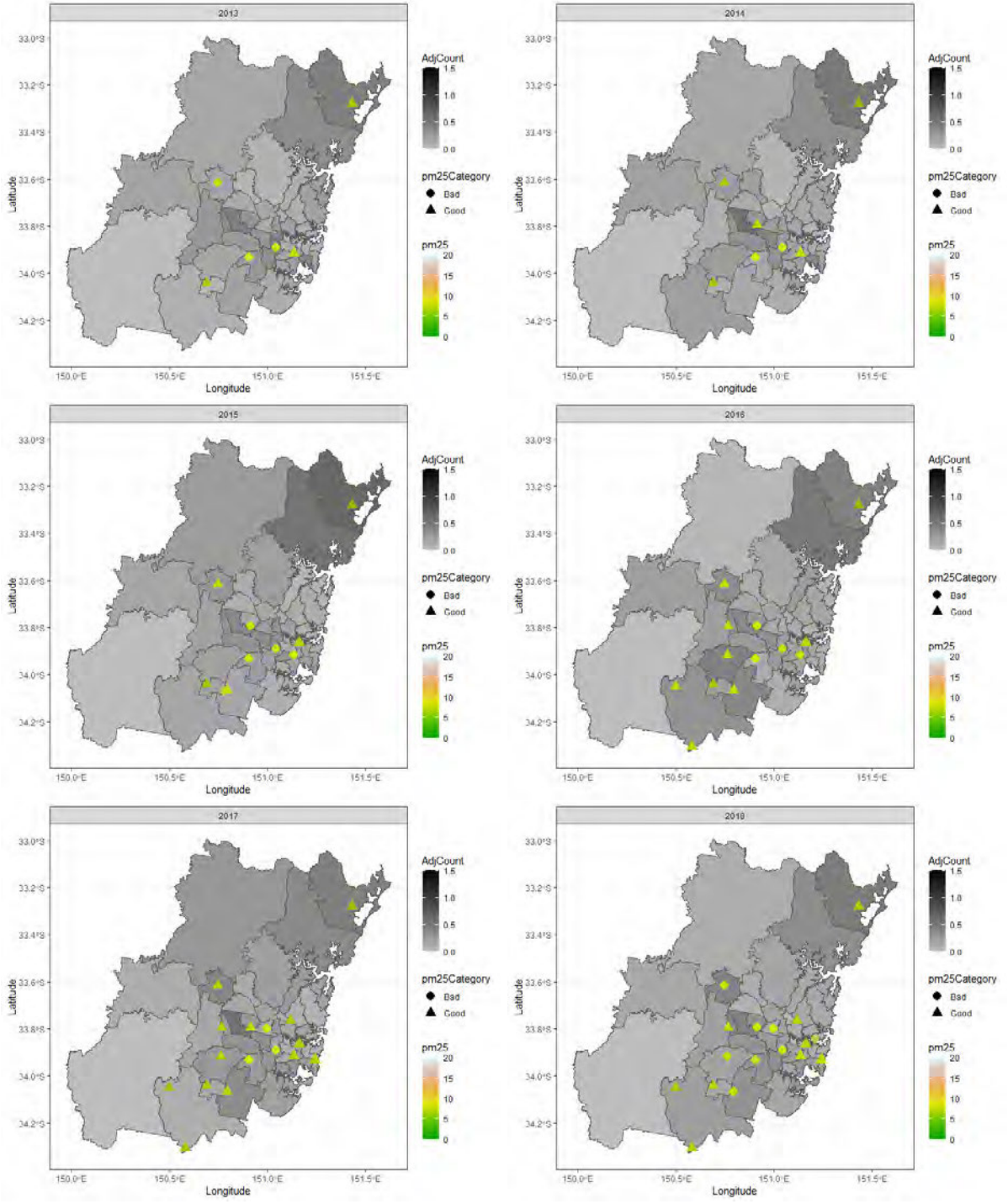


FIGURE 9.18: Annual PM 2.5 concentrations and the number of type II diabetes hospital admissions per 1000 population (Adj. count) in Greater Sydney from 2013-2018.

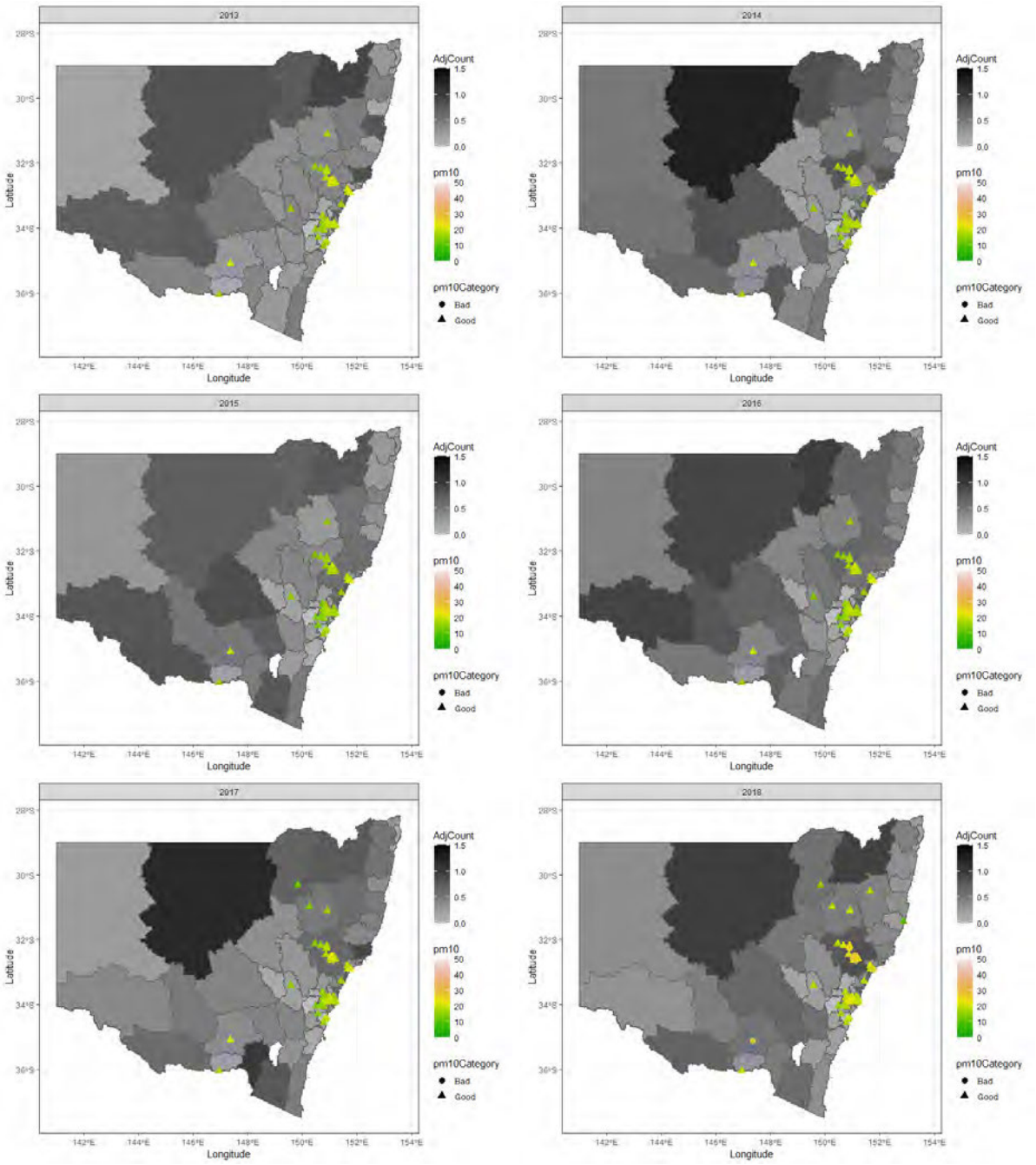


FIGURE 9.19: Annual PM 10 concentrations and the number of type II diabetes hospital admissions per 1000 population (Adj. count) in NSW from 2013-2018.

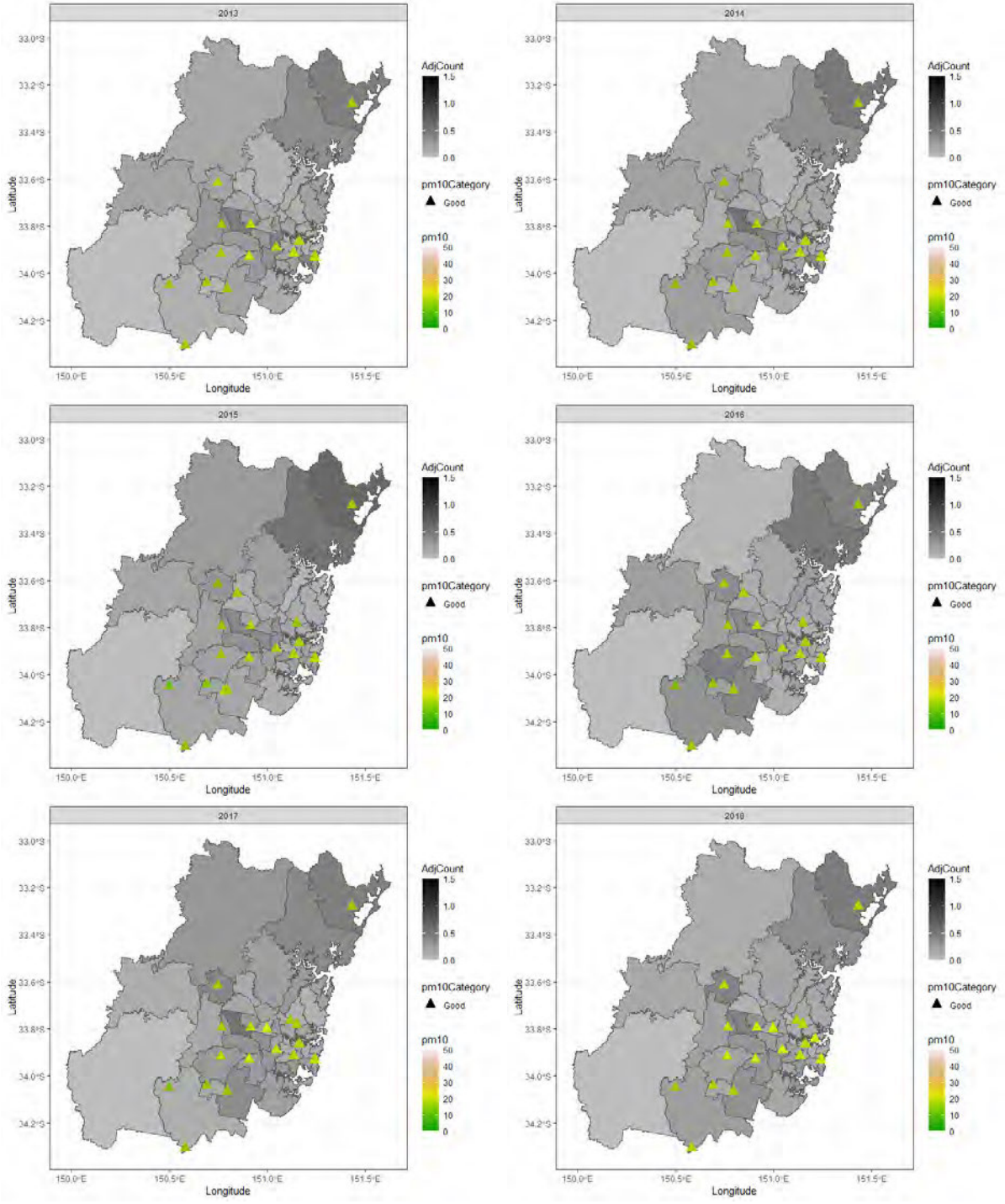


FIGURE 9.20: Annual PM 10 concentrations and the number of type II diabetes hospital admissions per 1000 population (Adj. count) in Greater Sydney from 2013-2018.

TABLE 9.1: Model output of Poisson regression model

	Estimate	Std. Error	z value	Pr(>|z|)	VIF
(Intercept)	-5.3958	2.0306	-2.66	0.0079	
AdjAge	0.0050	0.0144	0.35	0.7279	2.31
humid	0.0200	0.0135	1.48	0.1383	2.48
neph	3.1780	2.2502	1.41	0.1579	2.48
no	0.3616	0.1877	1.93	0.0541	25.68
no2	-0.6589	0.3687	-1.79	0.0739	23.83
ozone	-0.2909	0.3311	-0.88	0.3796	5.32
pm10	0.0179	0.0252	0.71	0.4780	10.35
pm25	-0.1646	0.0830	-1.98	0.0473	16.68
sd1	-0.0104	0.0060	-1.72	0.0855	3.78
so2	-0.2118	1.4253	-0.15	0.8819	8.08
tmp	-0.0382	0.0983	-0.39	0.6975	6.78

acceptable. Moreover, the dispersion parameter for the model was 0.901 indicating under-dispersion. Therefore Quasi-Poisson model was fitted for the data. Variable selection was done using both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). NO2, PM2.5, NO, Humidity, and SD1 were the best subset of variables in AIC whereas only Humidity was selected in BIC (Appendix B.11).

The output of the model with the best subset of variables in AIC is given in Table 9.2. Humidity, NO and SD1 are significant with smaller p-values less than 0.05 at a 5% level of significance). NO2 and PM2.5 are not significant and their standard error estimates are high compared to the estimated coefficients. P-value of the deviance goodness of fit test is 0.795 indicating the model is a good fit. P-value of the Pearson goodness of fit test is 0.725 again confirming a good fit. Model diagnostic plots were checked and an influential point with Cook's distance 1.2 was found (refer B.12). After removing that observation, the model was fitted again and the diagnostic plots were investigated. Then the P-values of the deviance goodness of fit and Pearson goodness fit test 0.886 and 0.839 were respectively. The model coefficients and the standard errors are given in Table 9.3. Except for Humidity, all the other variables were significant. However, it is interesting that NO has a positive effect whereas NO2 and PM2.5 and SD1 have negative effects on the disease counts. Model diagnostic plots were better after removing the influential points (refer B.12). After removing the humidity variable,

all the remaining variables were significant (refer B.14).

The model selected using AIC is better suited for prediction while BIC is better suited for explanation (Shmueli, 2010). Since the main focus is to understand the effect of air pollution and weather on diseases, the model with the best subset of variables using BIC was further analysed (Table 9.4). It can be seen that the p-value is very small indicating a Humidity is significant. The standard error of the coefficient is also low compared to the estimated coefficients. Coefficient estimate of the Humidity is 0.0332 and the Incident rate ratio is 1.03 (95% C.I. 1.02-1.05). This gives that the change in the mean response for one unit increase of the humidity level is 1.03. In other words, the expected number of type II diabetes hospital admissions will increase by 3% per unit increase in the humidity. Deviance and Person goodness of fit test p-values are 0.529 and 0.464 respectively. However, only 23% of the deviance is explained by the model which is very poor.

Table 9.2: Model output of Quasi-Poisson regression model with the best subset of variables using AIC

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-5.9651	0.8645	-6.90	0.0000
humid	0.0201	0.0089	2.25	0.0292
no	0.2823	0.1246	2.27	0.0283
no2	-0.3406	0.2196	-1.55	0.1279
pm25	-0.0911	0.0546	-1.67	0.1017
sd1	-0.0128	0.0045	-2.84	0.0067

Table 9.3: Model output of Quasi-Poisson regression model with the best subset of variables using AIC (after removing influential point)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.8779	0.9349	-5.22	0.0000
humid	0.0086	0.0097	0.89	0.3794
no	0.3969	0.1271	3.12	0.0032
no2	-0.5448	0.2258	-2.41	0.0200
pm25	-0.1199	0.0534	-2.25	0.0297
sd1	-0.0123	0.0043	-2.87	0.0063

TABLE 9.4: Model output of Quasi-Poisson regression model with the best subset of variables using BIC (after removing influential point)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-8.1190	0.6060	-13.40	0.0000
humid	0.0332	0.0089	3.73	0.0005

9.5 Conclusion

Based on the spatial exploratory analysis, it was found that the top middle area (Bourke-Cobar-Coonamble and Inverell-Tenterfield) of NSW exhibits high rates of hospital admissions for type II diabetes. Unfortunately, these areas were excluded from the modelling process due to the unavailability of weather and pollution data. It is advisable to prioritize these regions when considering the placement of new sensors, enabling future studies to be conducted in these areas. Additionally, the analysis revealed a positive correlation between humidity and type II diabetes hospital admissions in NSW; however, it is important to note that correlation does not imply causation.

Chapter 10

Discussion and Conclusions

This thesis has two-fold aims; to improve data cleaning methodology in preparing environmental data for the downstream analysis and to identify associations of environmental variables in diabetes. Chapters 4 -6 in the thesis address the first objective of data cleaning methodology. Chapter 7 presented an application of the data cleaning process and further data exploration to identify air pollution patterns. Chapter 8 includes a case study assessing the applicability of machine learning algorithms to predict health outcomes using environmental variables. Finally, Chapter 9 investigates the associations of diabetes with environmental factors.

10.1 Summary of the findings and their implications

The main contribution of Chapter 4 is the development of a data-cleaning framework for air pollution and weather data connecting some of the existing algorithms. This is available to use by any researcher who uses Sydney air pollution and weather data. Through the graphical user interface, one can apply existing methods without having to learn R codes. It also provides a facility for the visualisation of missing data patterns so one could analyse those things before deciding on a method. Moreover, after applying a method, one could visually as well as statistically compare the performances.

Chapter 5, first, discussed six well-established methods of dealing with missing values in a univariate time series context and compare their performance on imputing missing values for air quality data in the Sydney region. The methods discussed

here are Mean Imputation, Spline Interpolation, Simple Moving Average, Exponentially Weighted Moving Average, Kalman Smoothing on Structural Time Series Models and Kalman Smoothing on ARIMA models. The performances of these methods were compared with three performance measures; Mean Squared Error (MSE), Coefficient of Determination (R2) and Index of Agreement (d). Among the six methods considered, Kalman Smoothing Method on Structural Time Series is the best method for imputing missing values in the context of air quality data where the missing mechanism is MCAR.

Chapter 5, next, proposed a novel algorithm (FBReg) to impute univariate air pollution datasets using a bidirectional regularised regression model. The proposed method was evaluated against two baseline models namely Mean imputation and LOCF as well as two well-established imputation methods namely Kalman smoothing and AutoARIMA. The evaluation was done under the MCAR mechanism with exponentially distributed missing values and with varying consecutive gap sizes. FBreg outperforms all the other selected methods regardless of the percentage of missing values and the gap sizes in the series.

Moreover, Chapter 5 proposed another algorithm (DesReg) extending the FBReg algorithm to seasonal variables to deal with missing values. It was evaluated using hourly temperature data. AutoARIMA and Kalman methods were the best methods to impute missing values under the MCAR mechanism with exponential distribution. However, their performances are not consistent in other missing situations. Especially, the Kalman method performs poorly when there are large gaps. DesReg was the best method to deal with large missing gaps. DesReg could be recommended to impute large gaps in univariate time series which show seasonal patterns As stated in the Introduction Chapter, the problem of missing values in environmental sensor data is ubiquitous. Technological solutions may be implemented to avoid this problem in the future. However, it is vital to make the most use of available data for many purposes, including assessing the environmental impact on health outcomes.

Chapter 6, further improved the missing data imputation algorithm by incorporating

the information from correlated variables. The proposed method performs reasonably well in recovering large gaps as it incorporates both the time series characteristics and the information on other correlated variables to recover data. This algorithm can be recommended especially for meteorological datasets, as large gaps and correlated variables are ubiquitous in these data. This method is not applicable in univariate situations.

Chapter 7, presented an application of cleaned environmental data to identify clusters in environmental monitoring sites. DTW distance-based clustering was applied in this task. Chapter 8 presented an application of machine learning methods in predicting the risk of asthma based on air pollution and weather. The objective was to identify classification models which could be used to predict vulnerability and risk of getting future episodes of asthma based on weather and pollution conditions. Random forest gave the best accuracy among the considered machine learning methods. The methods applied in this chapter may apply to other diseases associated with environmental factors. In future work, these methods could be applied in the categorization of individuals who may be at risk of developing diabetes or areas at high risk using environmental conditions. Also, the analysis of this chapter used readily available integrated data related to a hospital in Victoria State to demonstrate the methodology and to highlight the value of data integration. The future direction is to carry out this to diabetes and in NSW State.

According to the spatial exploratory analysis in Chapter 9, the top middle area of NSW, more specifically Bourke-Cobar-Coonamble and Inverell-Tenterfield were identified as areas with high type II diabetes hospital admission rates. However, these areas were not included in the modelling due to the lack of weather and pollution records in these areas. It is recommended to pay more attention to these areas when placing new sensors so that further studies can be carried out in these areas. Moreover, according to Chapter 9, humidity showed a positive relationship with type II diabetes hospital admissions in NSW. However, the quantification of this relationship is questionable due to many reasons. This is because there may be several confounding factors that have not been investigated in this study. For example, type II diabetes is influenced

by people's lifestyles and food consumption patterns. However, the focus of this study was not to measure the effectiveness of those factors in developing diabetes. Instead, the focus was to identify whether there is an association between environmental factors in the number of hospital admissions due to diabetes-related issues. However, further analysis is required to conclude the effect of environmental factors on type II diabetes. For example, analysis of individual-level exposure to environmental factors and checking the associations of incidence and prevalence of diabetes may reveal the causal relationships. There may be people with diabetes hospitalised due to other complications. Since the analysis was done using NSW-admitted patient data collection categorised as diabetes, these people may not include in the analysis.

10.2 Limitations and future research

One of the major disadvantages of this study was the lack of pollution and weather data in most of the areas in NSW. The proposed methodologies throughout this thesis provided remedies for some of the situations. However, the problem still exists. This study used the data obtained through the NSW air quality monitoring network governed by the Department of Planning and Environment. However, not all the variables are measured in the sites and some of the sites were not functioning in some of the years. In addition to these data, low-cost sensors monitoring the air quality at council levels also could be used to reduce the problem of lack of data. Moreover, missing value-filling approaches could also be extended incorporating both space and time. Another limitation of this study is taking the annual averages to measure the long-term exposure to environmental variables. Research can be extended to identify more appropriate measures to represent the exposure. Moreover, an individual's exposure to the environment was obtained by considering the individual's residential SA2 level and linking its centroid to the nearest monitoring site. This may not represent their accurate exposure.

Diabetes data for this study was obtained from the NSW admitted patient data collection only. NSW emergency patient data collection can also be used to get more insights

into type II diabetes-related issues. Moreover, the focus of this study was given to type II diabetes patients only. Type I diabetes patients also can be analysed in the future. Moreover, if the patient-level data can be used with a measure of individual exposure to the environmental variables, a more comprehensive analysis could be carried out.

The problem of ecological fallacy is present in this study as the characteristics of a population as a whole are investigated and the individual-level analysis was not carried out. However, due to data limitations and ethical issues, individual-level analysis was not possible. This analysis was carried out using 2016 SA3-level aggregated data. However, different geographical areal units may apply to different purposes. For example, if some decisions are to be made based on local government areas (LGA), LGA-wise analysis would be helpful. Both the size and spatial arrangement matter for aggregated data. Also, sometimes spatial boundaries may change. Therefore, attention should be given to which areal unit is considered for the analysis when making decisions.

This study followed an inductive approach starting from the data to determine the associations of environmental factors on diabetes. To understand the causal relationships between diabetes and the environment, deductive approaches may be carried out in the future starting from the hypothesis that there is a relationship between diabetes and the environment and collecting data accordingly. Since this study mainly followed CRISP-DM (Cross Industry Standard Process for Data Mining), the findings of the research could be used to update the knowledge of the problem understanding itself. These results could be used to design causal studies effectively and efficiently. In general, the future direction is to identify environmental profiles relating to any non-communicable diseases such as diabetes, cataract etc. Further, the studies can be extended to other states and countries.

10.3 Conclusions

The existing data imputation methods show limited effectiveness in filling large gaps. However, the algorithms proposed in this thesis are specifically designed to address this issue and can successfully handle large gaps (if daily data, up to 100 missing

values) of air pollution and weather data for downstream analysis. Despite these advancements, there is still room for improvement in solving the missing data problem related to environmental variables. The data cleaning framework and algorithms in this thesis have wider applicability beyond the specific context of air pollution and weather data. They can be implemented in various research areas involving similar data requirements. The quasi-poison regression model applied to the type II diabetes hospital admission data showed a positive relationship between the number of hospital admissions and the humidity level of the environment. Further research is needed to make causal inferences. Overall, the methodologies presented in this thesis contribute to the understanding and management of missing data in environmental variables. Moreover, the findings of this research hold significant value for policymakers offering valuable insights for the effective management and reduction of the impact of diabetes.

Appendix A

Exploratory Analysis of the missing values of air pollution and weather

Figures below show the distributions of missing values over time from 1994.01.01 to 31.12.2018 for each variable.

Figures below show the distributions of missing values for each variable for site from 1994.01.01 to 31.12.2018.

FIGURE A.1: Distribution of PM2.5 missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

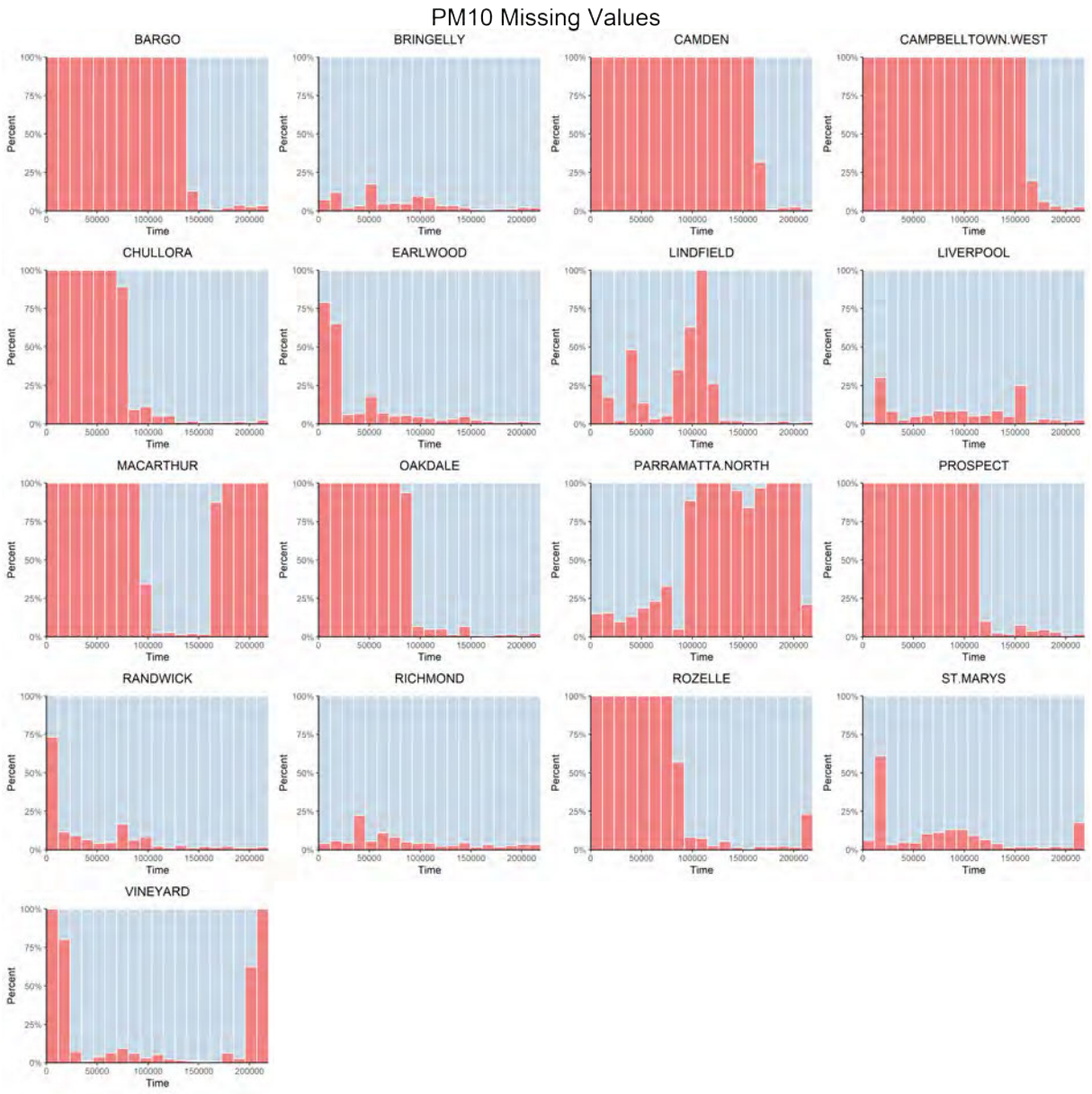


FIGURE A.2: Distribution of PM10 missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

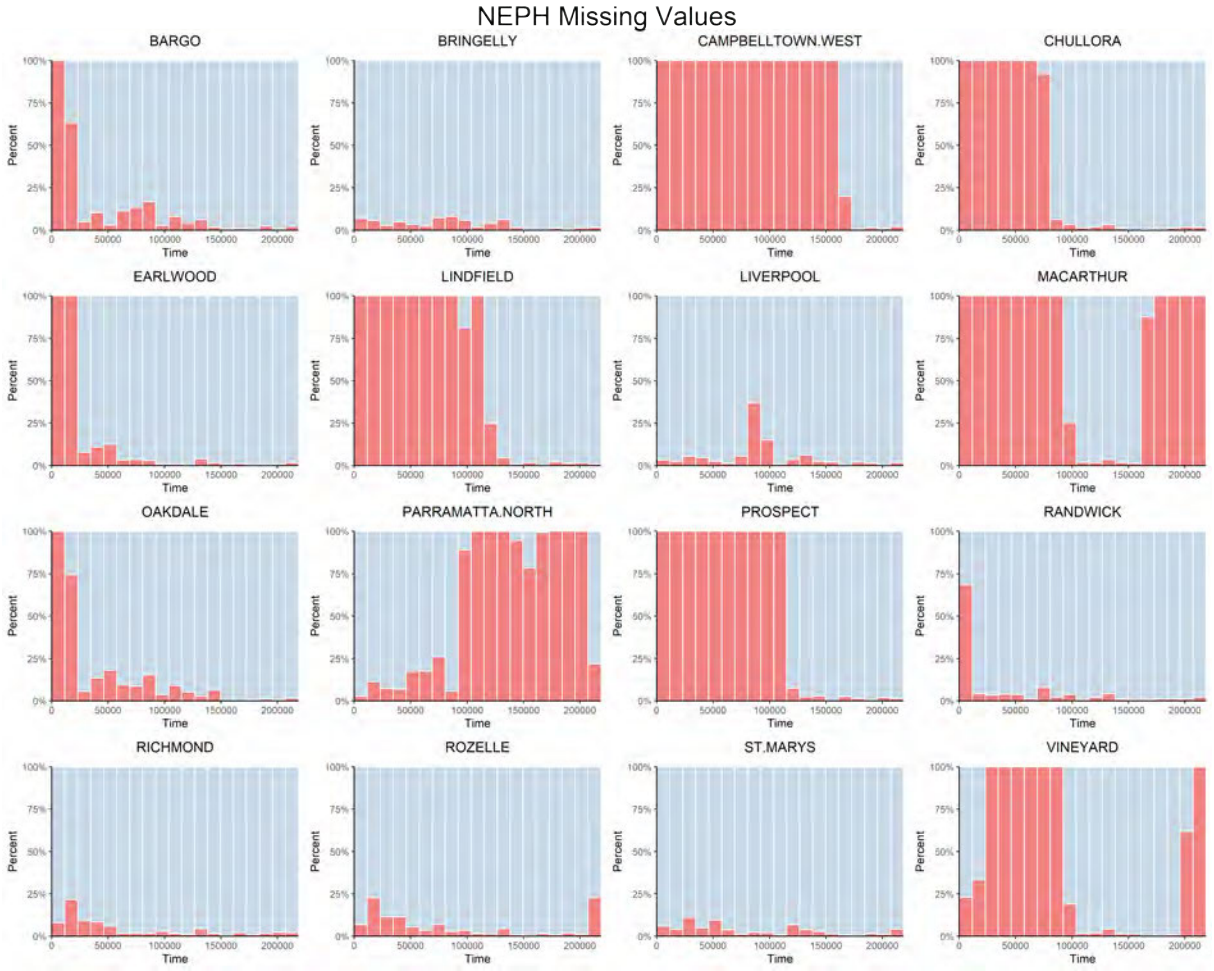


FIGURE A.3: Distribution of NEPH missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

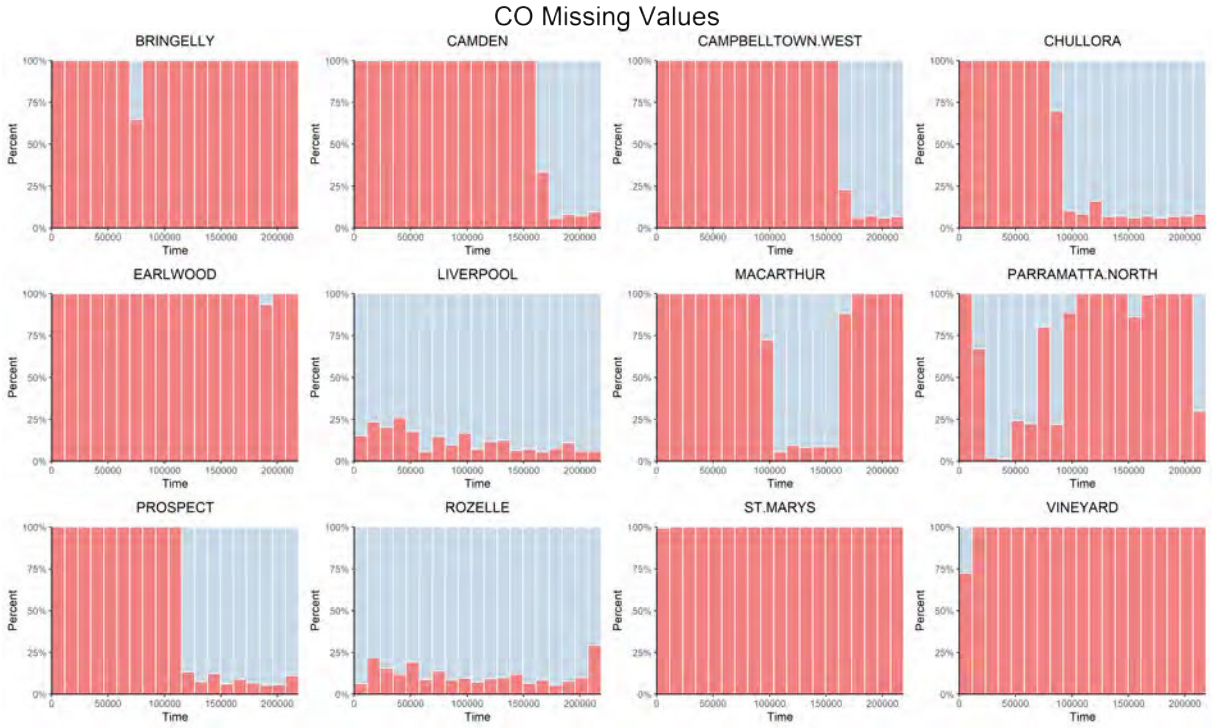


FIGURE A.4: Distribution of CO missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

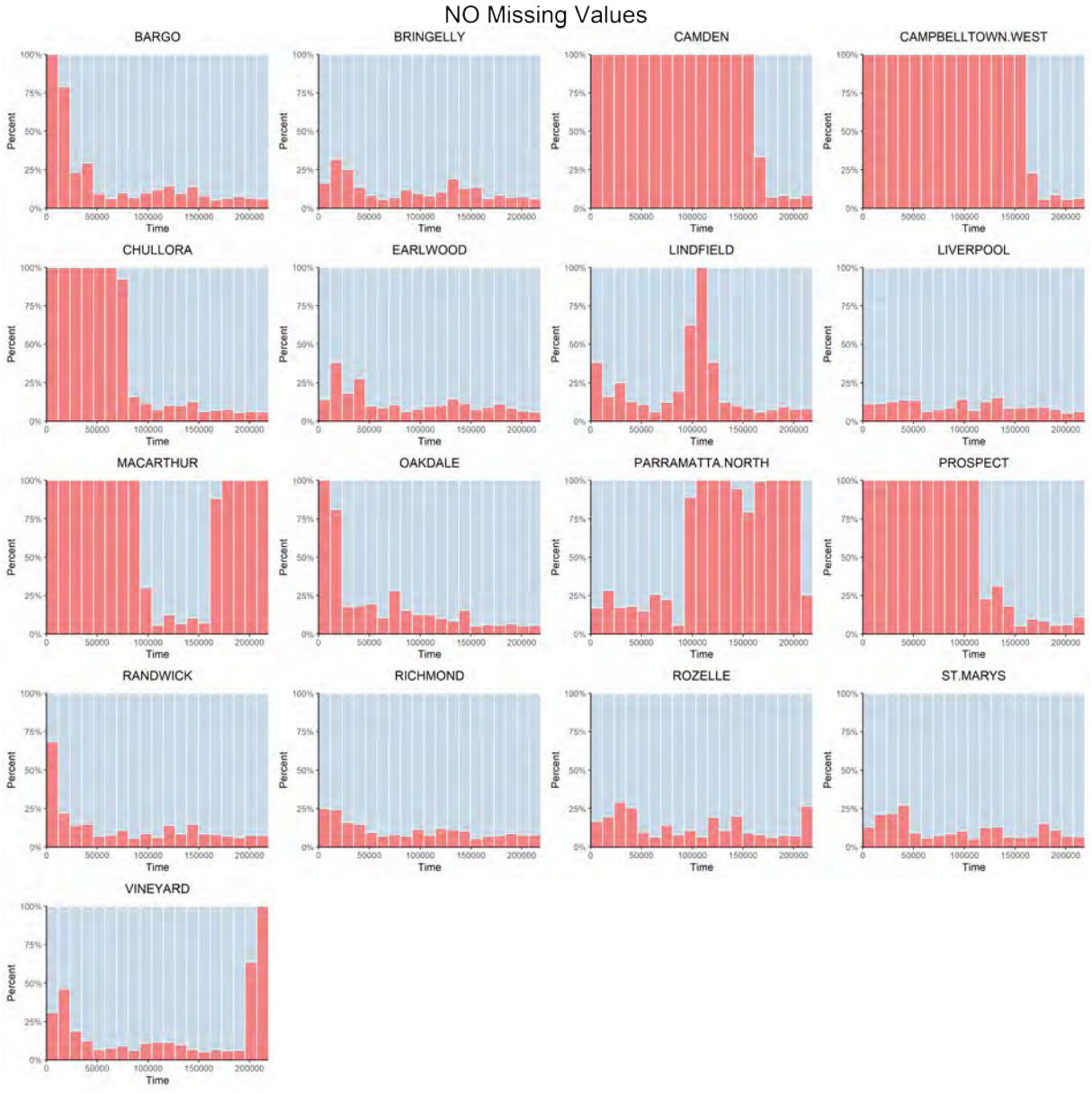


FIGURE A.5: Distribution of NO missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

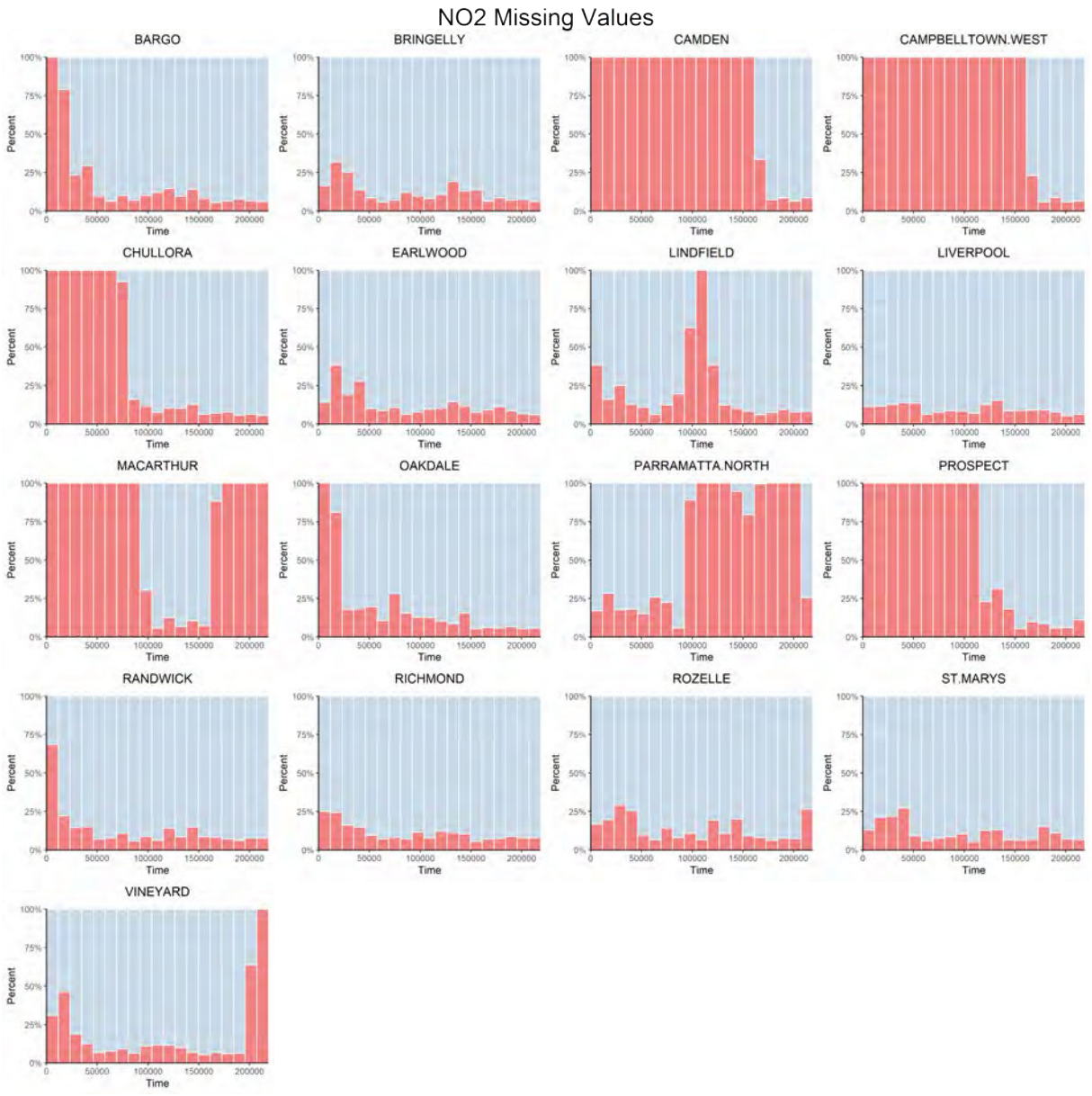


FIGURE A.6: Distribution of NO2 missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

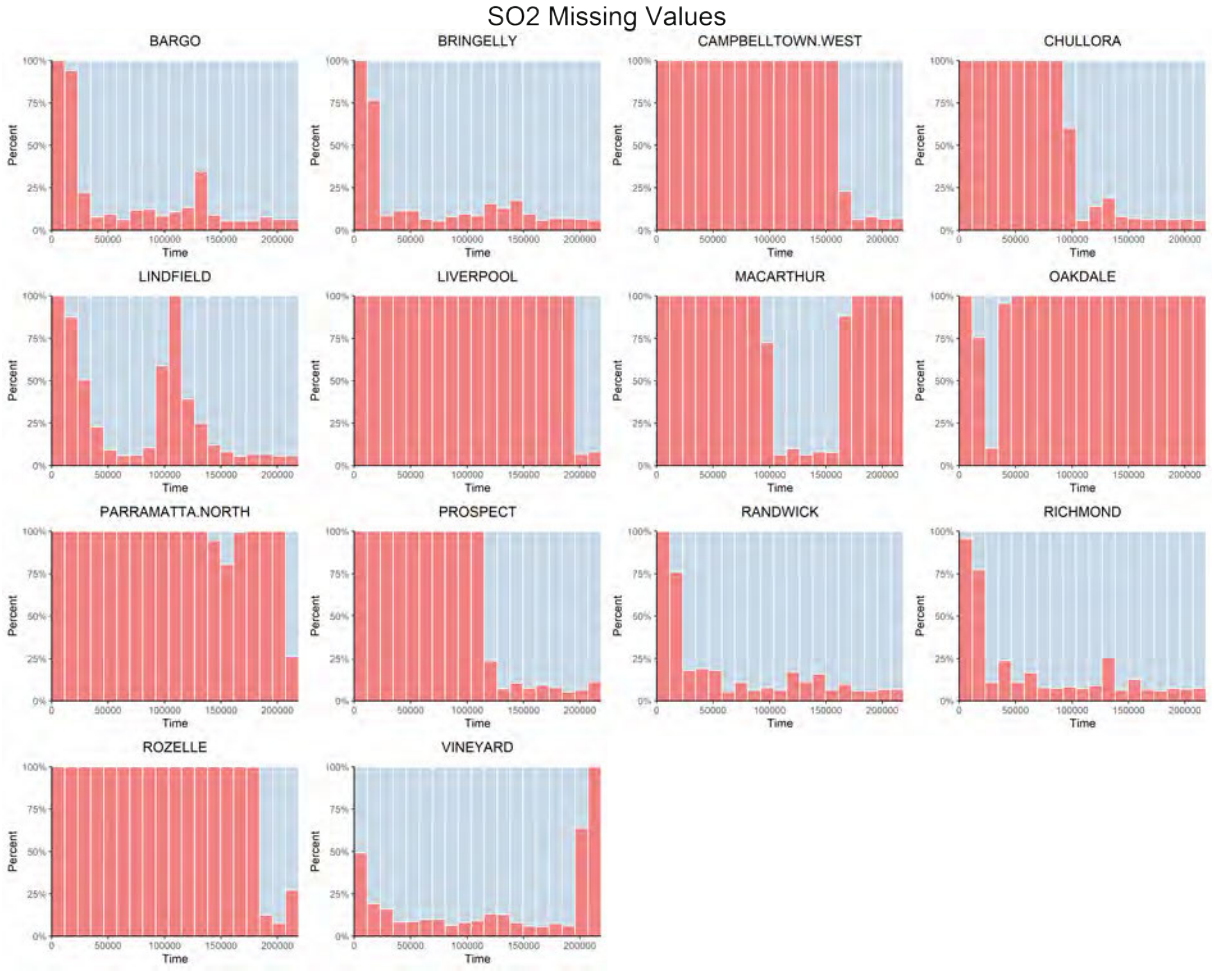


FIGURE A.7: Distribution of SO2 missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

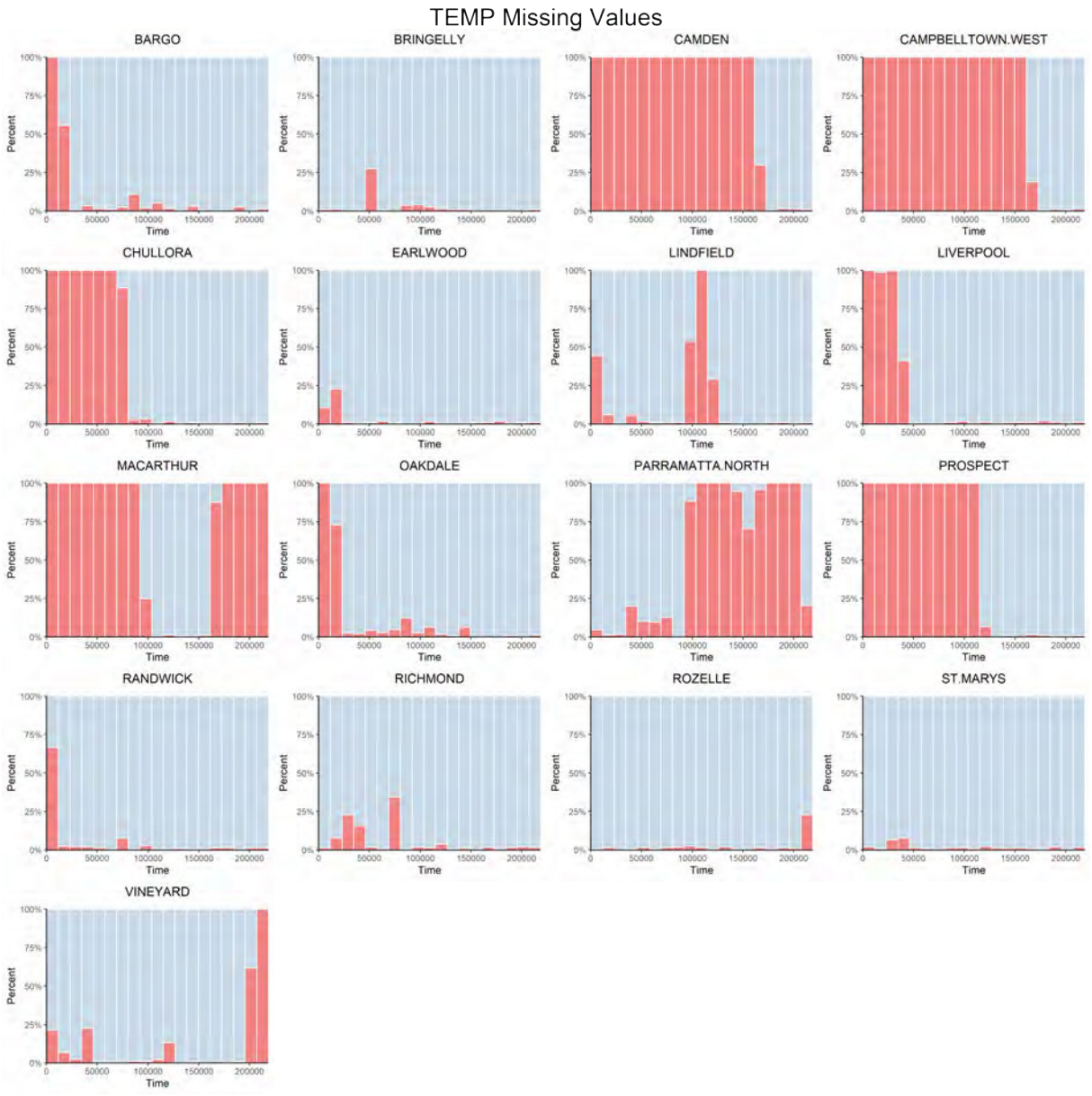


FIGURE A.8: Distribution of TEMP missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

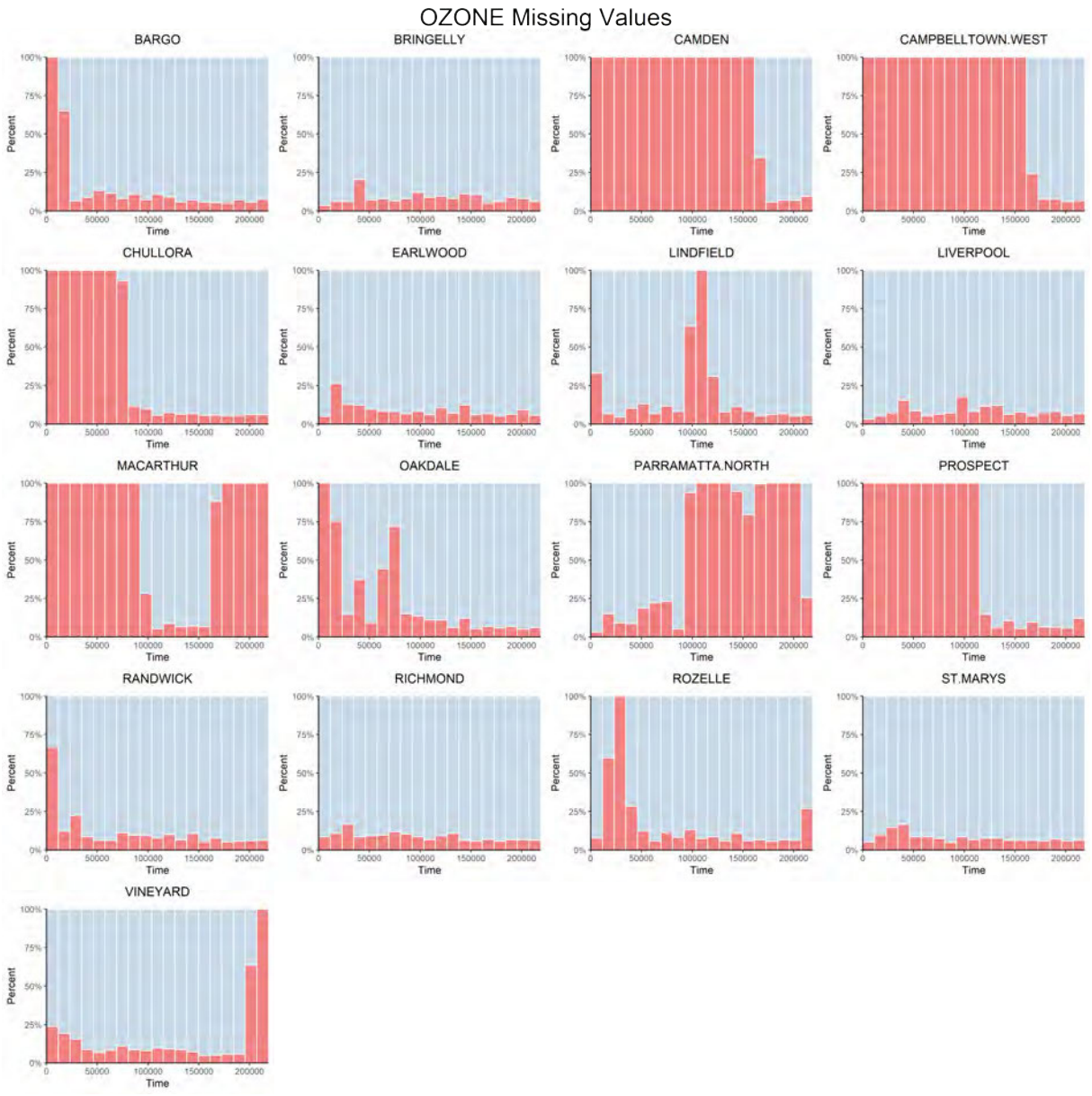


FIGURE A.9: Distribution of OZONE missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

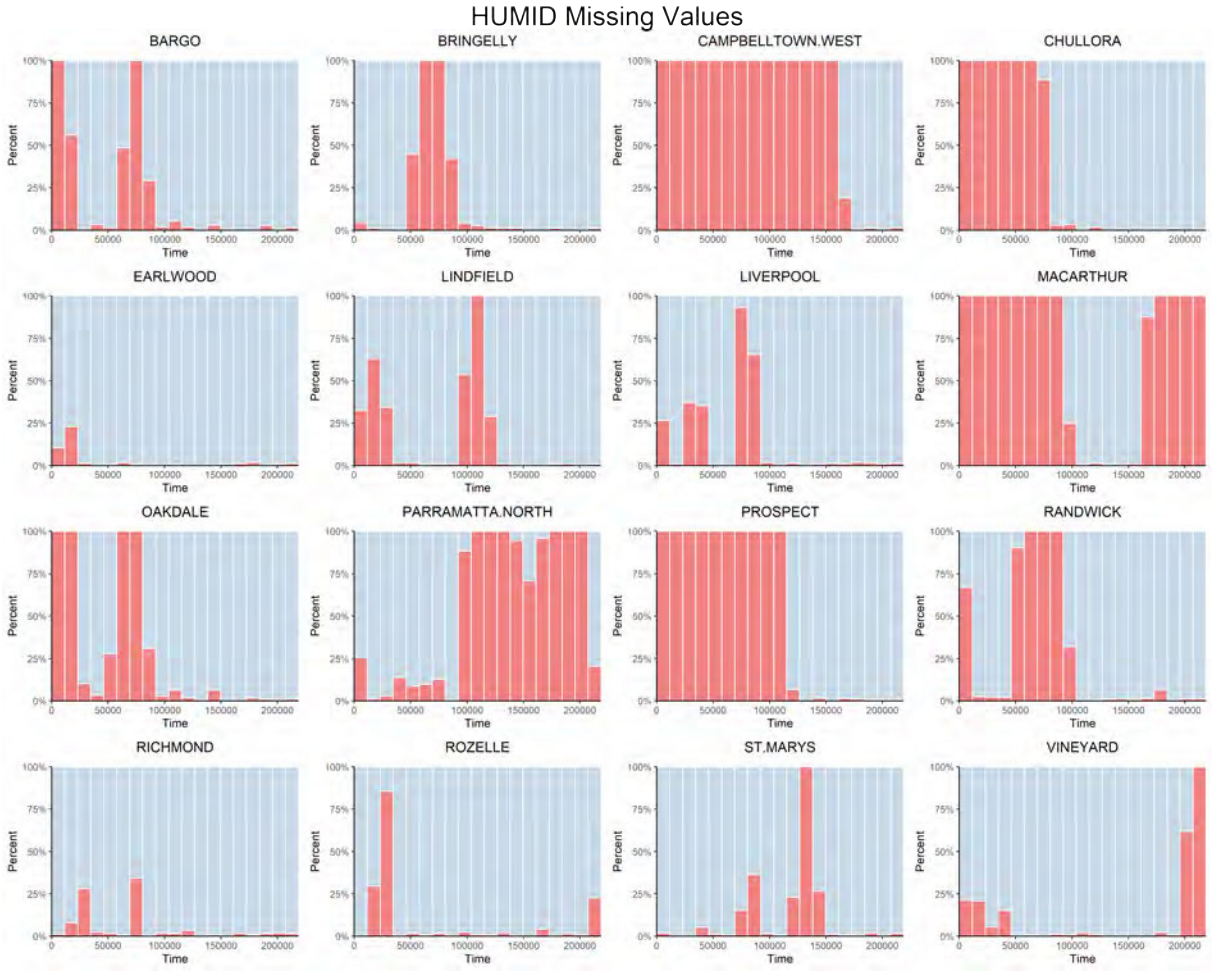


FIGURE A.10: Distribution of HUMID missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

FIGURE A.11: Distribution of SOLAR missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

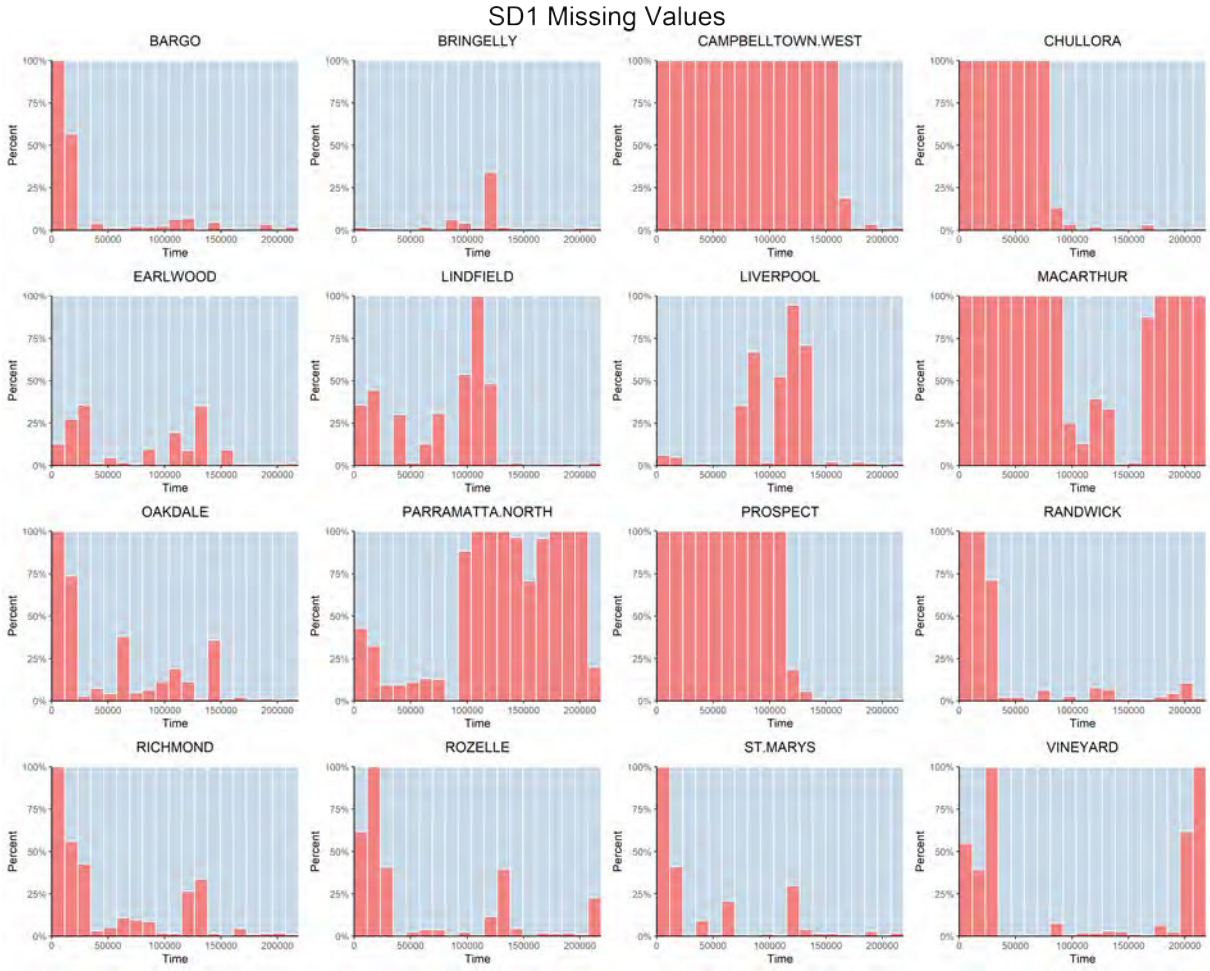


FIGURE A.12: Distribution of SD1 missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

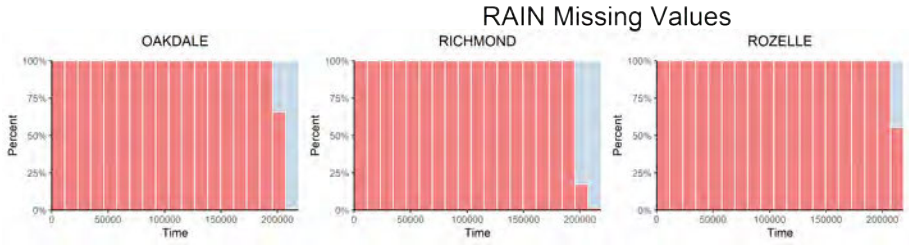


FIGURE A.13: Distribution of RAIN missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

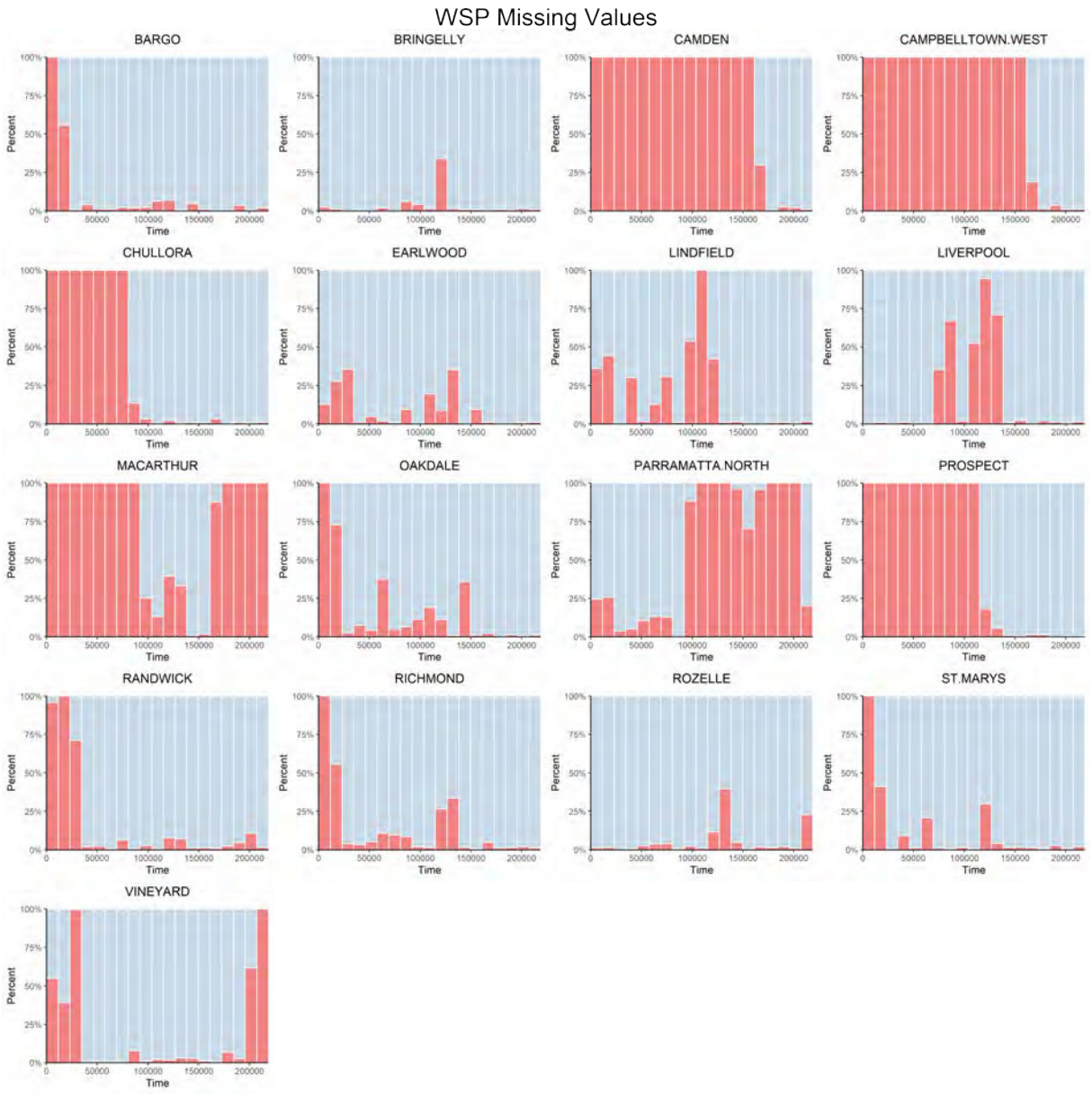


FIGURE A.14: Distribution of WSP missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

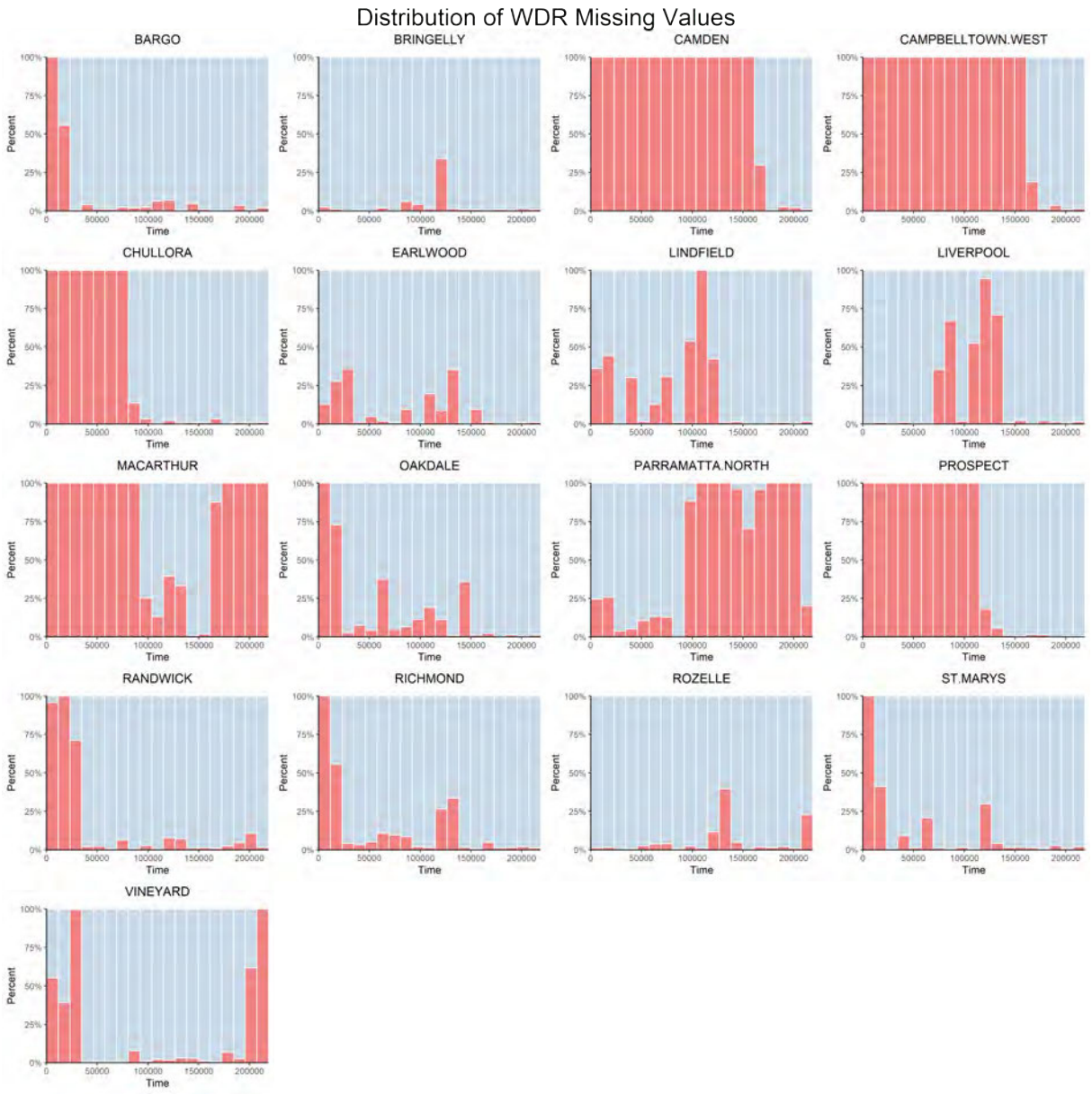


FIGURE A.15: Distribution of WDR missing values : X axis represents the time index for hourly data from 1994.01.01 1:00:00 AM to 31.12.2018 12:00:00 AM. Each red bar shows the percentage of missing values for each time interval whereas blue bar shows the percentage of observed values for the same interval.

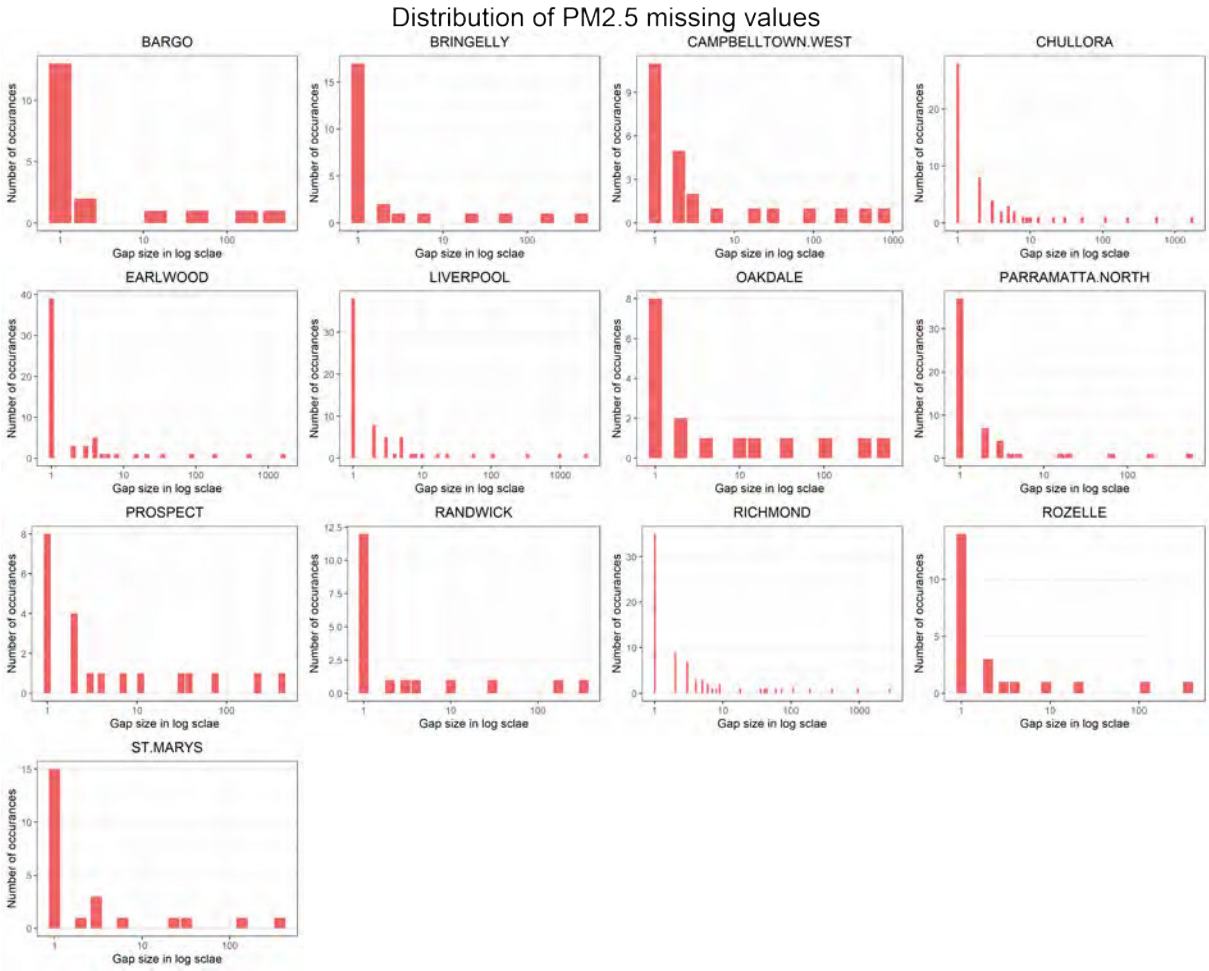


FIGURE A.16: Distribution of PM2.5 missing values

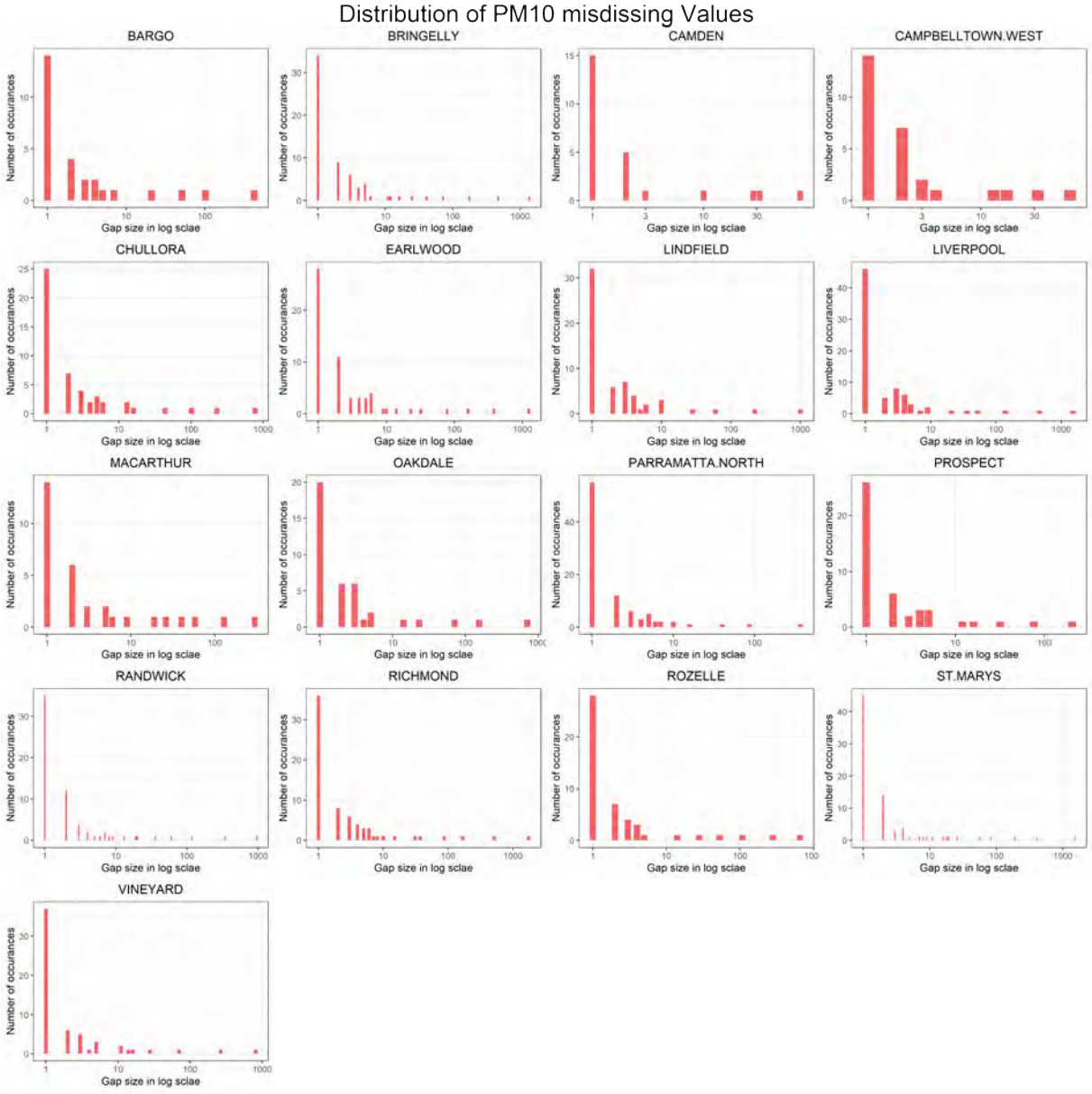


FIGURE A.17: Distribution of PM10 missing values

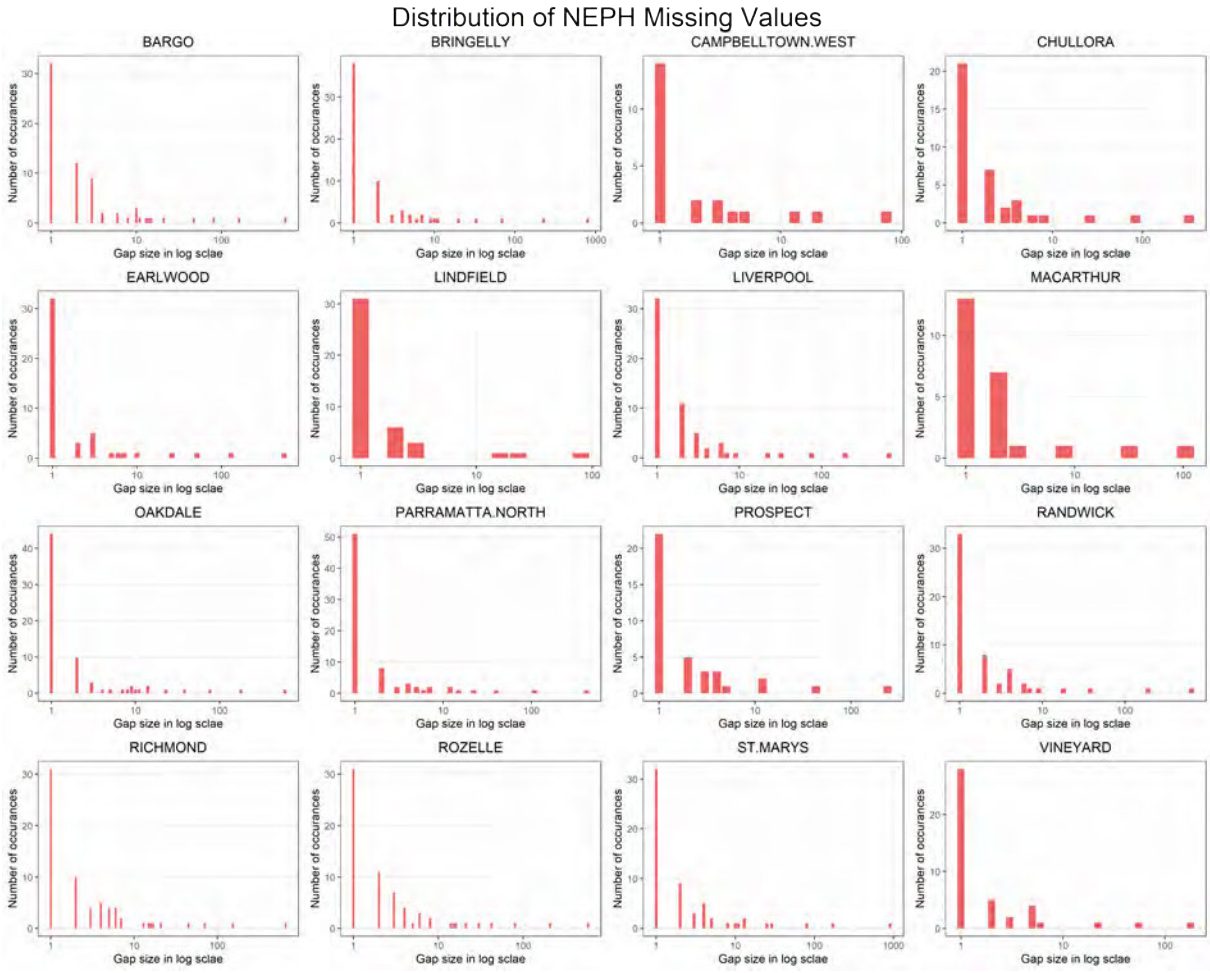


FIGURE A.18: Distribution of NEPH missing values

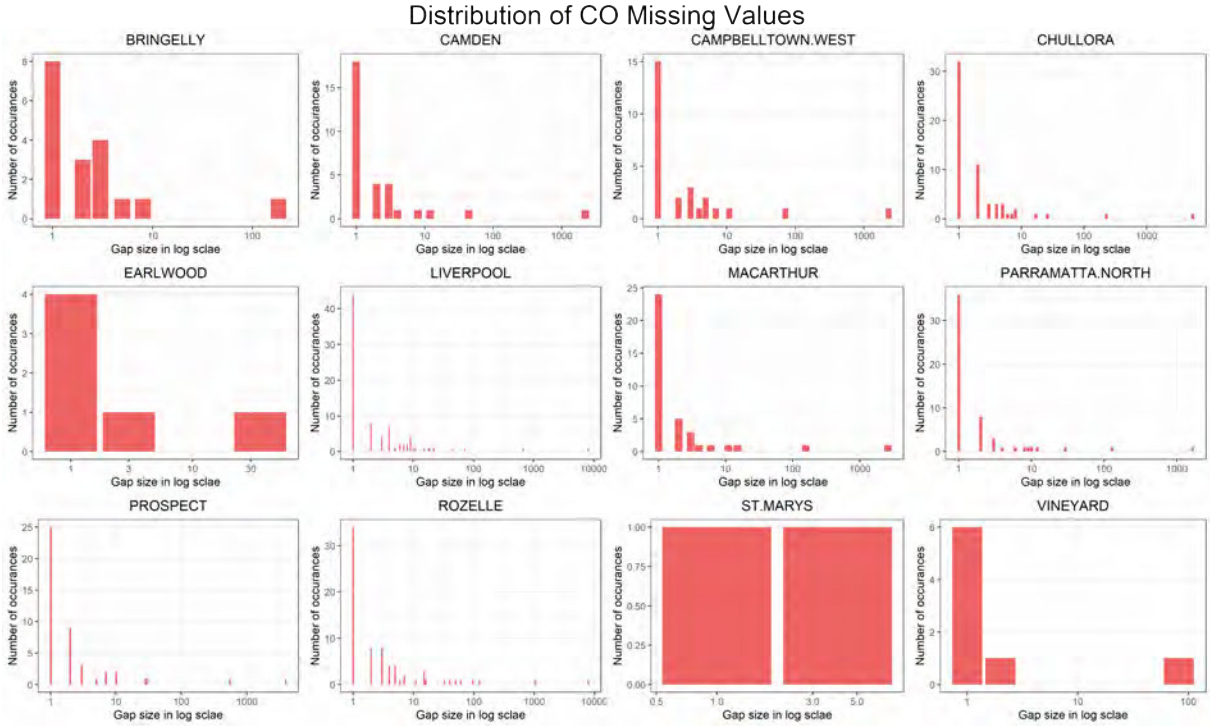


FIGURE A.19: Distribution of CO missing values

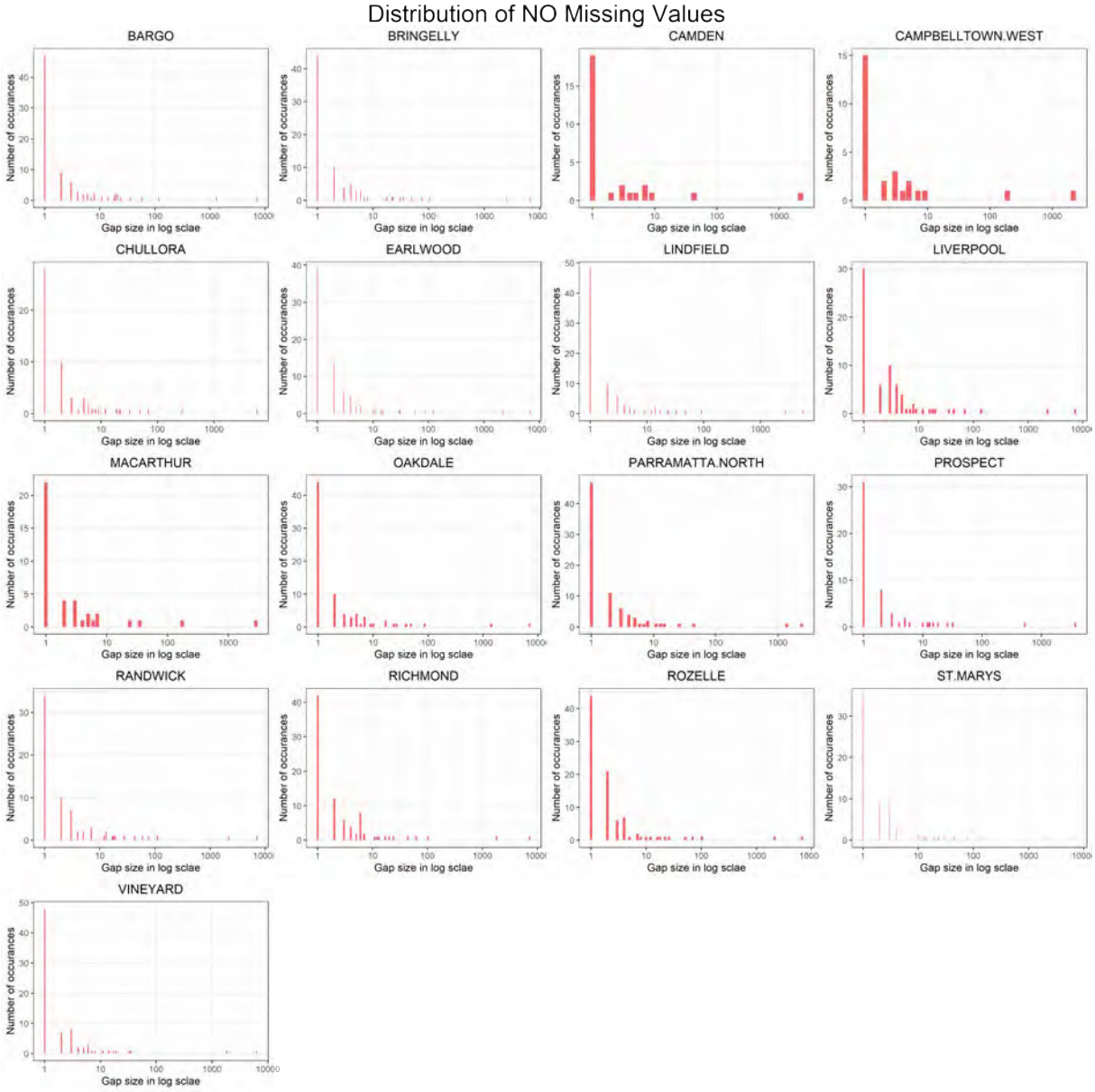


FIGURE A.20: Distribution of NO missing values

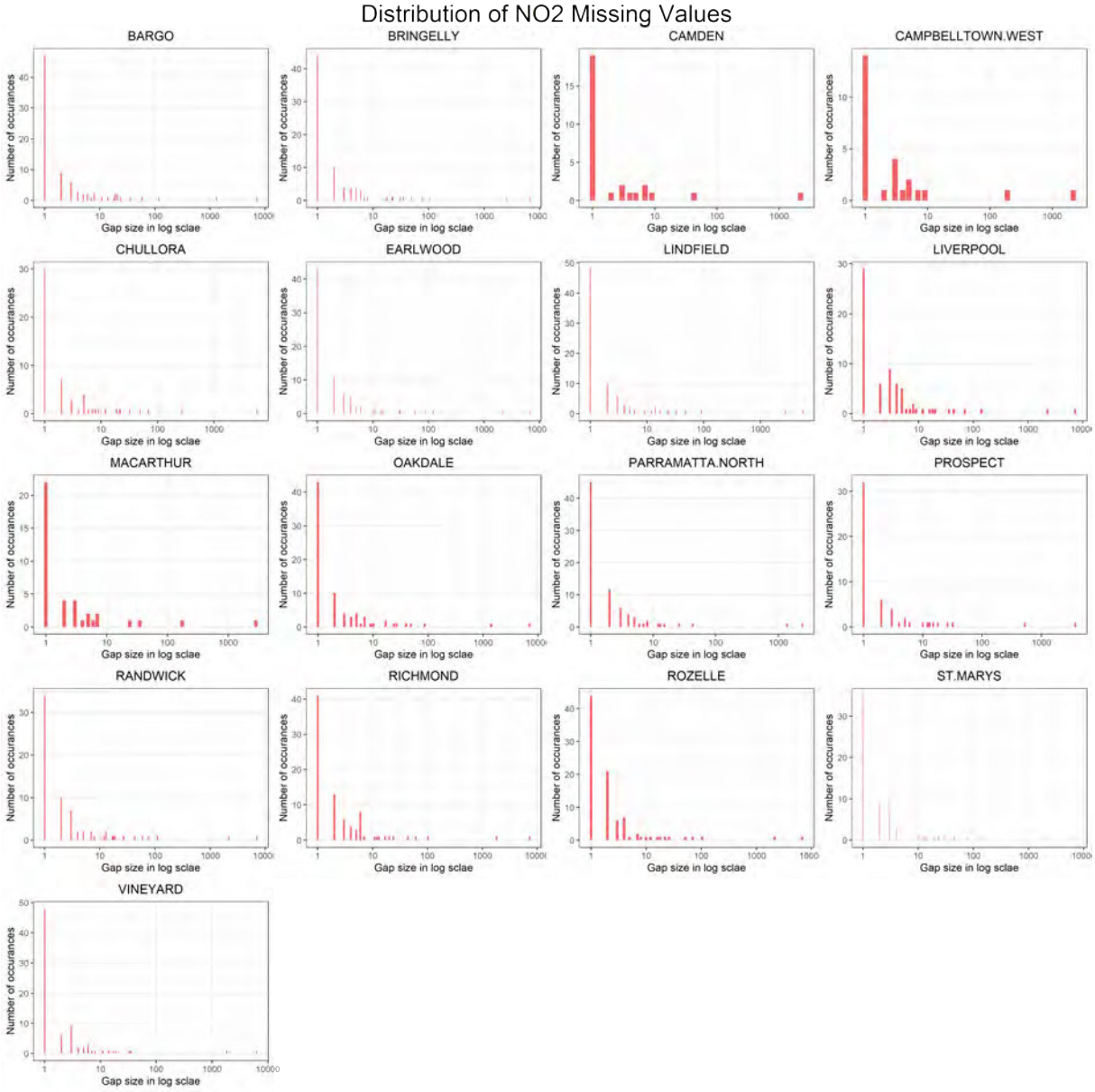


FIGURE A.21: Distribution of NO2 missing values

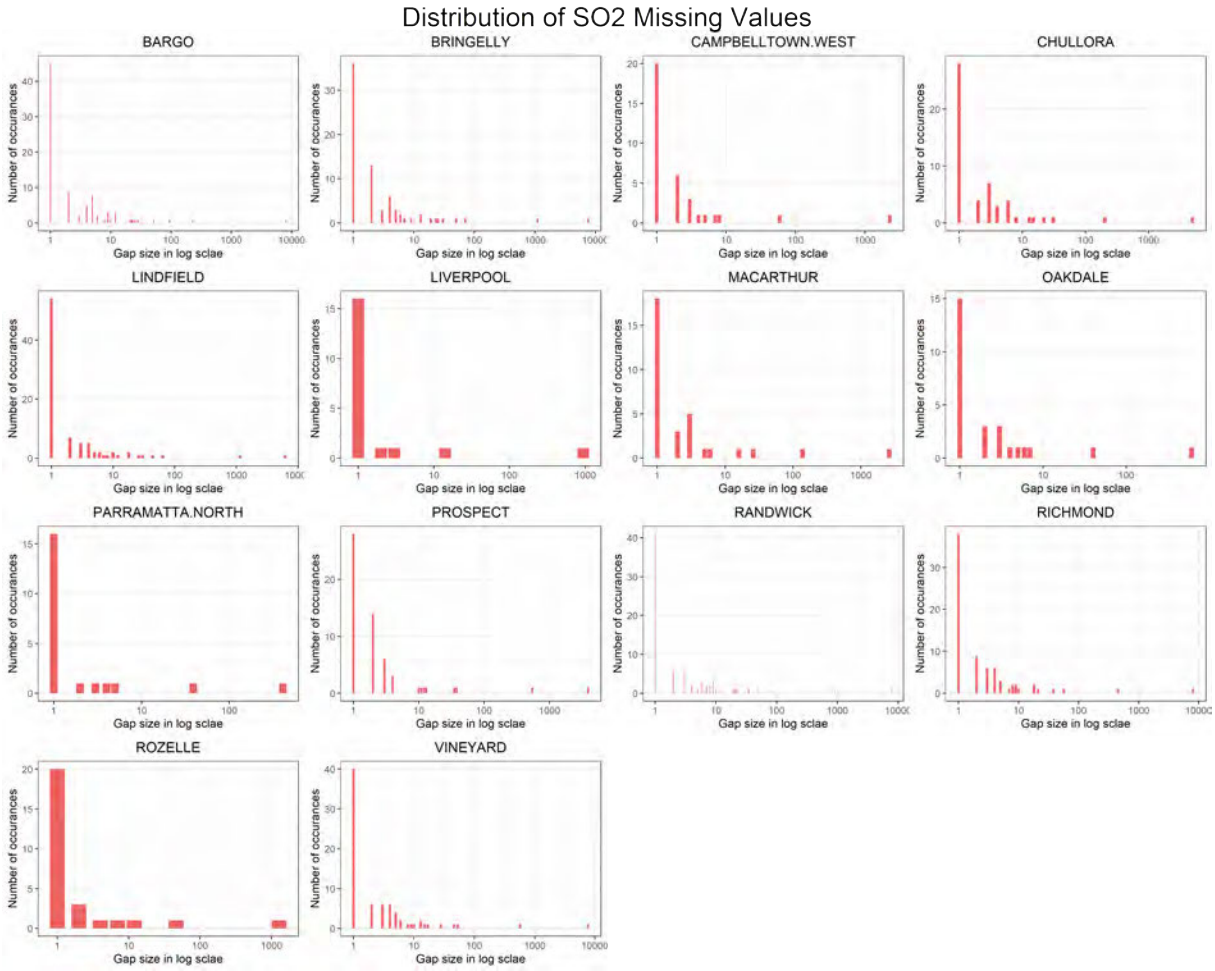


FIGURE A.22: Distribution of SO2 missing values

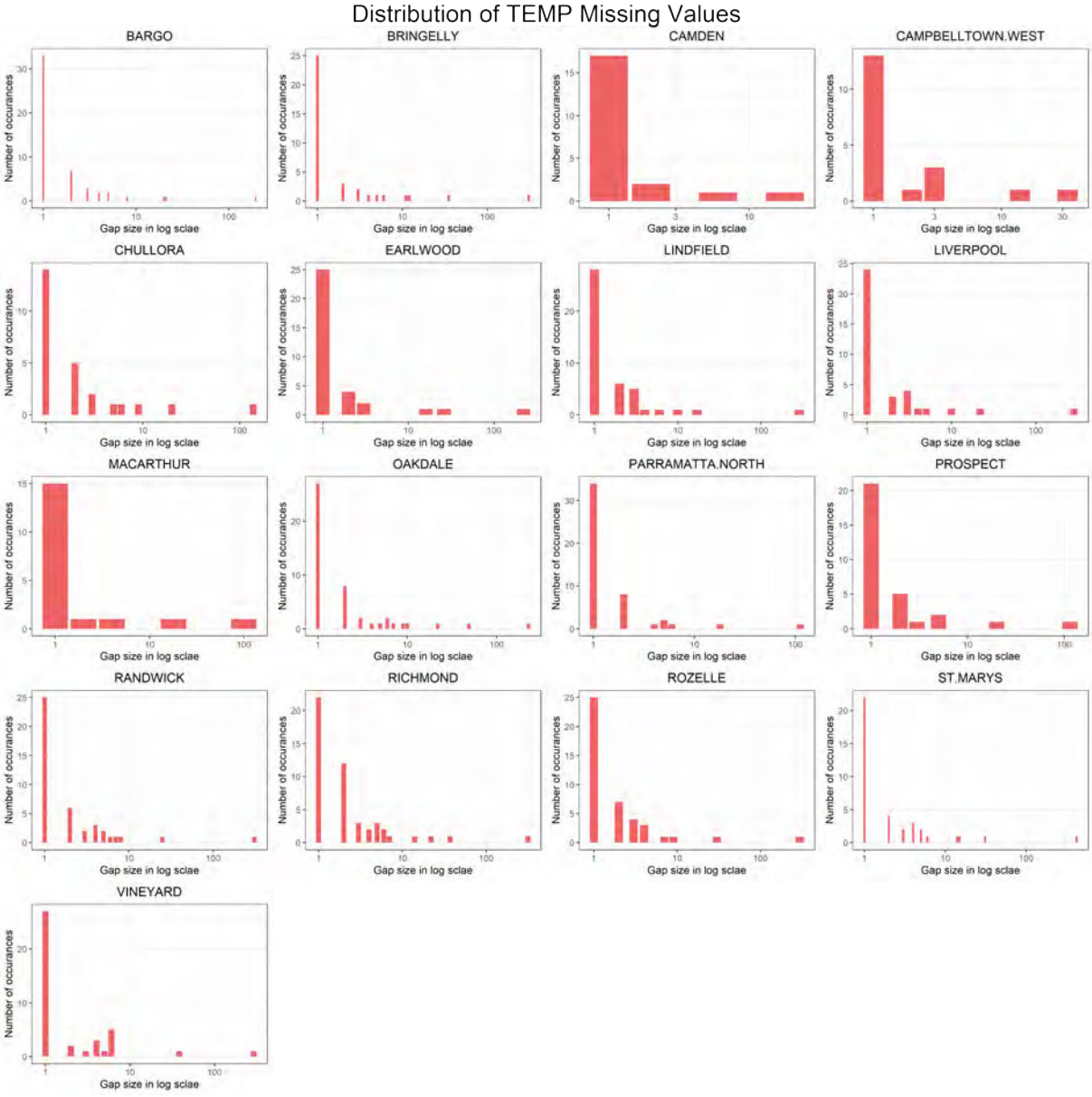


FIGURE A.23: Distribution of TEMP missing values

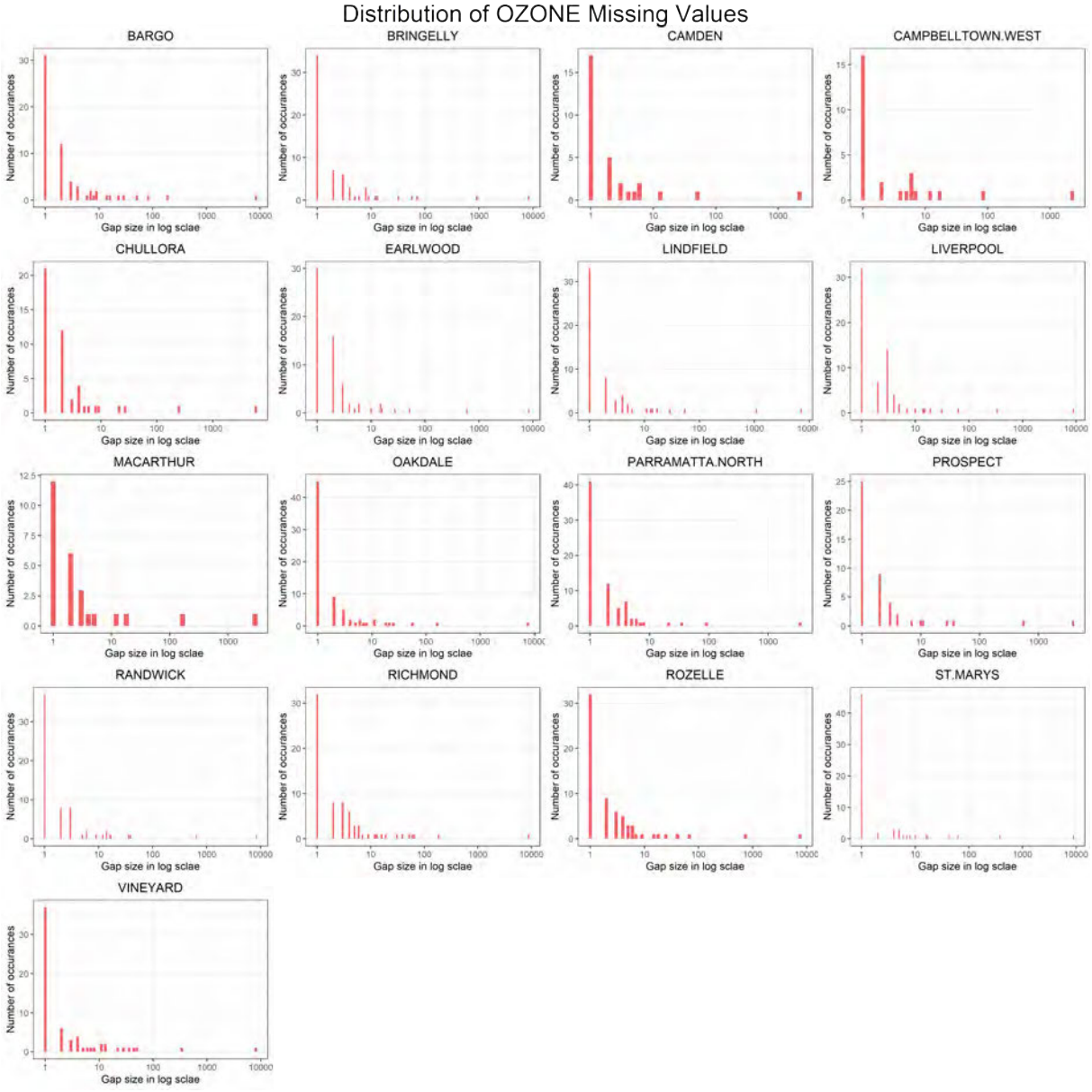


FIGURE A.24: Distribution of OZONE missing values

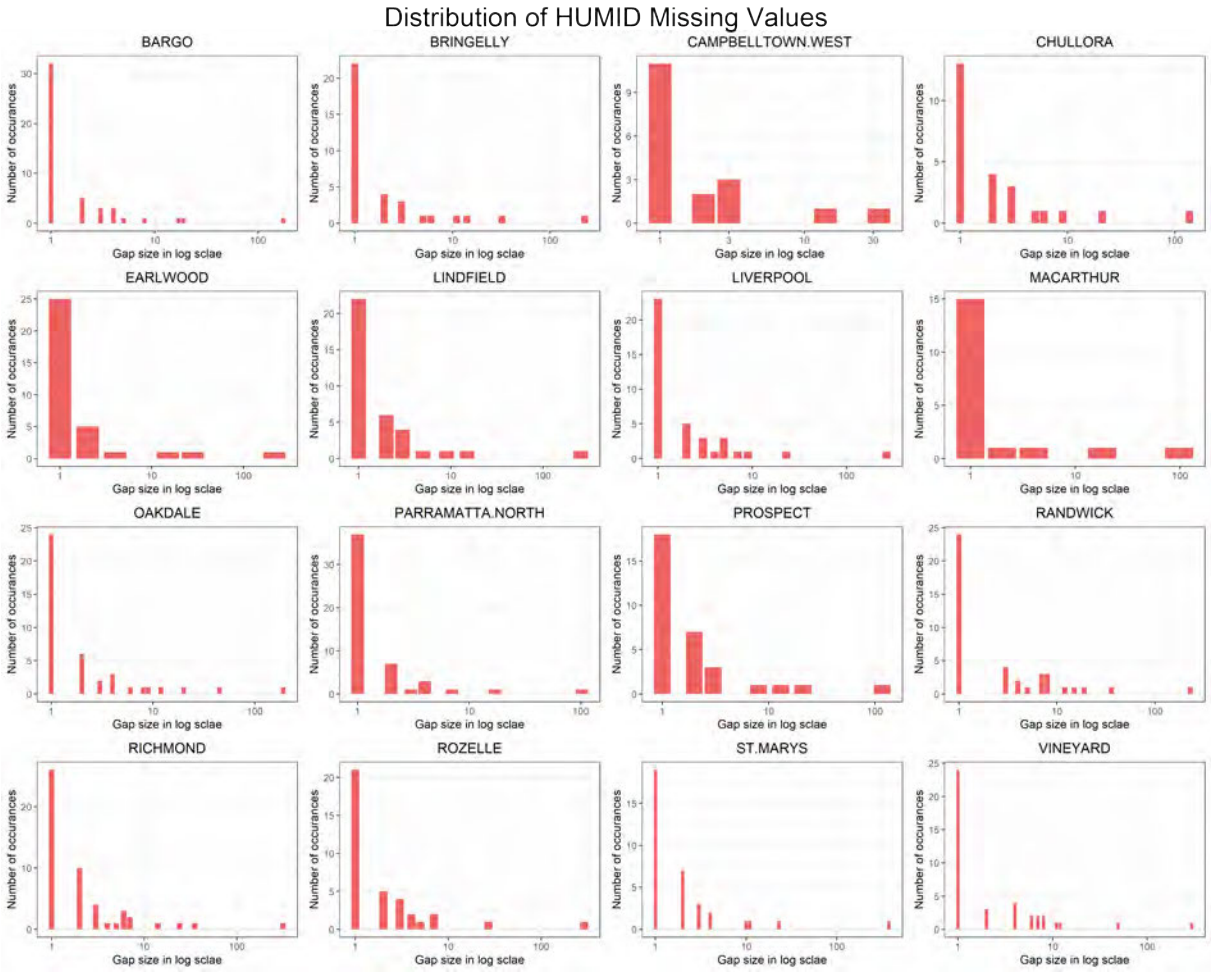


FIGURE A.25: Distribution of HUMID missing values

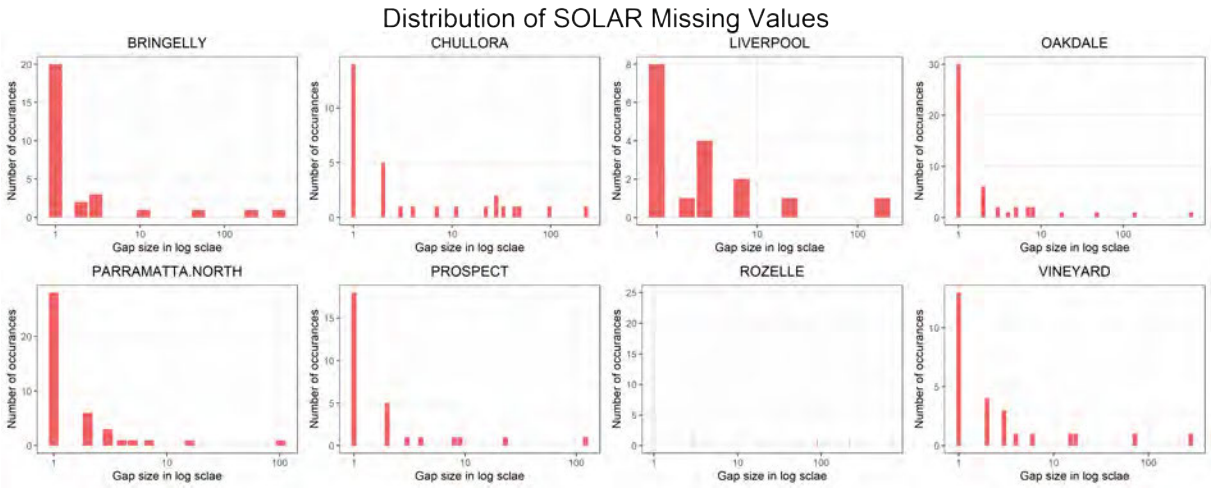


FIGURE A.26: Distribution of SOLAR missing values

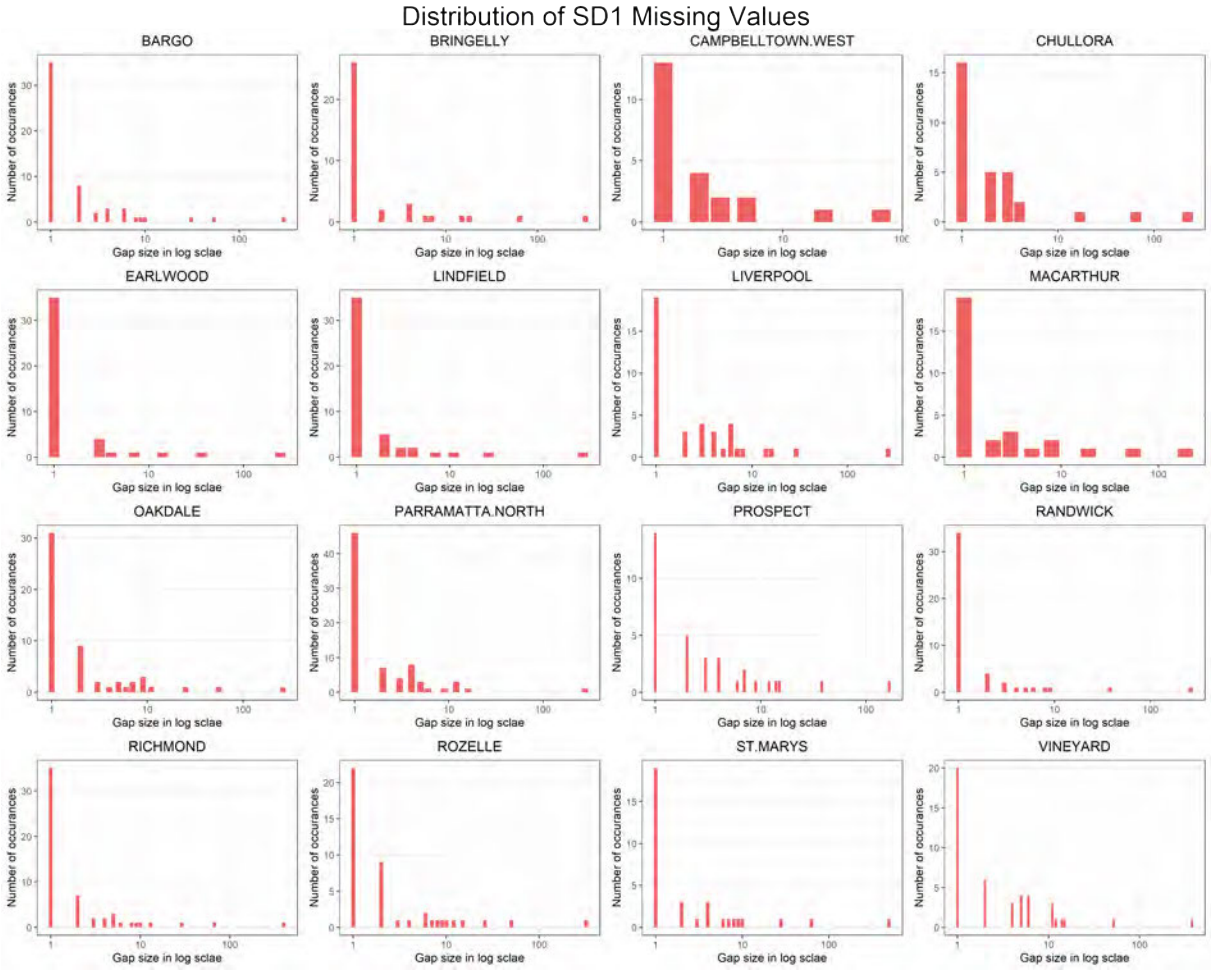


FIGURE A.27: Distribution of SD1 missing values

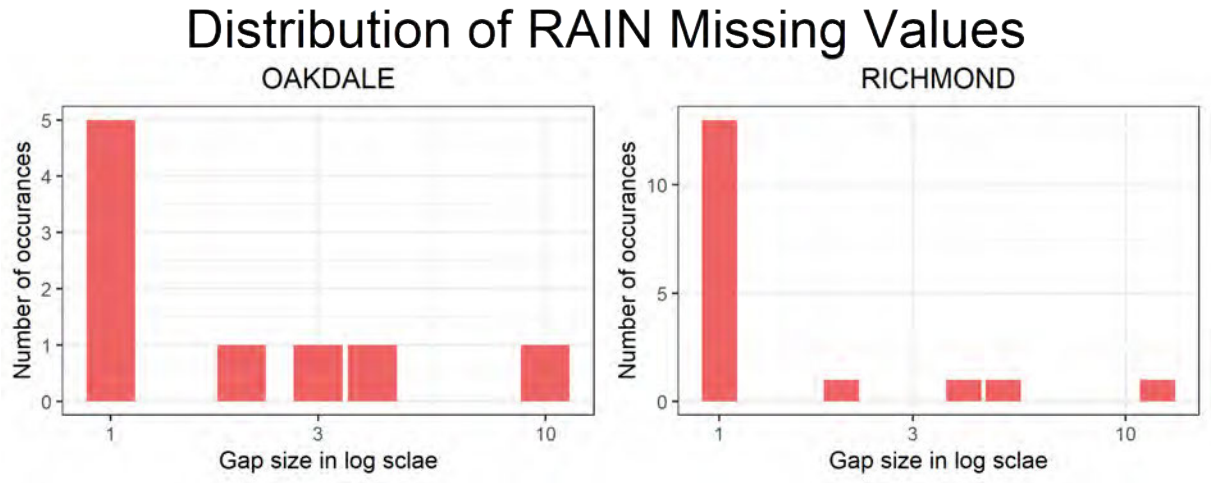


FIGURE A.28: Distribution of RAIN missing values

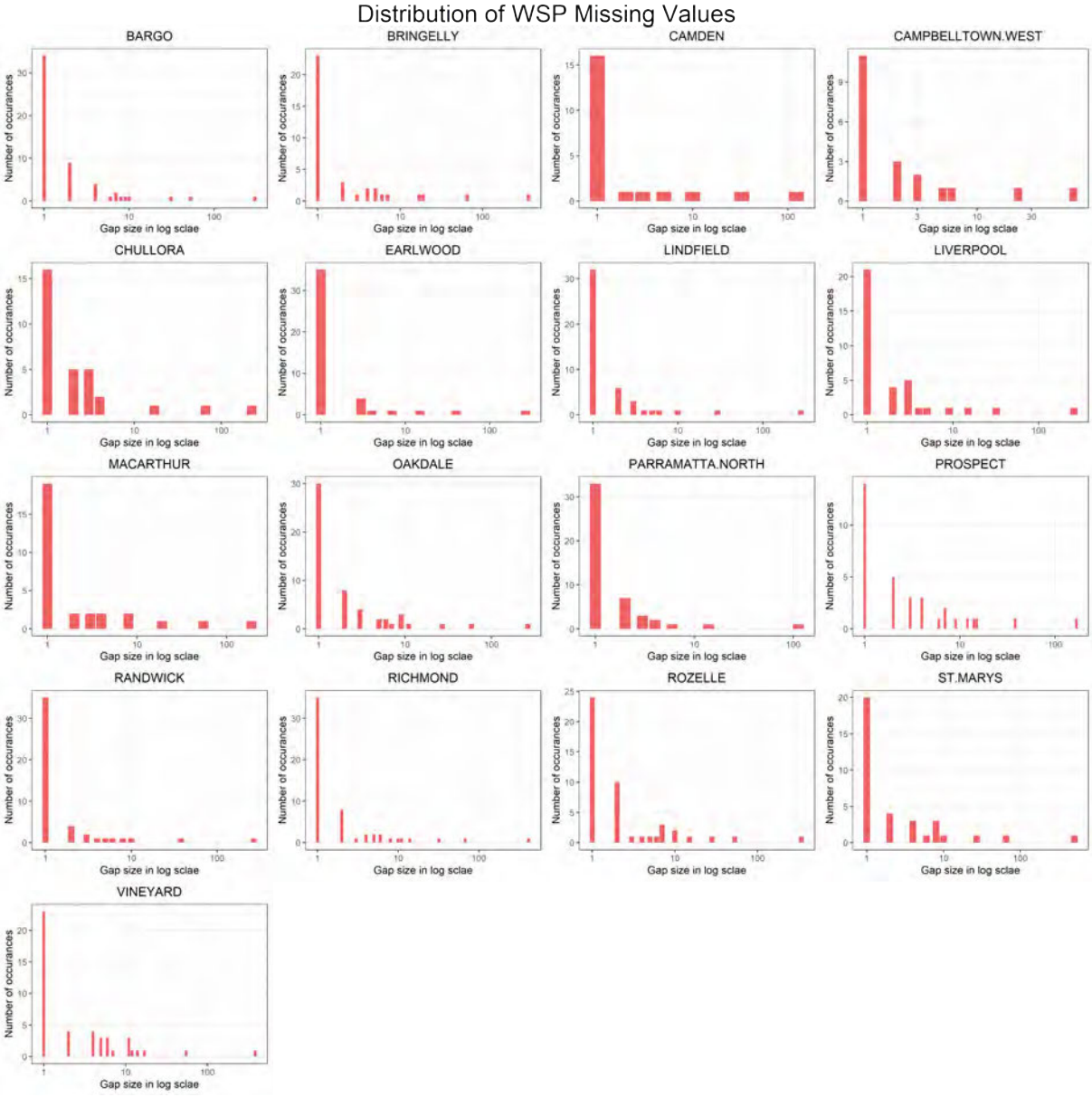


FIGURE A.29: Distribution of WSP missing values

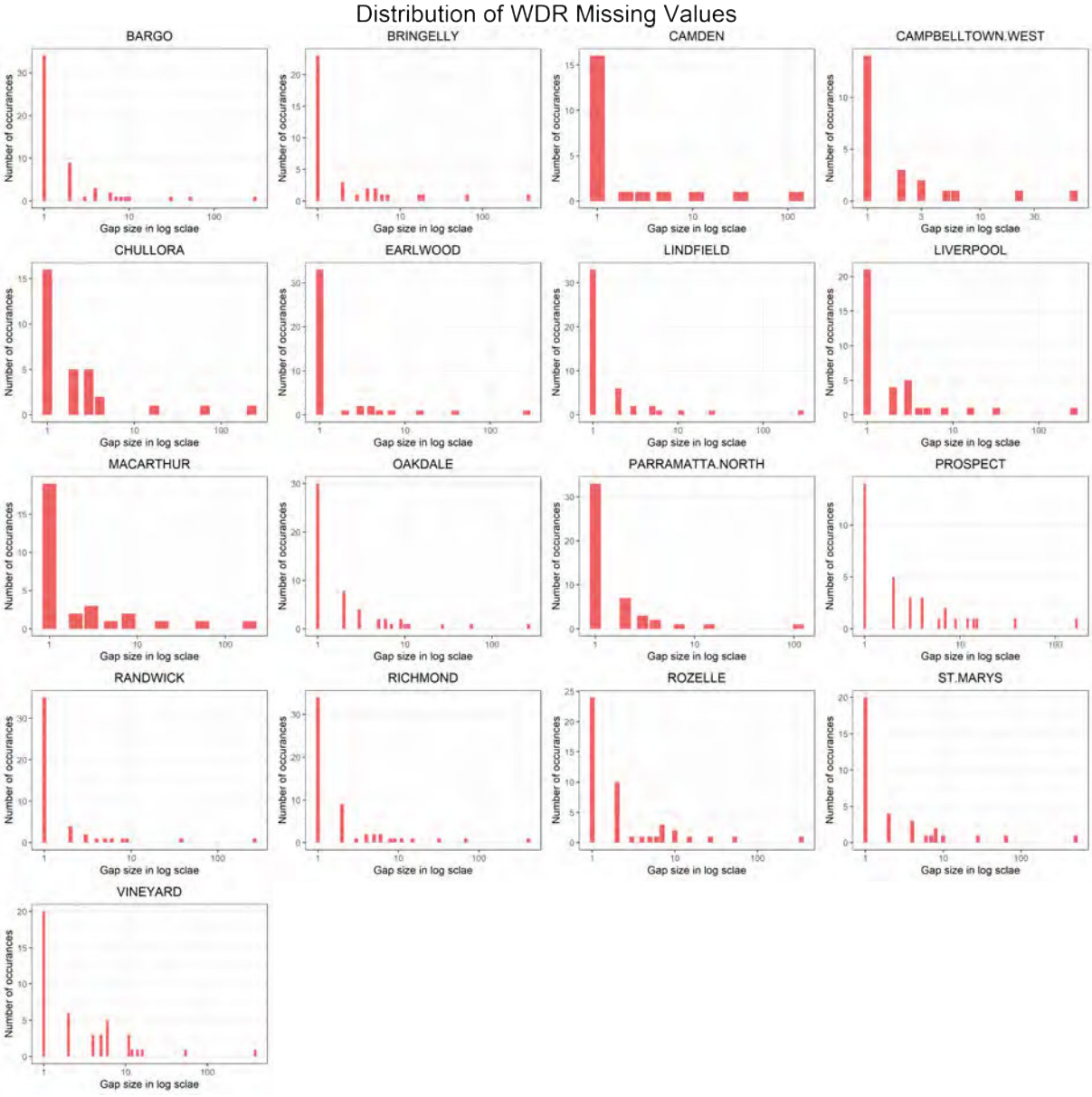


FIGURE A.30: Distribution of WDR missing values

	Station	Variable	Missing percentage	Logest gap size	Most frequent gap	Total missing values	No.of gap sizes
1	BARGO	TEMP	10.2%	17928	1	22273	318
2	BARGO	WSP	10.3%	17928	1	22578	500
3	BARGO	SO2	19.8%	17928	1	43302	8832
4	BARGO	NO	19.3%	17928	1	42128	8908
5	BARGO	NO2	19.3%	17928	1	42130	8908
6	BARGO	OZONE	15.8%	17928	1	34626	8904
7	BARGO	OZONE4	59.3%	106136	3	129839	218
8	BARGO	PM10	64.6%	139550	1	141322	644
9	BARGO	PM2.5	92.2%	200556	1	201751	618
10	BARGO	HUMID	18.8%	20219	1	41125	287
11	BARGO	NEPH	13.3%	17928	1	29158	1065
12	BARGO	SD1	10.4%	17928	1	22666	497
13	BRINGELLY	WDR	3.15%	3702	1	6903	525
14	BRINGELLY	TEMP	2.53%	3142	1	5530	418
15	BRINGELLY	WSP	3.15%	3702	1	6903	525
16	BRINGELLY	SO2	17.6%	19310	1	38447	9060

17	BRINGELLY	NO	12.1%	1538	1	26440	9696
18	BRINGELLY	NO2	12.1%	1538	1	26464	9696
19	BRINGELLY	CO	98.1%	142071	1	214779	221
20	BRINGELLY	OZONE	8.46%	1228	1	18515	9531
21	BRINGELLY	OZONE4	63.5%	106136	3	138855	211
22	BRINGELLY	PM10	4.87%	932	1	10660	2304
23	BRINGELLY	PM2.5	90.7%	196817	1	198478	712
24	BRINGELLY	HUMID	16.2%	32716	1	35360	333
25	BRINGELLY	NEPH	3.45%	311	1	7558	1288
26	BRINGELLY	SOLAR	29%	53243	1	63534	739
27	BRINGELLY	SD1	3.09%	2927	1	6752	485
28	CAMDEN	WDR	75.6%	164663	1	165390	192
29	CAMDEN	TEMP	75.5%	164663	1	165164	45
30	CAMDEN	WSP	75.6%	164663	1	165389	192
31	CAMDEN	NO	77.1%	164663	1	168614	2326
32	CAMDEN	NO2	77.1%	164663	1	168614	2326
33	CAMDEN	CO	77%	164663	1	168575	2331
34	CAMDEN	OZONE	77%	164663	1	168548	2330
35	CAMDEN	OZONE4	76.1%	164666	3	166540	119

36	CAMDEN	PM10	75.7%	164663	1	165722	172
37	CAMPBELLTOWN.WEST	WDR	75.1%	163367	1	164263	126
38	CAMPBELLTOWN.WEST	TEMP	74.9%	163367	1	163850	71
39	CAMPBELLTOWN.WEST	WSP	75%	163367	1	164223	127
40	CAMPBELLTOWN.WEST	SO2	76.3%	163367	1	167063	2405
41	CAMPBELLTOWN.WEST	NO	76.3%	163367	1	166965	2432
42	CAMPBELLTOWN.WEST	NO2	76.3%	163367	1	166964	2432
43	CAMPBELLTOWN.WEST	CO	76.3%	163367	1	166902	2427
44	CAMPBELLTOWN.WEST	OZONE	76.4%	163367	1	167267	2444
45	CAMPBELLTOWN.WEST	OZONE4	75.5%	163370	3	165311	164
46	CAMPBELLTOWN.WEST	PM10	75.4%	163367	1	164998	162
47	CAMPBELLTOWN.WEST	PM2.5	88.6%	189922	1	193865	1622
48	CAMPBELLTOWN.WEST	HUMID	74.9%	163367	1	163848	71
49	CAMPBELLTOWN.WEST	NEPH	75%	163367	1	164105	141
50	CAMPBELLTOWN.WEST	SD1	75.1%	163367	1	164264	126
51	CAMPBELLTOWN.WEST	CO8	77.2%	163373	9	168938	53
52	CHULLORA	WDR	38.3%	81925	1	83744	356
53	CHULLORA	TEMP	36.8%	79288	1	80635	208
54	CHULLORA	WSP	38.3%	81925	1	83745	356

55	CHULLORA	SO2	49.8%	98751	1	108881	5344
56	CHULLORA	NO	42.1%	79695	1	92167	6428
57	CHULLORA	NO2	42.1%	79695	1	92091	6429
58	CHULLORA	CO	45.2%	87867	1	98963	5834
59	CHULLORA	OZONE	40.8%	79755	1	89292	6325
60	CHULLORA	OZONE4	69.8%	79757	3	152818	156
61	CHULLORA	PM10	38.4%	79288	1	84115	1276
62	CHULLORA	PM2.5	40.6%	78864	1	88941	2827
63	CHULLORA	HUMID	36.8%	79288	1	80633	207
64	CHULLORA	NEPH	37.6%	79695	1	82326	504
65	CHULLORA	SOLAR	44%	92002	1	96343	576
66	CHULLORA	SD1	38.3%	81925	1	83745	357
67	CHULLORA	CO8	88.4%	192479	10	193436	35
68	EARLWOOD	WDR	9%	3453	1	19694	383
69	EARLWOOD	TEMP	2.44%	2657	1	5342	302
70	EARLWOOD	WSP	8.98%	3453	1	19642	379
71	EARLWOOD	NO	12.5%	2657	1	27253	9557
72	EARLWOOD	NO2	12.5%	2657	1	27385	9551
73	EARLWOOD	CO	99.7%	184261	1	218089	43

74	EARLWOOD	OZONE	9.01%	2657	1	19722	9260
75	EARLWOOD	OZONE4	69.7%	106136	3	152490	149
76	EARLWOOD	PM10	11.5%	8568	1	25232	2058
77	EARLWOOD	PM2.5	16.6%	22608	1	36232	2586
78	EARLWOOD	HUMID	2.41%	2657	1	5268	300
79	EARLWOOD	NEPH	13.4%	23603	1	29249	835
80	EARLWOOD	SD1	8.97%	3458	1	19632	353
81	LINDFIELD	WDR	18.9%	19722	1	41385	382
82	LINDFIELD	TEMP	13%	20213	1	28493	379
83	LINDFIELD	WSP	18.9%	19722	1	41397	387
84	LINDFIELD	SO2	29.8%	19722	1	65205	7673
85	LINDFIELD	NO	21.6%	20155	1	47321	8778
86	LINDFIELD	NO2	21.6%	20155	1	47321	8778
87	LINDFIELD	OZONE	18.1%	20155	1	39688	8403
88	LINDFIELD	OZONE4	58.4%	106136	3	127885	181
89	LINDFIELD	PM10	18.8%	20185	1	41142	1456
90	LINDFIELD	HUMID	17%	20213	1	37161	333
91	LINDFIELD	NEPH	53.6%	94505	1	117372	171
92	LINDFIELD	SOLAR	100%	117215	117215	218817	2

93	LINDFIELD	SD1	19.2%	12394	1	42054	370
94	LIVERPOOL	WDR	17.6%	19018	1	38467	365
95	LIVERPOOL	TEMP	18.6%	14464	1	40625	367
96	LIVERPOOL	WSP	17.6%	19018	1	38460	366
97	LIVERPOOL	SO2	90.3%	195842	1	197505	978
98	LIVERPOOL	NO	9.92%	767	1	21711	9985
99	LIVERPOOL	NO2	9.61%	362	1	21031	10021
100	LIVERPOOL	CO	12.1%	1346	1	26526	9245
101	LIVERPOOL	OZONE	8.1%	889	1	17733	9474
102	LIVERPOOL	OZONE4	69.8%	106136	3	152759	136
103	LIVERPOOL	PM10	7.32%	2863	1	16027	2372
104	LIVERPOOL	PM2.5	23.5%	35040	1	51384	4030
105	LIVERPOOL	HUMID	14.3%	18247	1	31362	365
106	LIVERPOOL	NEPH	5.15%	4409	1	11278	1094
107	LIVERPOOL	SOLAR	56.9%	123641	1	124526	233
108	LIVERPOOL	SD1	18.1%	19018	1	39612	418
109	LIVERPOOL	CO8	88.3%	192479	9	193129	25
110	MACARTHUR	WDR	73.8%	94943	1	161418	317
111	MACARTHUR	TEMP	69.3%	94943	1	151647	139

112	MACARTHUR	WSP	73.8%	94943	1	161417	317
113	MACARTHUR	SO2	73.6%	100287	1	161144	2862
114	MACARTHUR	NO	71.7%	94943	1	156802	3164
115	MACARTHUR	NO2	71.7%	94943	1	156802	3164
116	MACARTHUR	CO	73.7%	100287	1	161329	2859
117	MACARTHUR	OZONE	71.1%	94943	1	155564	3127
118	MACARTHUR	OZONE4	77.7%	94945	3	170121	125
119	MACARTHUR	PM10	70.1%	94943	1	153470	627
120	MACARTHUR	PM2.5	100%	189252	189252	218814	2
121	MACARTHUR	HUMID	69.3%	94943	1	151645	137
122	MACARTHUR	NEPH	69.6%	95002	1	152338	175
123	MACARTHUR	SD1	73.8%	94943	1	161418	317
124	MACARTHUR	CO8	79.2%	100292	7	173210	81
125	OAKDALE	WDR	16.9%	19800	1	37068	474
126	OAKDALE	TEMP	11.7%	19800	1	25646	395
127	OAKDALE	WSP	16.9%	19800	1	37066	474
128	OAKDALE	SO2	93.7%	124905	1	205147	707
129	OAKDALE	NO	20.3%	19983	1	44483	8766
130	OAKDALE	NO2	20.3%	19983	1	44488	8764

131	OAKDALE	OZONE	24.1%	19936	1	52721	8004
132	OAKDALE	OZONE4	62.9%	110156	3	137576	204
133	OAKDALE	PM10	43.5%	90880	1	95187	1032
134	OAKDALE	PM2.5	92.5%	200579	1	202494	998
135	OAKDALE	HUMID	26%	28605	1	56935	345
136	OAKDALE	NEPH	14.6%	19936	1	31848	1079
137	OAKDALE	SOLAR	13.3%	19800	1	29025	962
138	OAKDALE	SD1	17%	19800	1	37151	469
139	OAKDALE	RAIN	93%	203365	1	203488	24
140	PARRAMATTA.NORTH	WDR	56.2%	55808	1	122907	196
141	PARRAMATTA.NORTH	TEMP	54.2%	55616	1	118629	200
142	PARRAMATTA.NORTH	WSP	56.2%	55808	1	122906	197
143	PARRAMATTA.NORTH	SO2	94.8%	149099	1	207410	476
144	PARRAMATTA.NORTH	NO	59.9%	55610	1	131106	4054
145	PARRAMATTA.NORTH	NO2	59.9%	55610	1	131139	4054
146	PARRAMATTA.NORTH	CO	69.6%	66019	1	152394	1920
147	PARRAMATTA.NORTH	OZONE	57.8%	56215	1	126470	3772
148	PARRAMATTA.NORTH	OZONE4	85.2%	80995	3	186497	169
149	PARRAMATTA.NORTH	PM10	58.9%	55660	1	128915	691

150	PARRAMATTA.NORTH	PM2.5	73.7%	116503	1	161277	913
151	PARRAMATTA.NORTH	HUMID	55%	55616	1	120282	192
152	PARRAMATTA.NORTH	NEPH	56.8%	55614	1	124218	745
153	PARRAMATTA.NORTH	SOLAR	57.4%	57108	1	125705	183
154	PARRAMATTA.NORTH	SD1	58%	55808	1	126998	465
155	PARRAMATTA.NORTH	CO8	84.9%	165975	6	185838	36
156	PROSPECT	WDR	54.2%	115546	1	118544	319
157	PROSPECT	TEMP	53.3%	114839	1	116577	179
158	PROSPECT	WSP	54.2%	115546	1	118545	320
159	PROSPECT	SO2	57.3%	115295	1	125449	4653
160	PROSPECT	NO	58.9%	116560	1	128976	4463
161	PROSPECT	NO2	58.9%	116560	1	128975	4463
162	PROSPECT	CO	56.7%	115367	1	124129	4654
163	PROSPECT	OZONE	56.7%	115295	1	124056	4639
164	PROSPECT	OZONE4	70.3%	115297	3	153907	313
165	PROSPECT	PM10	54.6%	115313	1	119394	411
166	PROSPECT	PM2.5	84.6%	183086	1	185102	816
167	PROSPECT	HUMID	53.3%	114839	1	116680	200
168	PROSPECT	NEPH	53.7%	114839	1	117559	364

169	PROSPECT	SOLAR	53.6%	115295	1	117217	196
170	PROSPECT	SD1	54.2%	115546	1	118544	319
171	PROSPECT	CO8	88.4%	192479	6	193523	28
172	RANDWICK	WDR	16.8%	25173	1	36714	384
173	RANDWICK	TEMP	4.98%	5544	1	10905	419
174	RANDWICK	WSP	16.8%	25173	1	36717	385
175	RANDWICK	SO2	18.6%	19399	1	40759	9104
176	RANDWICK	NO	13.1%	5358	1	28767	9803
177	RANDWICK	NO2	13.2%	5358	1	28789	9800
178	RANDWICK	OZONE	11.8%	5365	1	25759	9302
179	RANDWICK	OZONE4	59.9%	106136	3	131033	190
180	RANDWICK	PM10	8.41%	6112	1	18403	1686
181	RANDWICK	PM2.5	93.6%	203412	1	204838	567
182	RANDWICK	HUMID	26.8%	48516	1	58732	407
183	RANDWICK	NEPH	6.02%	5365	1	13169	993
184	RANDWICK	SD1	17%	31216	1	37208	382
185	RICHMOND	WDR	14.4%	17351	1	31546	645
186	RICHMOND	TEMP	5.36%	4047	1	11737	483
187	RICHMOND	WSP	14.4%	17351	1	31534	645

188	RICHMOND	SO2	18.7%	18472	1	40853	8776
189	RICHMOND	NO	11.1%	1710	1	24226	9570
190	RICHMOND	NO2	11.1%	1710	1	24247	9567
191	RICHMOND	OZONE	8.7%	294	1	19040	9535
192	RICHMOND	OZONE4	65%	106136	3	142292	208
193	RICHMOND	PM10	5.35%	1783	1	11711	2751
194	RICHMOND	PM2.5	20.4%	20559	1	44591	4906
195	RICHMOND	HUMID	4.91%	3820	1	10745	485
196	RICHMOND	NEPH	4.2%	724	1	9199	1156
197	RICHMOND	SD1	16.5%	17351	1	36009	618
198	RICHMOND	RAIN	90.5%	197560	1	197967	36
199	ROZELLE	WDR	5.44%	2501	1	11896	538
200	ROZELLE	TEMP	2.27%	2502	1	4958	406
201	ROZELLE	WSP	5.44%	2501	1	11899	539
202	ROZELLE	SO2	86.7%	185115	1	189764	1376
203	ROZELLE	NO	13.8%	2512	1	30268	9384
204	ROZELLE	NO2	13.8%	2512	1	30268	9383
205	ROZELLE	CO	11.7%	2512	1	25578	9581
206	ROZELLE	OZONE	17.8%	18716	1	38876	8569

207	ROZELLE	OZONE4	65.2%	106136	3	142609	227
208	ROZELLE	PM10	42.9%	86651	1	93829	1206
209	ROZELLE	PM2.5	86.4%	185534	1	189179	531
210	ROZELLE	HUMID	8.25%	13065	1	18062	386
211	ROZELLE	NEPH	5.9%	2501	1	12912	1131
212	ROZELLE	SOLAR	5.1%	2523	1	11156	1559
213	ROZELLE	SD1	16%	23225	1	34912	516
214	ROZELLE	CO8	89.7%	192479	28	196236	44
215	ROZELLE	RAIN	97.7%	213640	213640	213695	5
216	ST.MARYS	WDR	11.6%	15599	1	25388	675
217	ST.MARYS	TEMP	1.82%	1562	1	3992	544
218	ST.MARYS	WSP	11.6%	15599	1	25390	675
219	ST.MARYS	NO	11.4%	2011	1	24987	9886
220	ST.MARYS	NO2	11.4%	2011	1	24987	9886
221	ST.MARYS	CO	100%	218737	1	218741	5
222	ST.MARYS	OZONE	7.89%	1026	1	17271	9572
223	ST.MARYS	OZONE4	63.2%	106136	3	138292	360
224	ST.MARYS	PM10	9.29%	5593	1	20321	2486
225	ST.MARYS	PM2.5	89.4%	194242	1	195616	599

226	ST.MARYS	HUMID	11.6%	17057	1	25323	471
227	ST.MARYS	NEPH	3.69%	543	1	8075	1361
228	ST.MARYS	SD1	11.6%	15599	1	25381	665
229	VINEYARD	WDR	20.4%	18616	1	44653	602
230	VINEYARD	TEMP	12.4%	18616	1	27171	408
231	VINEYARD	WSP	20.4%	18616	1	44598	603
232	VINEYARD	SO2	19.4%	18616	1	42467	8548
233	VINEYARD	NO	19.9%	18616	1	43448	8481
234	VINEYARD	NO2	19.9%	18616	1	43474	8481
235	VINEYARD	CO	98.5%	213094	1	215651	90
236	VINEYARD	OZONE	17.3%	18616	1	37912	8671
237	VINEYARD	OZONE4	67.7%	106136	3	148240	158
238	VINEYARD	PM10	21.1%	20601	1	46210	1319
239	VINEYARD	HUMID	12.4%	18616	1	27148	453
240	VINEYARD	NEPH	44.7%	74604	1	97828	325
241	VINEYARD	SOLAR	51.2%	74507	1	111934	423
242	VINEYARD	SD1	20.4%	18616	1	44626	591

Figures below show the distributions of variables.

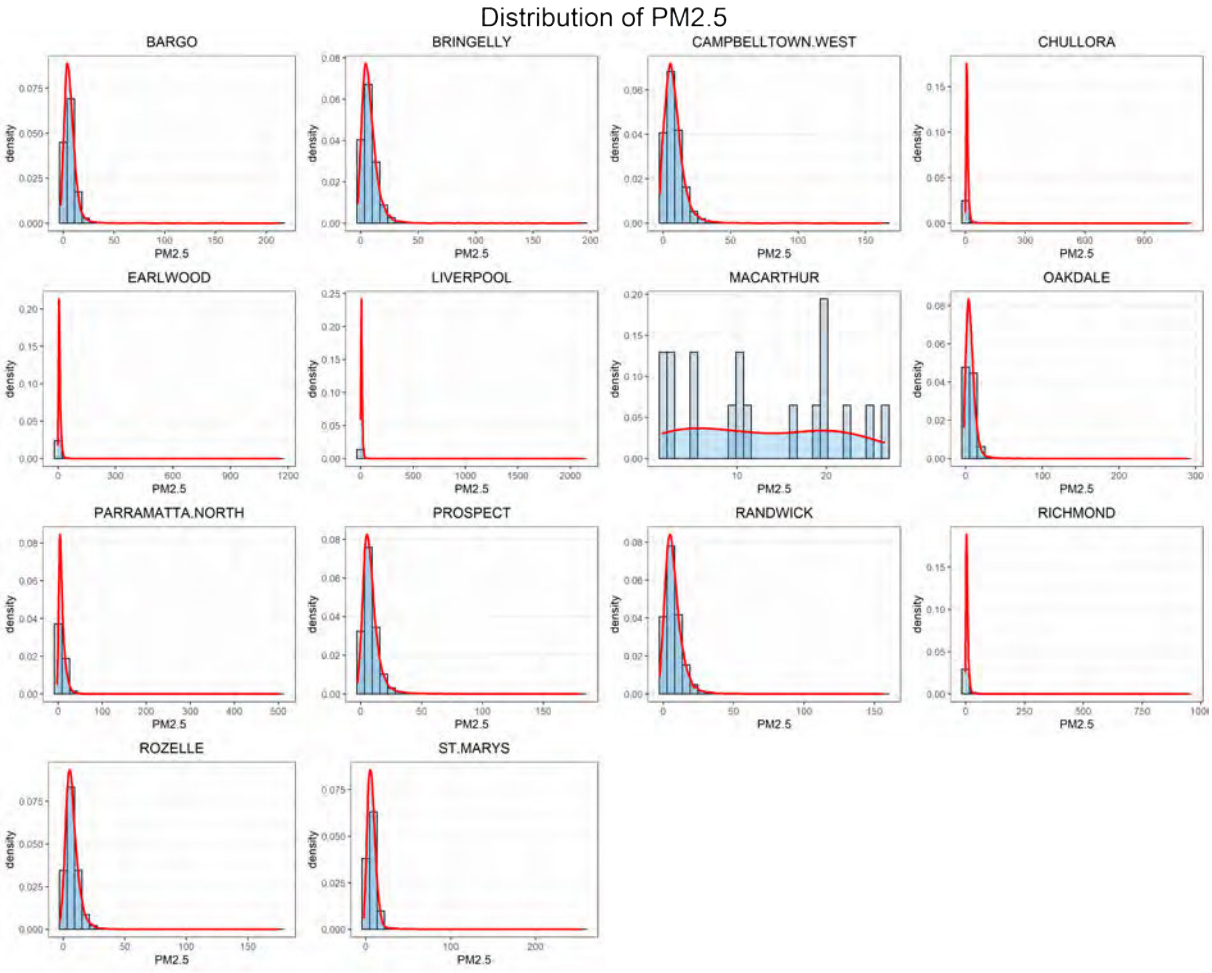


FIGURE A.31: Distribution of PM2.5 missing values

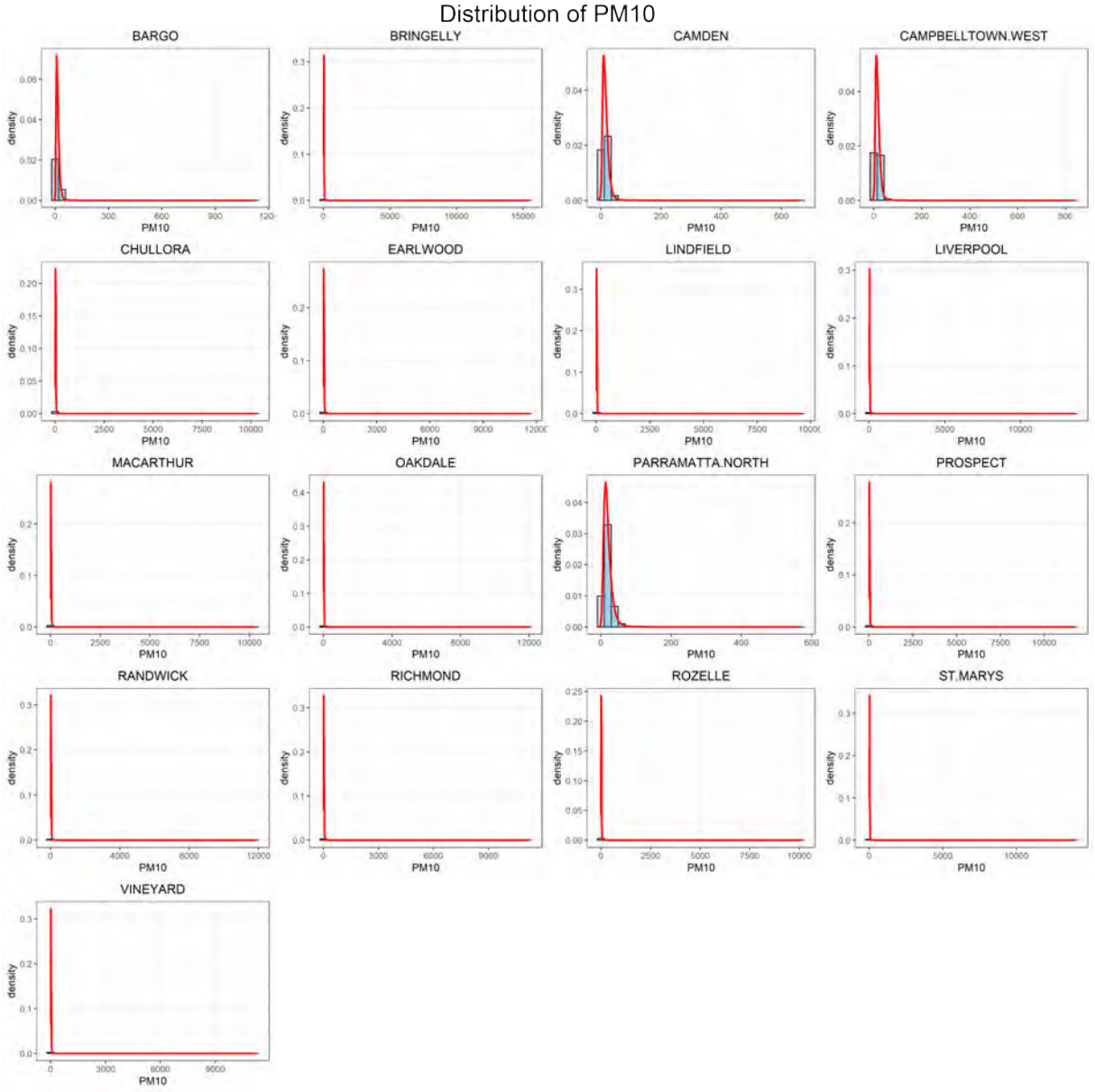


FIGURE A.32: Distribution of PM10 missing values

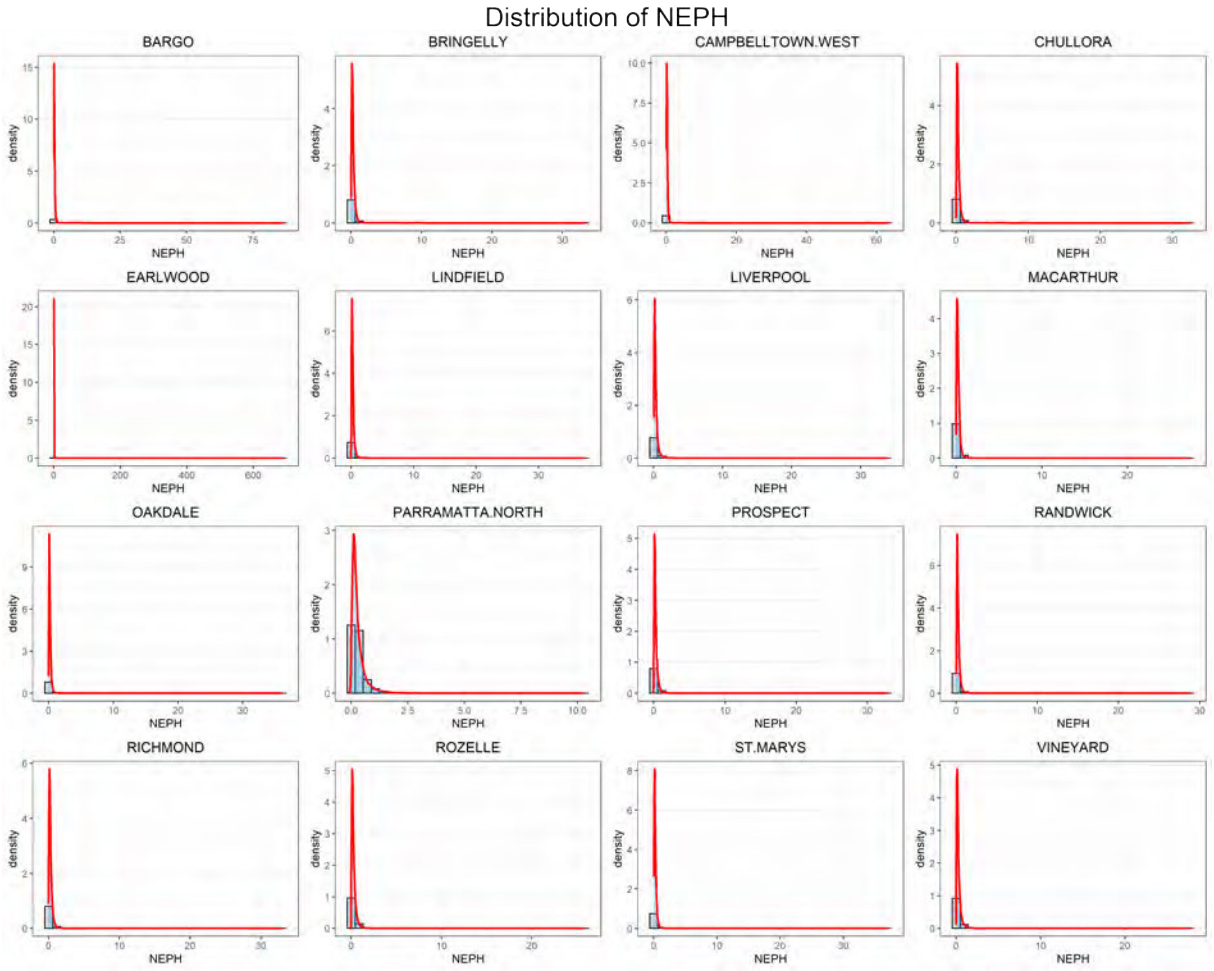


FIGURE A.33: Distribution of NEPH missing values

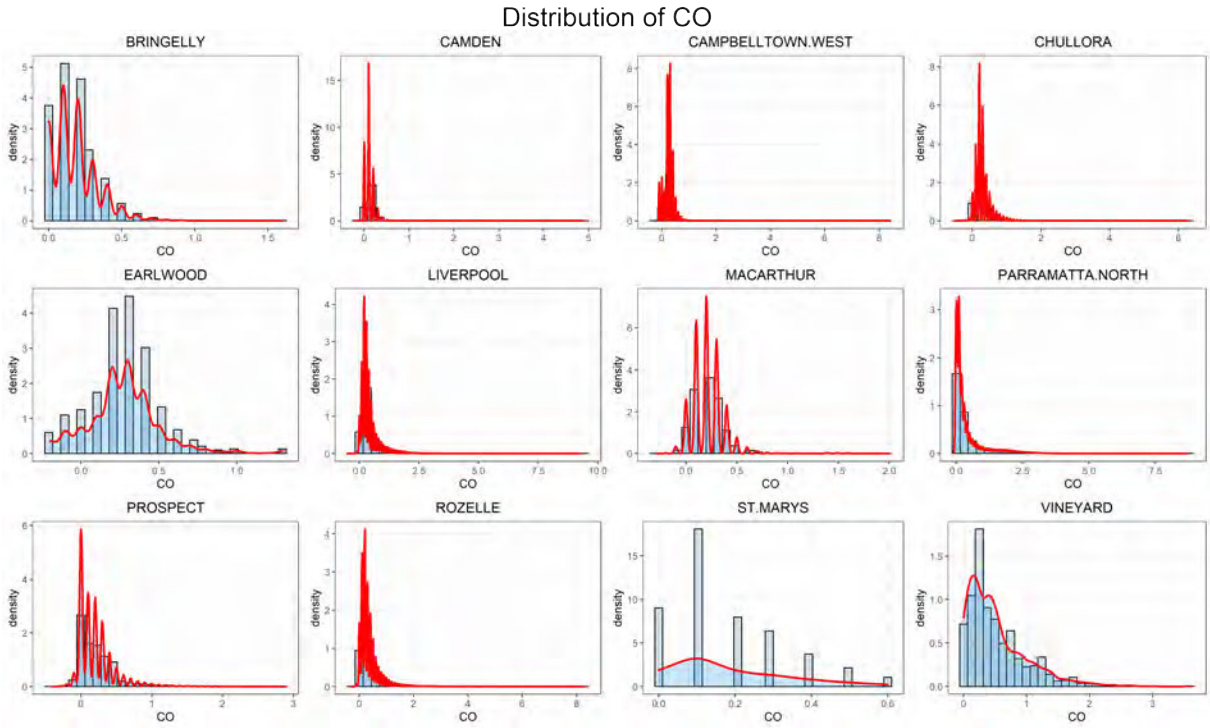


FIGURE A.34: Distribution of CO missing values

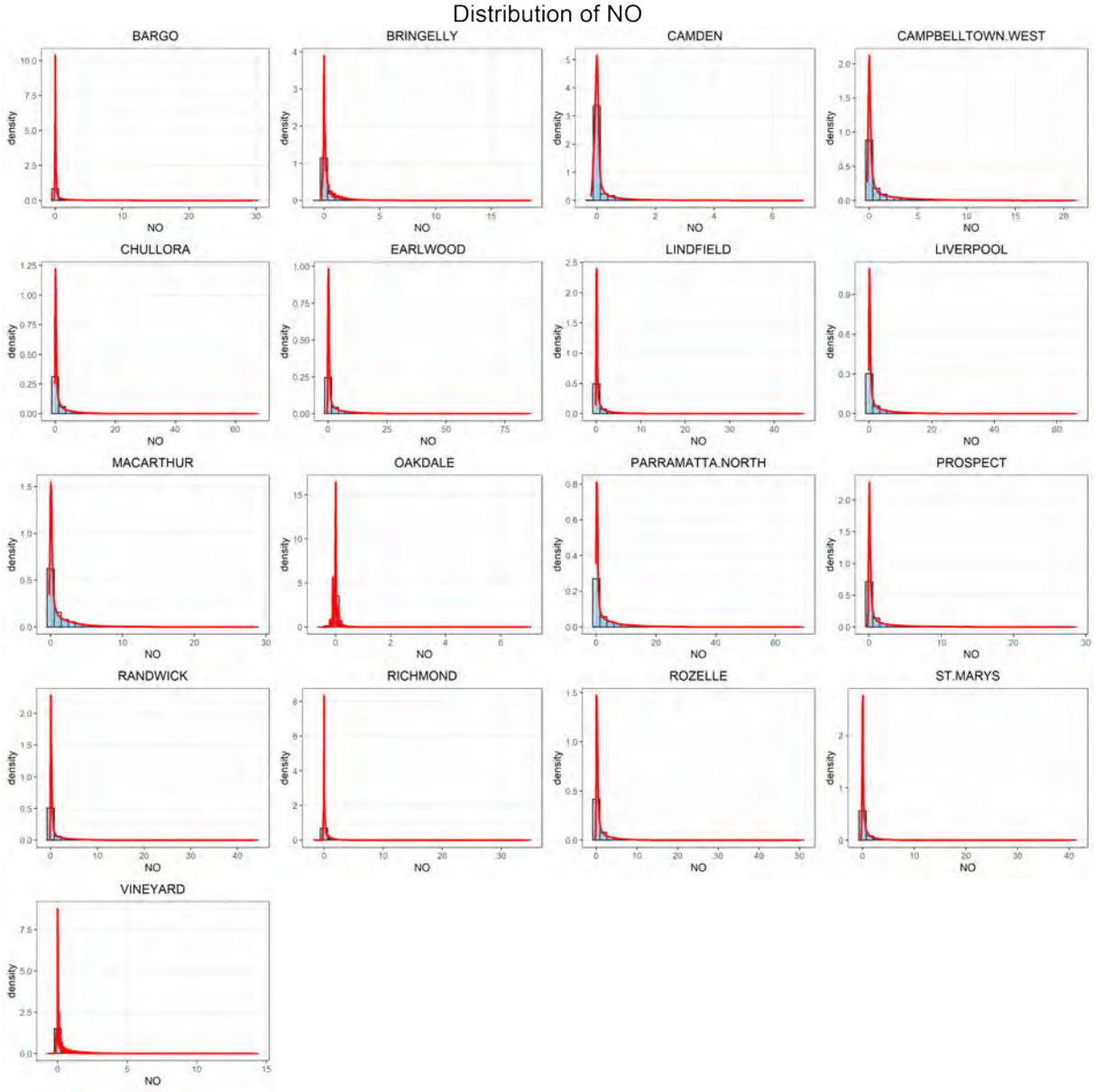


FIGURE A.35: Distribution of NO missing values

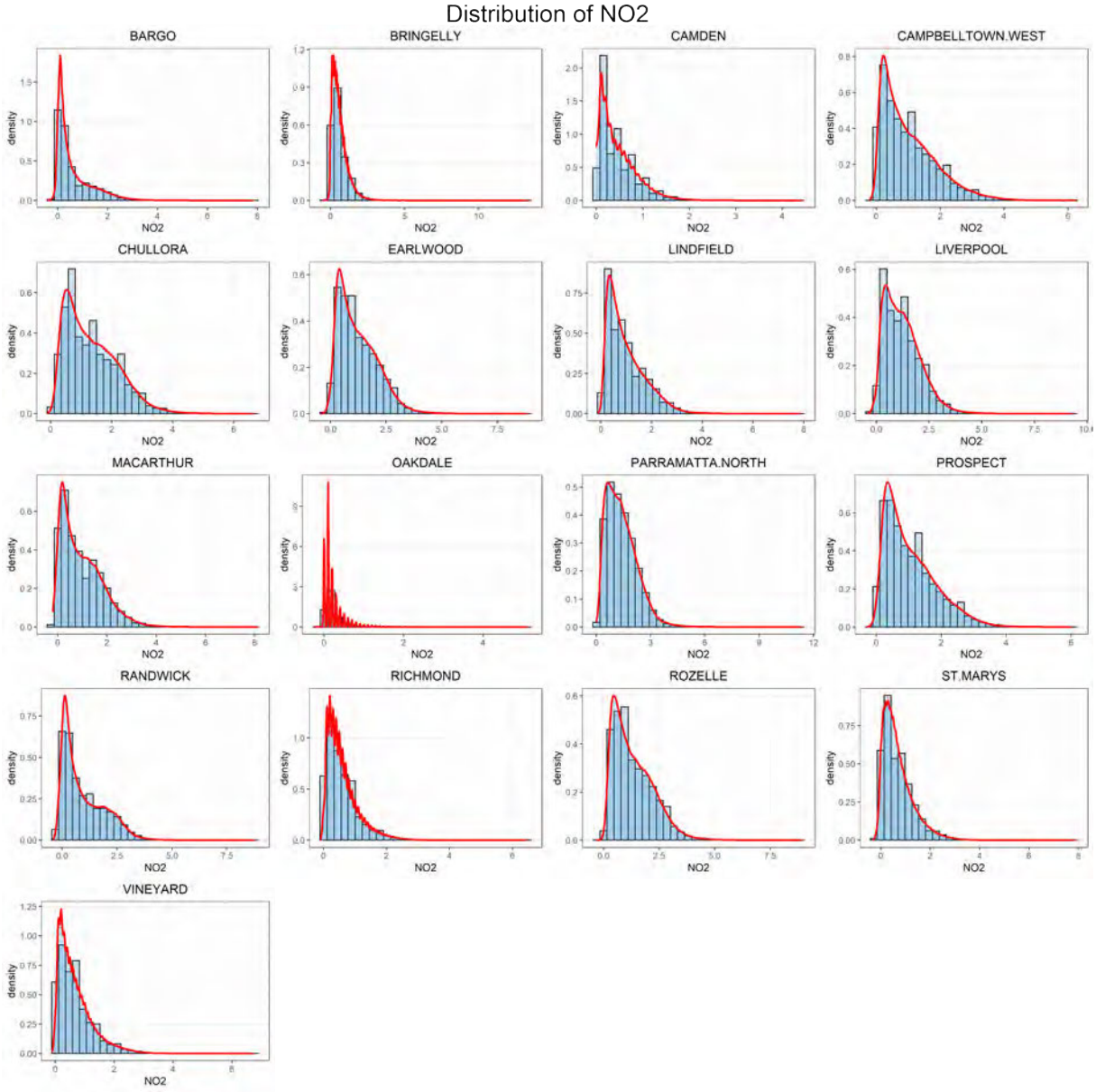


FIGURE A.36: Distribution of NO2 missing values

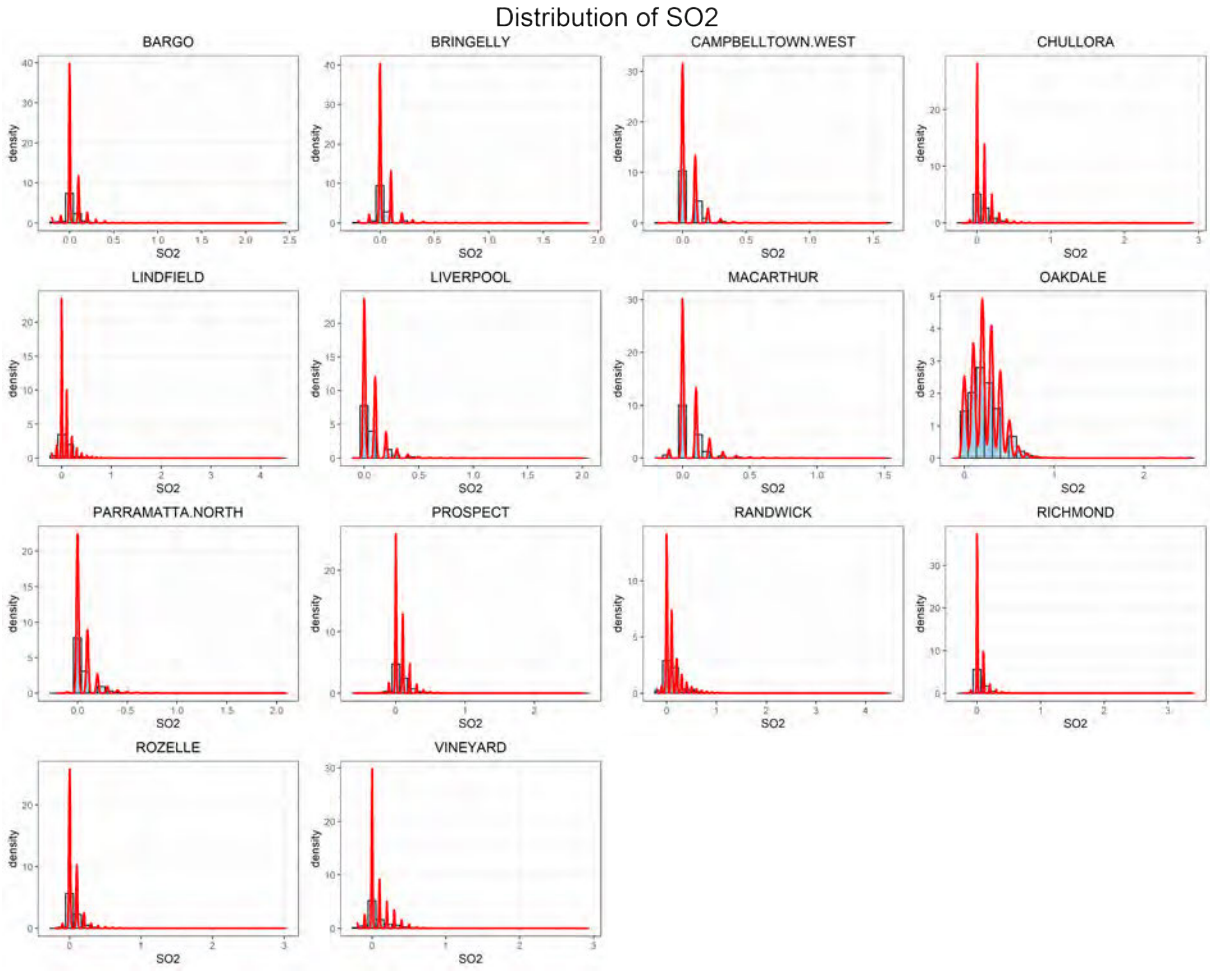


FIGURE A.37: Distribution of SO2 missing values

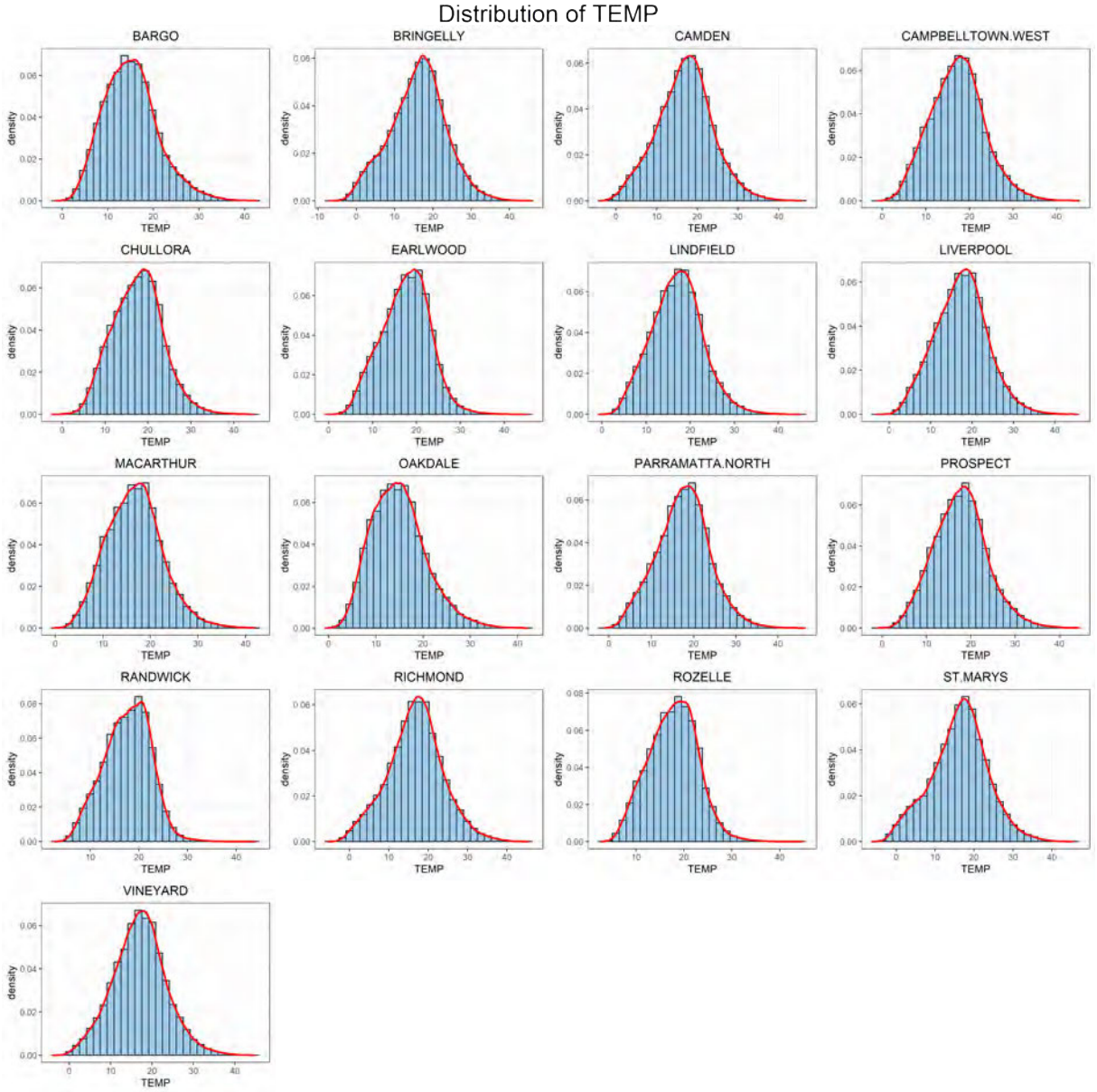


FIGURE A.38: Distribution of TEMP missing values

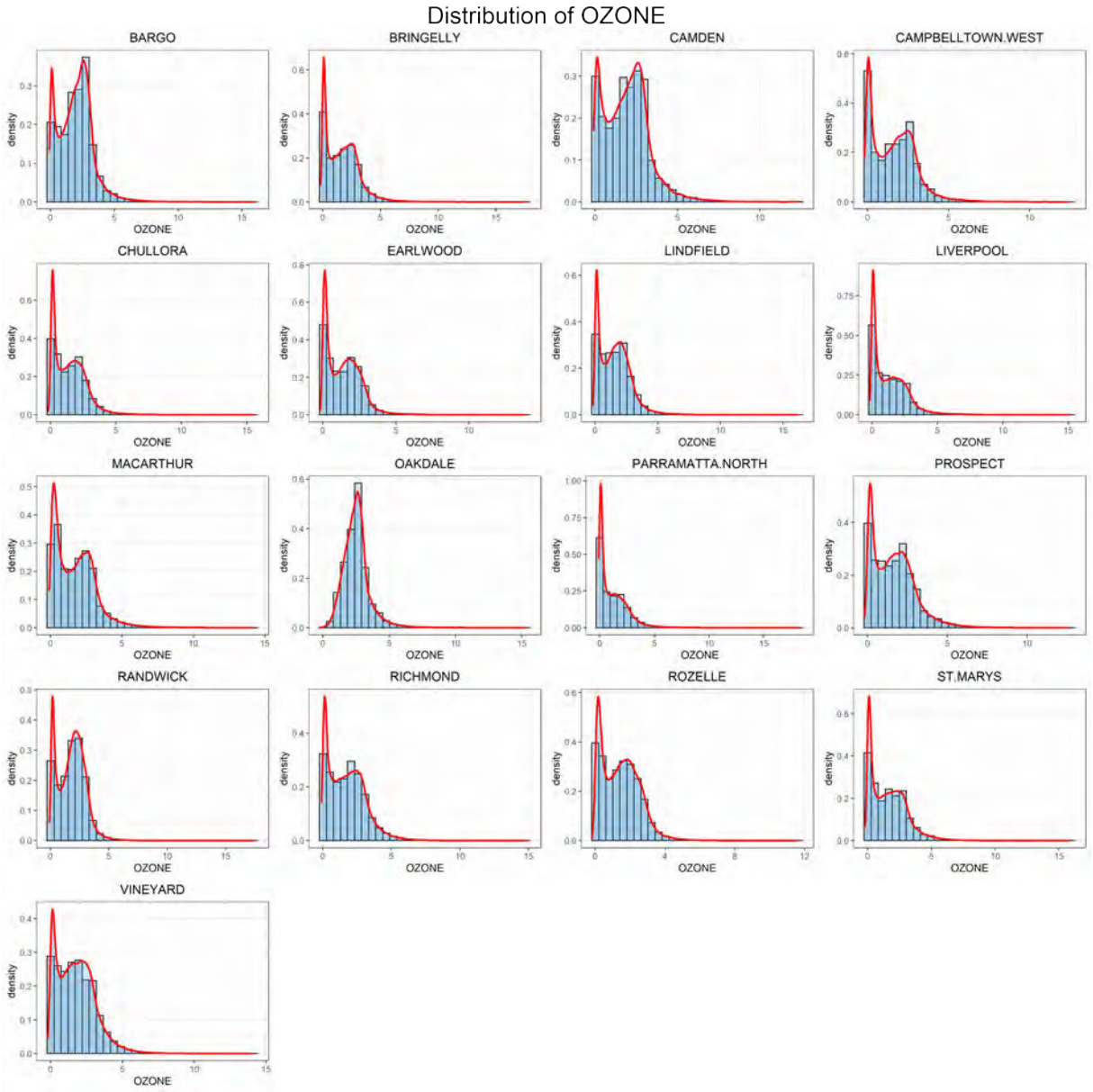


FIGURE A.39: Distribution of OZONE missing values

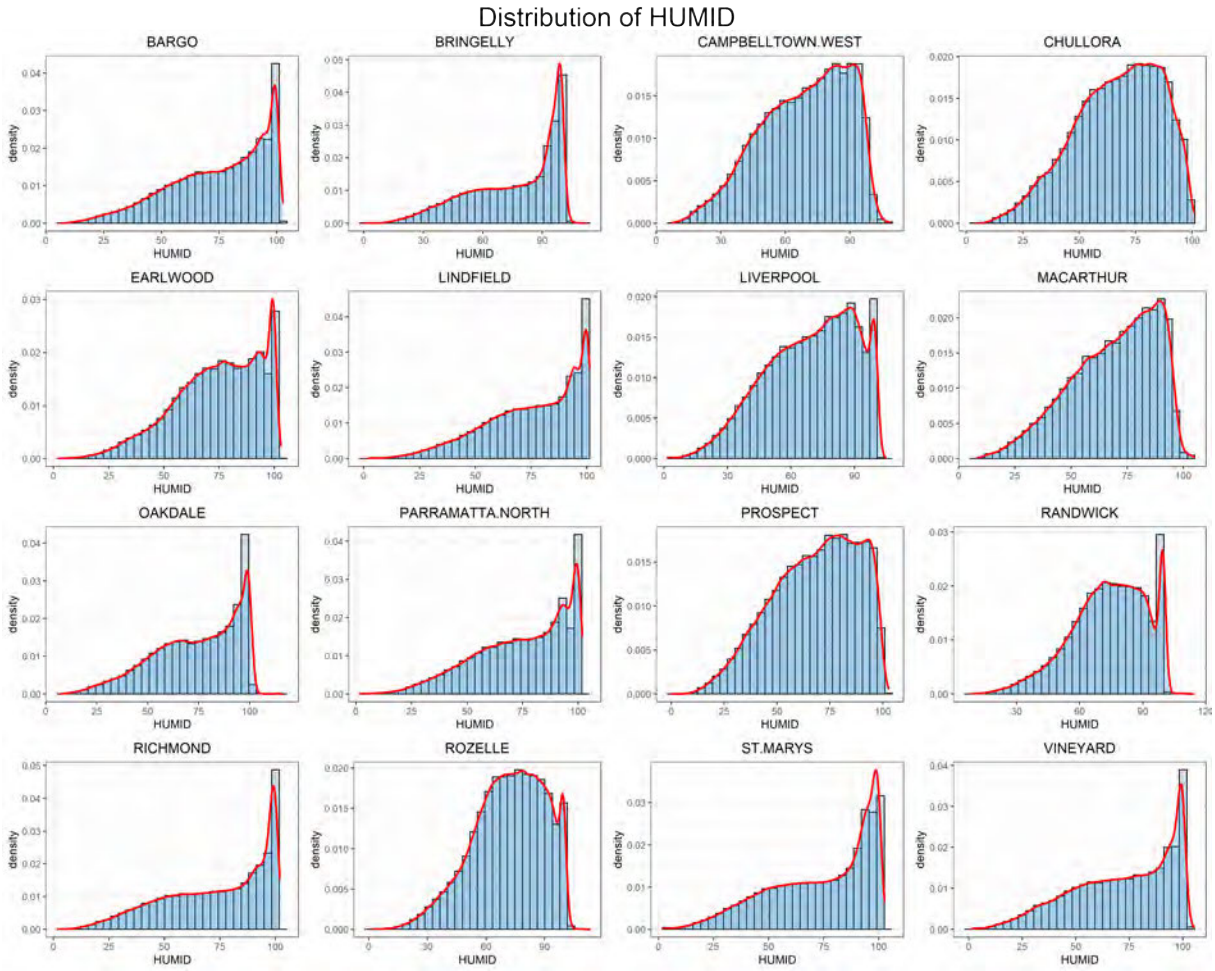


FIGURE A.40: Distribution of HUMID missing values

FIGURE A.41: Distribution of SOLAR missing values

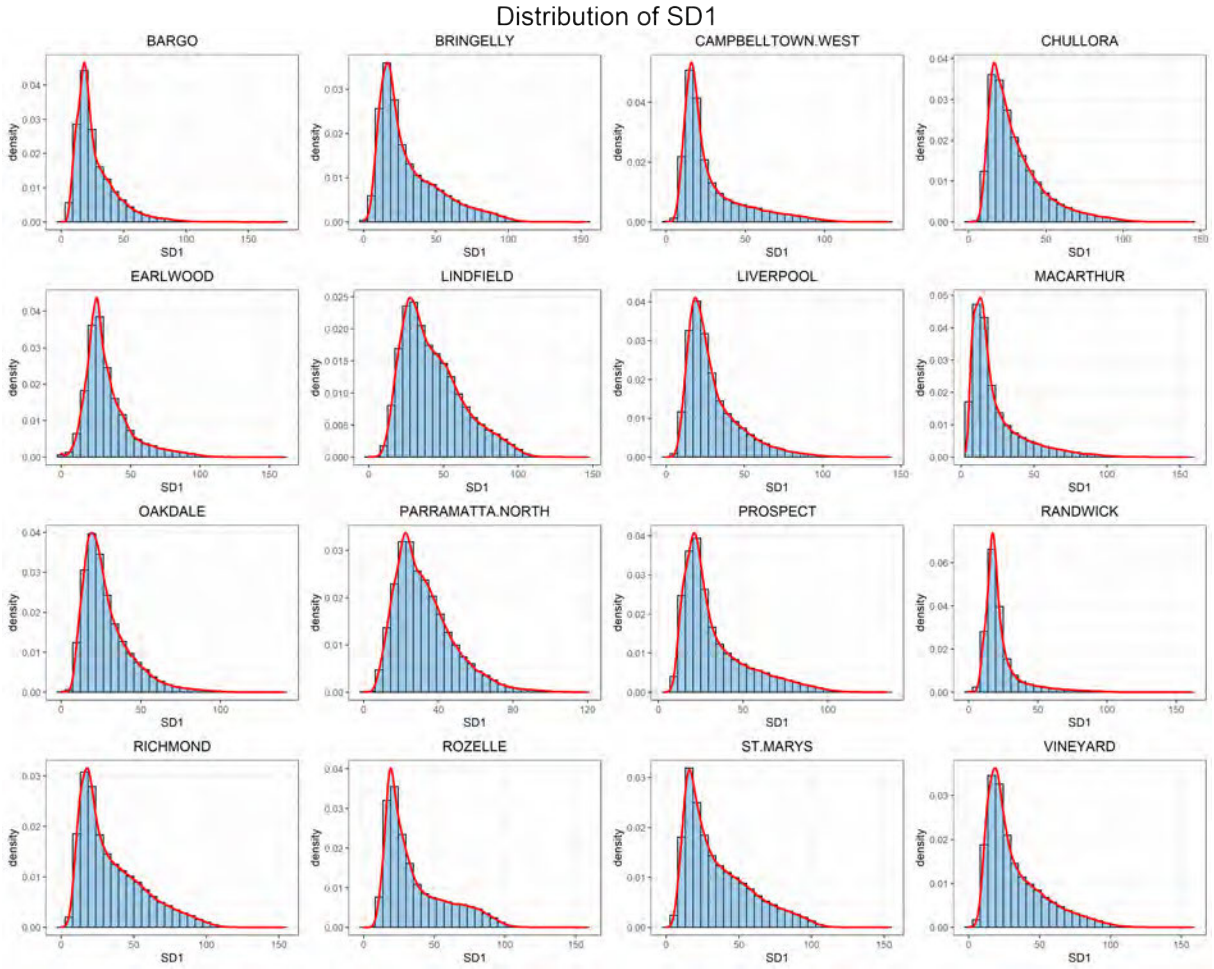


FIGURE A.42: Distribution of SD1 missing values

FIGURE A.43: Distribution of RAIN missing values

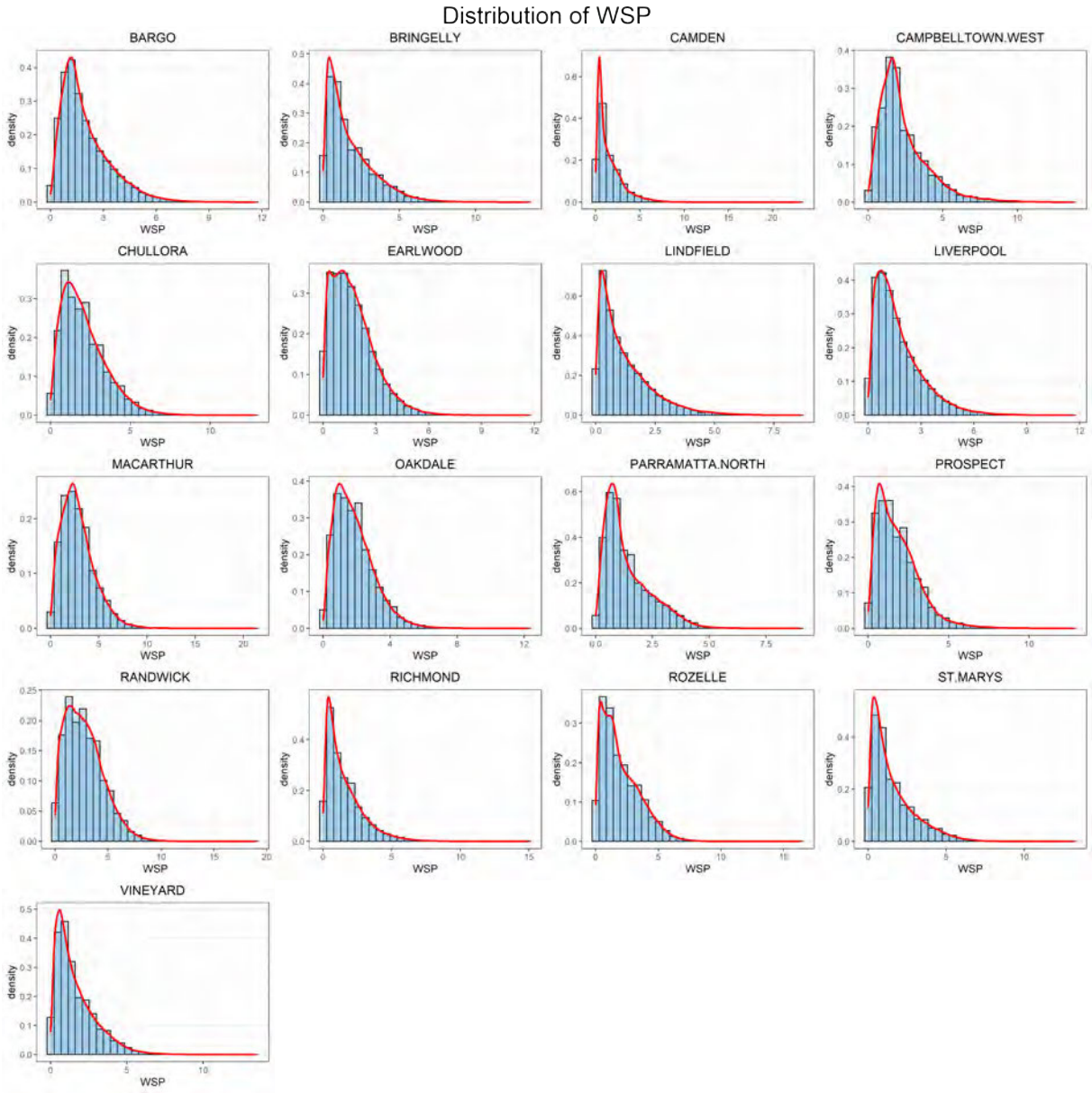


FIGURE A.44: Distribution of WSP missing values

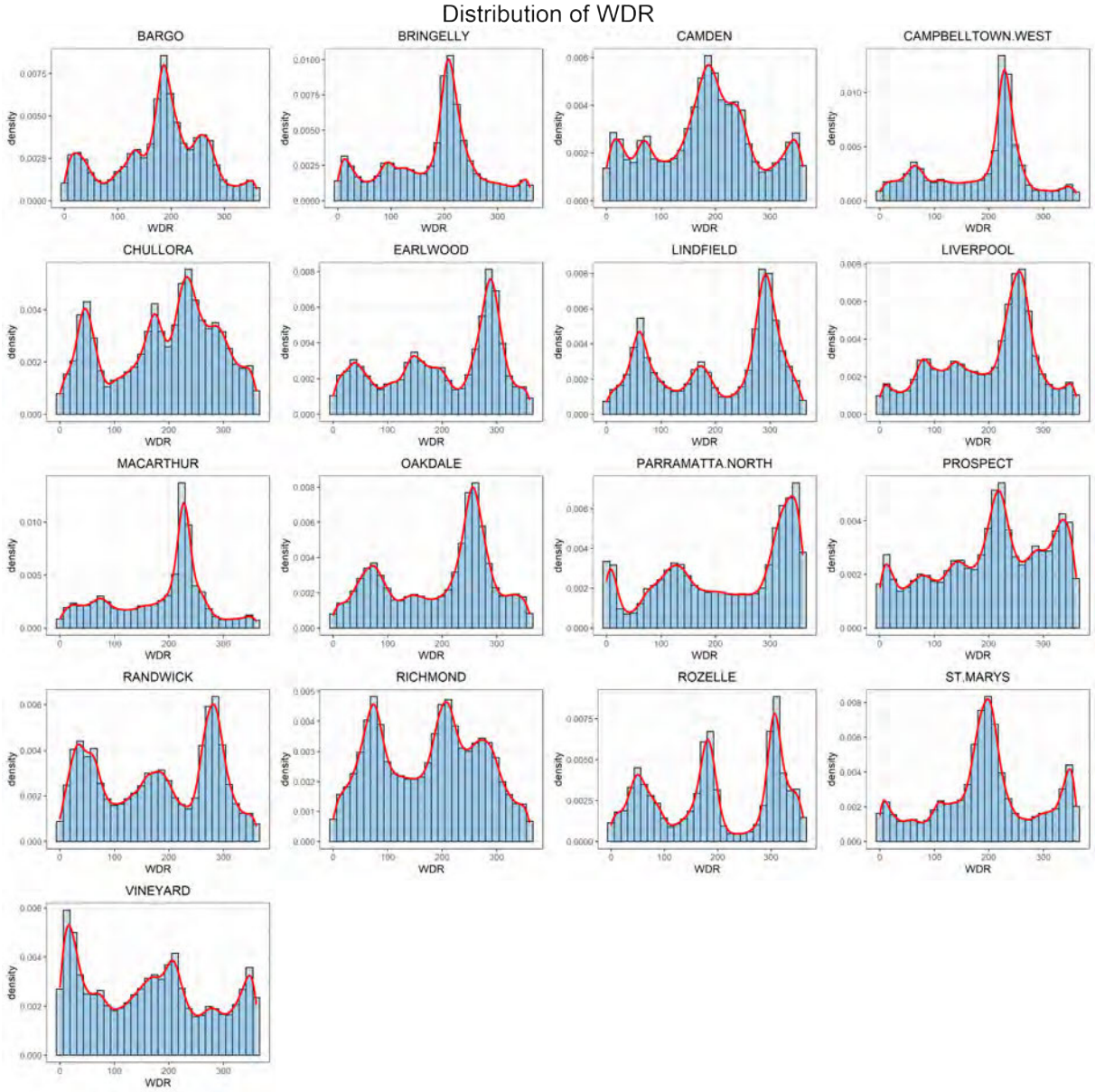


FIGURE A.45: Distribution of WDR missing values

Appendix B

Supplementary charts

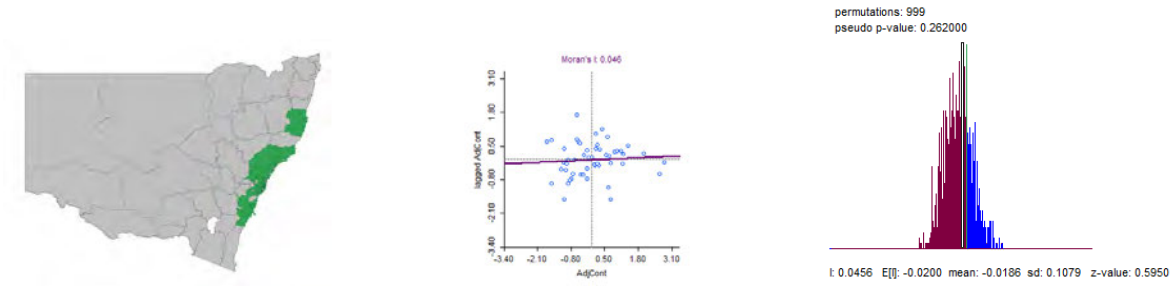


FIGURE B.1: (a) SA3 areas included in the model. (b) Moran scatter plot. Global Moran's I score is 0.046 (c) Distribution of Moran's I under the null hypothesis of spatial randomness.

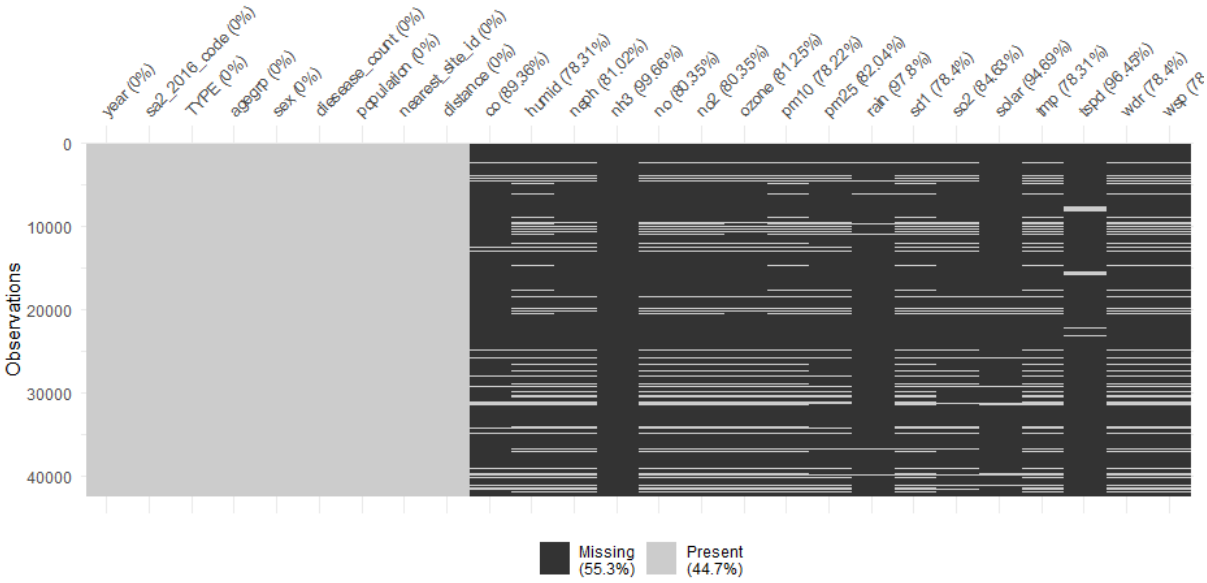


FIGURE B.2: Missing values in the dataset used for the analysis in chapter 9

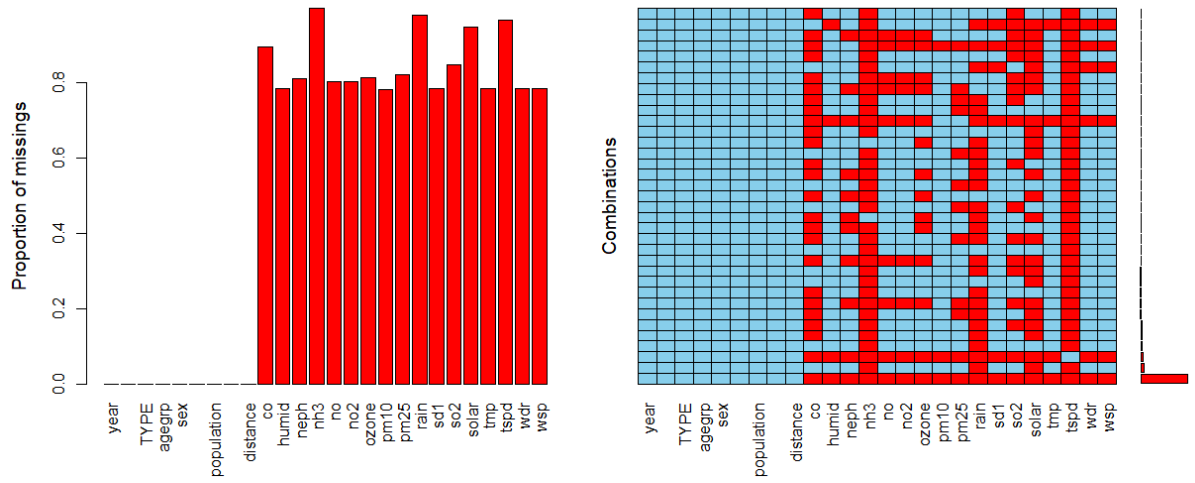


FIGURE B.3: Proportions and the combinations of missing values in the dataset used in chapter 9

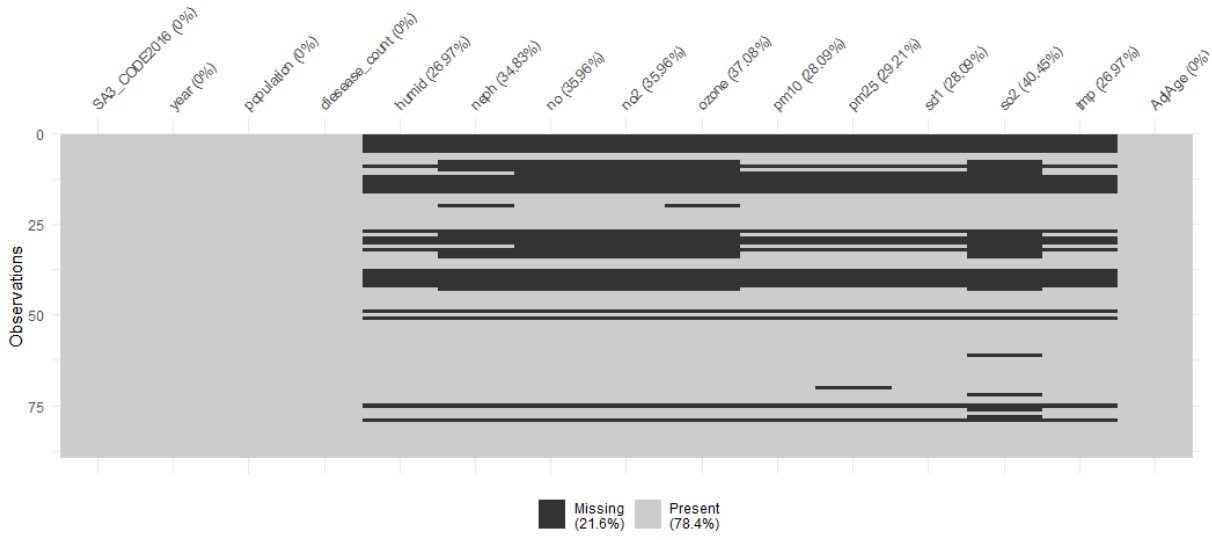


FIGURE B.4: Missing values in the dataset filtered for 2018 used for the analysis in chapter 9

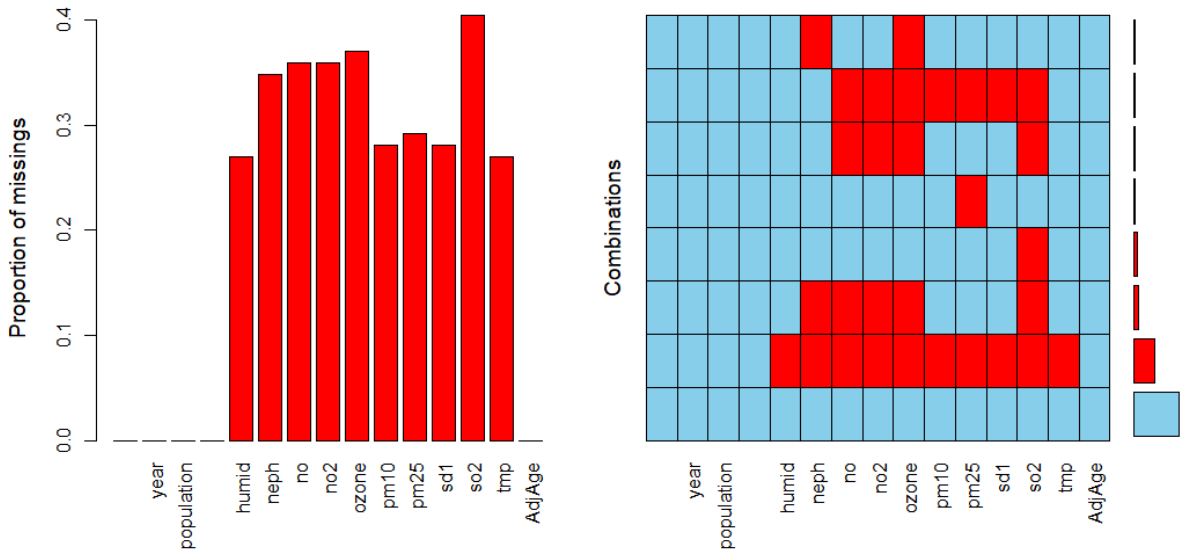


FIGURE B.5: Proportions and the combinations of missing values in the dataset filtered for 2018 used in chapter 9

	population	diesease_count	humid	neph	no	no2	ozone	pm10	pm25	sd1	so2	tmp	AdjAge	AdjCount
population	1.00	0.93	-0.13	0.03	-0.02	0.03	-0.13	0.24	0.11	-0.22	0.10	0.06	-0.29	-0.23
diesease_count	0.93	1.00	0.04	-0.03	-0.09	-0.06	-0.09	0.26	0.02	-0.31	0.18	0.04	-0.19	0.05
humid	-0.13	0.04	1.00	0.21	-0.07	0.02	0.06	0.34	0.07	-0.46	0.63	0.22	0.06	0.27
neph	0.03	-0.03	0.21	1.00	0.65	0.80	-0.48	0.64	0.83	-0.09	0.46	0.73	-0.53	-0.11
no	-0.02	-0.09	-0.07	0.65	1.00	0.93	-0.88	0.38	0.88	0.45	0.26	0.85	-0.14	-0.21
no2	0.03	-0.06	0.02	0.80	0.93	1.00	-0.82	0.51	0.91	0.27	0.38	0.86	-0.34	-0.24
ozone	-0.13	-0.09	0.06	-0.48	-0.88	-0.82	1.00	-0.37	-0.76	-0.47	-0.30	-0.83	0.02	0.23
pm10	0.24	0.26	0.34	0.64	0.38	0.51	-0.37	1.00	0.68	-0.49	0.86	0.58	-0.22	-0.04
pm25	0.11	0.02	0.07	0.83	0.88	0.91	-0.76	0.68	1.00	0.13	0.51	0.85	-0.29	-0.22
sd1	-0.22	-0.31	-0.46	-0.09	0.45	0.27	-0.47	-0.49	0.13	1.00	-0.55	0.21	0.14	-0.25
so2	0.10	0.18	0.63	0.46	0.26	0.38	-0.30	0.86	0.51	-0.55	1.00	0.54	-0.04	0.07
tmp	0.06	0.04	0.22	0.73	0.85	0.86	-0.83	0.58	0.85	0.21	0.54	1.00	-0.16	-0.16
AdjAge	-0.29	-0.19	0.06	-0.53	-0.14	-0.34	0.02	-0.22	-0.29	0.14	-0.04	-0.16	1.00	0.17
AdjCount	-0.23	0.05	0.27	-0.11	-0.21	-0.24	0.23	-0.04	-0.22	-0.25	0.07	-0.16	0.17	1.00

FIGURE B.6: Correlation matrix of the variables

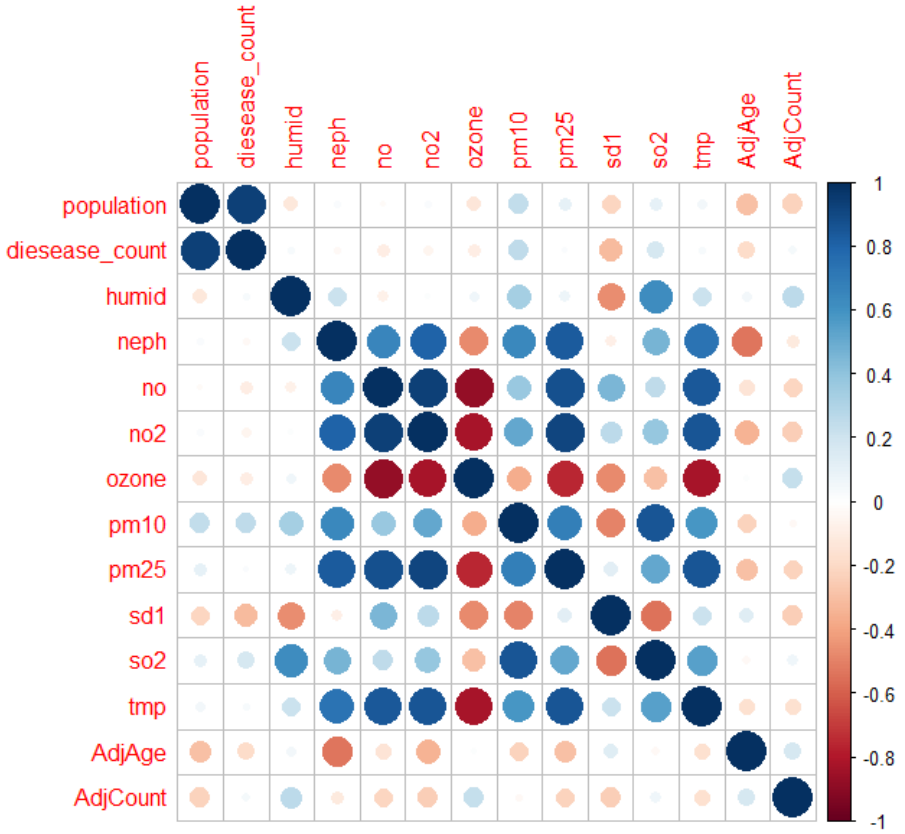


FIGURE B.7: Graphical representation of the correlation matrix of the variables

AdjAge	humid	neph	no	no2	ozone	pm10	pm25	sd1	so2	tmp
2.313320	2.476768	10.908790	25.684796	23.825316	5.324067	10.354392	16.684239	3.775682	8.079908	6.777995

FIGURE B.8: Variance inflation factors of the predictors

```
Call:
glm(formula = diesease_count ~ AdjAge + humid + neph + no + no2 +
    ozone + pm10 + pm25 + sd1 + so2 + tmp + offset(log(population)),
    family = poisson(link = "log"), data = SA3_2018)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.51163  -0.61414  -0.06331   0.53201   2.19799

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.395832   2.030570  -2.657  0.00788 **
AdjAge       0.005007   0.014393   0.348  0.72794
humid        0.020035   0.013519   1.482  0.13835
neph         3.178048   2.250218   1.412  0.15785
no           0.361554   0.187715   1.926  0.05409 .
no2         -0.658889   0.368683  -1.787  0.07391 .
ozone       -0.290895   0.331085  -0.879  0.37961
pm10         0.017866   0.025181   0.709  0.47802
pm25        -0.164567   0.082970  -1.983  0.04732 *
sd1         -0.010362   0.006025  -1.720  0.08548 .
so2         -0.211760   1.425270  -0.149  0.88189
tmp         -0.038215   0.098298  -0.389  0.69745
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 61.900  on 50  degrees of freedom
Residual deviance: 33.303  on 39  degrees of freedom
AIC: 307.27

Number of Fisher Scoring iterations: 4
```

FIGURE B.9: Poisson regression model output in R software

```
Call:
glm(formula = diesease_count ~ AdjAge + humid + neph + no + no2 +
    ozone + pm10 + pm25 + sd1 + so2 + tmp + offset(log(population)),
    family = "quasipoisson", data = SA3_2018)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.51163  -0.61414  -0.06331   0.53201   2.19799

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.395832   1.927632  -2.799  0.00792 **
AdjAge       0.005007   0.013663   0.366  0.71602
humid        0.020035   0.012834   1.561  0.12658
neph         3.178048   2.136145   1.488  0.14486
no           0.361554   0.178199   2.029  0.04933 *
no2         -0.658889   0.349992  -1.883  0.06723 .
ozone       -0.290895   0.314301  -0.926  0.36038
pm10         0.017866   0.023905   0.747  0.45932
pm25        -0.164567   0.078764  -2.089  0.04325 *
sd1         -0.010362   0.005720  -1.812  0.07775 .
so2         -0.211760   1.353017  -0.157  0.87644
tmp         -0.038215   0.093315  -0.410  0.68440
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 0.9011814)

    Null deviance: 61.900  on 50  degrees of freedom
Residual deviance: 33.303  on 39  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

FIGURE B.10: Quasi-Poisson regression model output in R software

```
diesease_count ~ humid + no + no2 + pm25 + sd1 + offset(log(population))

         Df Deviance    AIC
<none>        37.040 299.00
- no2     1   39.134 299.10
- pm25    1   39.475 299.44
+ neph    1   35.627 299.59
+ pm10    1   35.810 299.77
+ ozone   1   36.810 300.77
+ tmp     1   36.888 300.85
+ AdjAge  1   36.896 300.86
+ so2     1   36.971 300.93
- no      1   41.431 301.39
- humid   1   41.475 301.44
- sd1     1   44.140 304.10

Call:  glm(formula = diesease_count ~ humid + no + no2 + pm25 + sd1 +
    offset(log(population)), family = poisson(link = "log"),
    data = SA3_2018)

Coefficients:
(Intercept)        humid           no          no2         pm25          sd1
   -5.96505      0.02013      0.28228     -0.34061     -0.09114     -0.01276

Degrees of Freedom: 50 Total (i.e. Null);  45 Residual
Null Deviance:      61.9
Residual Deviance: 37.04        AIC: 299
```

```
diesease_count ~ humid + offset(log(population))

         Df Deviance    AIC
<none>        47.619 305.44
+ no2     1   44.309 306.07
+ sd1     1   44.515 306.27
+ pm25    1   44.858 306.62
+ no      1   45.308 307.06
+ AdjAge  1   45.894 307.65
+ tmp     1   46.088 307.85
+ neph    1   46.182 307.94
+ ozone   1   46.769 308.53
+ so2     1   47.568 309.33
+ pm10    1   47.580 309.34
- humid   1   61.900 315.79

Call:  glm(formula = diesease_count ~ humid + offset(log(population)),
    family = poisson(link = "log"), data = SA3_2018)

Coefficients:
(Intercept)        humid
   -8.11904      0.03315

Degrees of Freedom: 50 Total (i.e. Null);  49 Residual
Null Deviance:      61.9
Residual Deviance: 47.62        AIC: 301.6
```

FIGURE B.11: (a) Best subset of variables using AIC (b) Best subset of variables using BIC

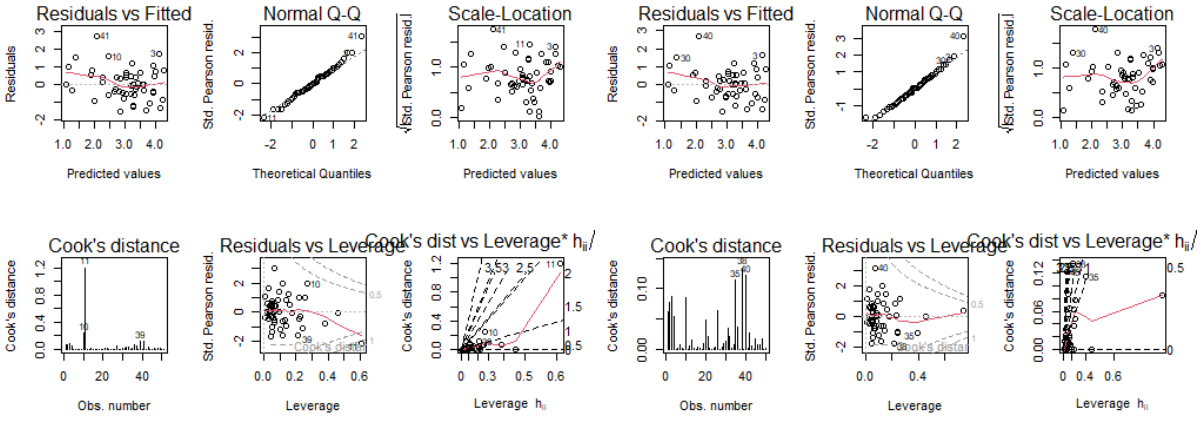


FIGURE B.12: (a) Quasi-Poisson model diagnostic plots (b)Quasi-Poisson model diagnostic plots after removing the influential observation

```
Call:
glm(formula = diesease_count ~ humid + no + no2 + pm25 + sd1 +
    offset(log(population)), family = "quasipoisson", data = SA3_2018[-11,
    ])

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.58875  -0.52378   0.01739   0.54206   2.40123

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.877944   0.934900  -5.218 4.68e-06 ***
humid        0.008626   0.009713   0.888  0.37937
no           0.396910   0.127087   3.123  0.00316 **
no2         -0.544789   0.225759  -2.413  0.02005 *
pm25        -0.119911   0.053355  -2.247  0.02968 *
sd1         -0.012302   0.004286  -2.870  0.00628 **
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 0.7897901)

    Null deviance: 56.915  on 49  degrees of freedom
Residual deviance: 33.071  on 44  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

FIGURE B.13: Best Quasi-Poisson regression model output in R software (using AIC and after removing influential points)

```
Call:
glm(formula = diesease_count ~ no + no2 + pm25 + sd1 + offset(log(population)),
    family = "quasipoisson", data = SA3_2018[-11, ])

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5984  -0.5799   0.0203   0.5697   2.4101

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.143590   0.432917  -9.571 2.01e-12 ***
no           0.432640   0.119915   3.608 0.000771 ***
no2         -0.600537   0.215320  -2.789 0.007722 **
pm25        -0.130084   0.052011  -2.501 0.016089 *
sd1         -0.013533   0.004055  -3.338 0.001703 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 0.7871706)

    Null deviance: 56.915  on 49  degrees of freedom
Residual deviance: 33.701  on 45  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

FIGURE B.14: Best Quasi-Poisson regression model output in R software (using AIC and after removing influential points and after removing humidity variable)

```
Call:
glm(formula = diesease_count ~ humid + offset(log(population)),
    family = "quasipoisson", data = SA3_2018)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.90804  -0.70189  -0.02492   0.69871   2.35757

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.119043   0.605991 -13.398  < 2e-16 ***
humid        0.033155   0.008884   3.732 0.000494 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.004864)

    Null deviance: 61.900  on 50  degrees of freedom
Residual deviance: 47.619  on 49  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

FIGURE B.15: Best Quasi-Poisson regression model output in R software (using BIC)

Bibliography

Abraham, S., & Li, X. (2014). A cost-effective wireless sensor network system for indoor air quality monitoring applications. *Procedia Computer Science*, *34*, 165–171.

Abril, J. C. (2011). Structural time series models. In M. Lovric (Ed.), *International encyclopedia of statistical science* (pp. 1555–1558). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_577

Aghabozorgi, S., Shirkhorshidi, A. S., & Wah, T. Y. (2015). Time-series clustering - a decade review. *Information systems*, *53*, 16–38.

Anselin, L. (1994). Exploratory spatial data analysis and geographic information systems. *New tools for spatial analysis*, *17*, 45–54.

Anselin, L. (1995). Local indicators of spatial association-LISA. *Geographical analysis*, *27*(2), 93–115.

Anselin, L. (2019). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In *Spatial analytical perspectives on GIS* (pp. 111–126). Routledge.

Anselin, L., Syabri, I., & Kho, Y. (2006). GeoDa : An introduction to spatial data analysis. *Geographical Analysis*, *38*(1), 5–22. https://doi.org/https://doi.org/10.1111/j.0016-7363.2005.00671.x

Australian Bureau of Statistics. (2017). Population estimates by age and sex, regions of New South Wales (ASGS 2016) [data set] [https://www.abs.gov.au/AUSSTATS/abs@.nsf/DetailsPage/3235.02016?OpenDocument, Last accessed on 2020-09-30].

Australian Bureau of Statistics. (2021). Digital boundary files [data set] [https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standa

rd-asgs-edition-3/jul2021-jun2026/access-and-downloads/digital-boundary-files, Last accessed on 2022-01-30].

Bénié, G., Kaboré, S., Goita, K., & Courel, M.-F. (2005). Remote sensing-based spatio-temporal modeling to predict biomass in Sahelian grazing ecosystem. *Ecological Modelling*, *184*(2-4), 341–354.

Blauw, L. L., Aziz, N. A., Tannemaat, M. R., Blauw, C. A., de Craen, A. J., Pijl, H., & Rensen, P. C. (2017). Diabetes incidence and glucose intolerance prevalence increase with higher outdoor temperature. *BMJ Open Diabetes Research and Care*, *5*(1), e000317.

Booth, G. L., Luo, J., Park, A. L., Feig, D. S., Moineddin, R., & Ray, J. G. (2017). Influence of environmental temperature on risk of gestational diabetes. *Cmaj*, *189*(19), E682–E689.

Bousquet, J., Anto, J. M., Annesi-Maesano, I., Dedeu, T., Dupas, E., Pépin, J.-L., Eyindanga, L. S. Z., Arnavielhe, S., Ayache, J., Basagana, X., et al. (2018). POLLAR: impact of air POLLution on Asthma and Rhinitis; a European institute of innovation and technology health (EIT health) project. *Clinical and translational allergy*, *8*(1), 1–13.

Bowe, B., Xie, Y., Li, T., Yan, Y., Xian, H., & Al-Aly, Z. (2018). The 2016 global and national burden of diabetes mellitus attributable to PM2.5 air pollution. *The Lancet Planetary Health*, *2*(7), e301–e312.

Breysse, P. N., Diette, G. B., Matsui, E. C., Butz, A. M., Hansel, N. N., & McCormack, M. C. (2010). Indoor air pollution and asthma in children. *Proceedings of the American Thoracic Society*, *7*(2), 102–106.

Chandrasekaran, S., Zaefferer, M., Moritz, S., Stork, J., Friese, M., Fischbach, A., & Bartz-Beielstein, T. (2016). Data preprocessing: A new algorithm for univariate imputation designed specifically for industrial needs. *PROCEEDINGS 26. WORKSHOP COMPUTATIONAL INTELLIGENCE*.

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2020). *shiny: Web Application Framework for R* [R package version 1.5.0].

Chen, H., Burnett, R. T., Kwong, J. C., Villeneuve, P. J., Goldberg, M. S., Brook, R. D., van Donkelaar, A., Jerrett, M., Martin, R. V., Brook, J. R., et al. (2013). Risk of incident diabetes in relation to long-term exposure to fine particulate matter in Ontario, Canada. *Environmental health perspectives*, *121*(7), 804–810.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). STL: A seasonal-trend decomposition. *Journal of official statistics*, *6*(1), 3–73.

Cliff, A. D., Ord, J., Haggett, P., Versey, G., et al. (1981). *Spatial diffusion: An historical geography of epidemics in an island community* (Vol. 14). CUP Archive.

Colagiuri, R., Director. (2013). Diabetes and climate change: Different drums-same orchestra. *Journal of public health policy*, *34*(1), 165–169.

Comer, K. F., Grannis, S., Dixon, B. E., Bodenhamer, D. J., & Wiehe, S. E. (2011). Incorporating geospatial capacity within clinical data systems to address social determinants of health. *Public health reports*, *126*(3_suppl), 54–61.

Compieta, P., Di Martino, S., Bertolotto, M., Ferrucci, F., & Kechadi, T. (2007). Exploratory spatio-temporal data mining and visualization. *Journal of Visual Languages & Computing*, *18*(3), 255–279.

Coogan, P. F., White, L. F., Jerrett, M., Brook, R. D., Su, J. G., Seto, E., Burnett, R., Palmer, J. R., & Rosenberg, L. (2012). Air pollution and incidence of hypertension and diabetes mellitus in black women living in Los Angeles. *Circulation*, *125*(6), 767–772.

Dadvand, P., Rushton, S., Diggle, P. J., Goffe, L., Rankin, J., & Pless-Mulloli, T. (2011). Using spatio-temporal modeling to predict long-term exposure to black smoke at fine spatial and temporal scale. *Atmospheric Environment*, *45*(3), 659–664.

Dendup, T., Feng, X., Clingan, S., & Astell-Burt, T. (2018). Environmental risk factors for developing type 2 diabetes mellitus: A systematic review. *International journal of environmental research and public health*, *15*(1), 78.

Dixon, M. F., Polson, N. G., & Sokolov, V. O. (2019). Deep learning for spatio-temporal modeling: Dynamic traffic flows and high frequency trading. *Applied Stochastic Models in Business and Industry*, *35*(3), 788–807.

Dixon, W. J. (1988). *BMDP statistical software manual to accompany the 1988 software release*. University of California Press.

Donaldson, K., Ian Gilmour, M., & MacNee, W. (2000). Asthma and PM10. *Respiratory research*, *1*(1), 12–15.

Donders, A. R. T., Van Der Heijden, G. J., Stijnen, T., & Moons, K. G. (2006). A gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*(10), 1087–1091.

Duc, H., Salter, D., Azzi, M., Jiang, N., Warren, L., Watt, S., Riley, M., White, S., Trieu, T., Tzu-Chi Chang, L., et al. (2021). The effect of lockdown period during the COVID-19 pandemic on air quality in Sydney region, Australia. *International Journal of Environmental Research and Public Health*, *18*(7), 3528.

Environment Protection Authority Victoria. (2021). PM10 particles in the air [https://www.epa.vic.gov.au/for-community/environmental-information/air-quality/pm10-particles-in-the-air, Last accessed on 2022-09-22].

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics New York.

The global asthma report 2018. (2018).

Guardian, T. (2012, November). Sustainable business: Ann keeling - diabetes and climate change, making the links.

Guarnieri, M., & Balmes, J. R. (2014). Outdoor air pollution and asthma. *The Lancet*, *383*(9928), 1581–1592.

Hansen, A. B., Ravnskjær, L., Loft, S., Andersen, K. K., Bräuner, E. V., Baastrup, R., Yao, C., Ketzel, M., Becker, T., Brandt, J., et al. (2016). Long-term exposure to fine particulate matter and incidence of diabetes in the Danish nurse cohort. *Environment international*, *91*, 243–250.

Howard, J. (2017). Is there a link between climate change and diabetes? [https://edition.cnn.com/2017/03/20/health/climate-change-type-2-diabetes-study/index.html, Last accessed on 2019-07-29].

Hryniewicz, O., & Kaczmarek, K. (2016). Bayesian analysis of time series using granular computing approach. *Applied Soft Computing*, *47*, 644–652.

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., & Wang, E. (2020). Package 'forecast'. *Online] https://cran. r-project. org/web/packages/forecast/forecast. pdf*.

International Diabetes Federation. (2012). *Diabetes and climate change report* (tech. rep.). International Diabetes Federation.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013a). *An introduction to statistical learning* (Vol. 112). Springer.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013b). *An introduction to statistical learning*. Springer. https://doi.org/10.1007/978-1-0716-1418-1_12

Jerrett, M., Brook, R., White, L. F., Burnett, R. T., Yu, J., Su, J., Seto, E., Marshall, J., Palmer, J. R., Rosenberg, L., et al. (2017). Ambient ozone and incident diabetes: A prospective analysis in a large cohort of African American women. *Environment international*, *102*, 42–47.

Joshi, S. K., & Shrestha, S. (2010). Diabetes mellitus: A review of its associations with different environmental factors. *Kathmandu University medical journal*, *8*(1), 109–115.

Jouanna, J. (2012). Water, health and disease in the Hippocratic treatise airs, waters, places. In *Greek medicine from hippocrates to galen* (pp. 155–172). Brill.

Junger, W., & De Leon, A. P. (2015). Imputation of missing data in time series for air pollutants. *Atmospheric Environment*, *102*, 96–104.

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., & Kolehmainen, M. (2004). Methods for imputation of missing values in air quality data sets. *Atmospheric Environment*, *38*(18), 2895–2907.

Kamarianakis, Y., & Prastacos, P. (2003). Spatial time series modeling: A review of the proposed methodologies. *The Regional Economics Applications Laboratory*.

Khagayi, S., Amek, N., Bigogo, G., Odhiambo, F., & Vounatsou, P. (2017). Bayesian spatio-temporal modeling of mortality in relation to malaria incidence in Western Kenya. *Plos one*, *12*(7), e0180516.

Kihal-Talantikite, W., Legendre, P., Le Nouveau, P., & Deguen, S. (2019). Premature adult death and equity impact of a reduction of NO2, PM10, and PM2. 5 levels

in Paris - a health impact assessment study conducted at the census block level. *International journal of environmental research and public health*, *16*(1), 38.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kline, D. M., & Berardi, V. L. (2005). Revisiting squared-error and cross-entropy functions for training neural network classifiers. *Neural Computing & Applications*, *14*(4), 310–318.

Koenig, J. Q. (1999). Air pollution and asthma. *Journal of allergy and clinical immunology*, *104*(4), 717–722.

Krämer, U., Herder, C., Sugiri, D., Strassburger, K., Schikowski, T., Ranft, U., & Rathmann, W. (2010). Traffic-related air pollution and incident type 2 diabetes: Results from the SALIA cohort study. *Environmental health perspectives*, *118*(9), 1273–1279.

Law, J., Quick, M., & Chan, P. (2014). Bayesian spatio-temporal modeling for analysing local patterns of crime over time at the small-area level. *Journal of quantitative criminology*, *30*, 57–78.

Lee, B.-J., Kim, B., & Lee, K. (2014). Air pollution exposure and cardiovascular disease. *Toxicological research*, *30*(2), 71–75.

Lee, C. M. Y., Colagiuri, R., Magliano, D. J., Cameron, A. J., Shaw, J., Zimmet, P., & Colagiuri, S. (2013). The cost of diabetes in adults in Australia. *Diabetes research and clinical practice*, *99*(3), 385–390. https://doi.org/10.1016/j.diabres.2012.12.002

Lee, J., & Wong, D. W. (2001). *Statistical analysis with ArcView GIS*. John Wiley & Sons.

Lei, K. S., & Wan, F. (2010). Pre-processing for missing data: A hybrid approach to air pollution prediction in Macau. *2010 IEEE International Conference on Automation and Logistics*, 418–422. https://doi.org/10.1109/ICAL.2010.5585320

Ling, C., & Groop, L. (2009). Epigenetics: A molecular link between environmental factors and type 2 diabetes. *Diabetes*, *58*(12), 2718–2725.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American statistical Association*, *83*(404), 1198–1202.

Liu, X., Jayaratne, R., Thai, P., Kuhn, T., Zing, I., Christensen, B., Lamont, R., Dunbabin, M., Zhu, S., Gao, J., et al. (2020). Low-cost sensors as an alternative for long-term air quality monitoring. *Environmental research*, *185*, 109438.

Liu, Y., Pan, J., Zhang, H., Shi, C., Li, G., Peng, Z., Ma, J., Zhou, Y., & Zhang, L. (2019). Short-term exposure to ambient air pollution and asthma mortality. *American Journal of Respiratory and Critical Care Medicine*, *200*(1), 24–32.

Liyanage, L., & Liyanage, S. H. (2010). Data integration and data mining framework to discover health impacts of climate change. *Annual International Academic Conference on Business Intelligence and Data Warehousing and Annual International Academic Conference on Data Analysis, Data Quality & Metadata Management: Singapore, 12-13 July 2010*.

Maantay, J. A., & McLafferty, S. (2011). Environmental health and geospatial analysis: An overview. In *Geospatial analysis of environmental health* (pp. 3–37). Springer.

Manda, S. O., Feltbower, R. G., & Gilthorpe, M. S. (2009). Investigating spatio-temporal similarities in the epidemiology of childhood leukaemia and diabetes. *European journal of epidemiology*, *24*, 743–752.

Martínez-Bello, D. A., López-Quílez, A., & Torres Prieto, A. (2018). Spatio-temporal modeling of zika and dengue infections within Colombia. *International Journal of Environmental Research and Public Health*, *15*(7), 1376.

Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, *101*, 1–9.

Montonen, J., Järvinen, R., Heliövaara, M., Reunanen, A., Aromaa, A., & Knekt, P. (2005). Food consumption and the incidence of type ii diabetes mellitus. *European Journal of Clinical Nutrition*, *59*(3), 441–448.

Moran, P. A. (1948). The interpretation of statistical maps. *Journal of the Royal Statistical Society. Series B (Methodological)*, *10*(2), 243–251.

Morawska, L., Thai, P. K., Liu, X., Asumadu-Sakyi, A., Ayoko, G., Bartonova, A., Bedini, A., Chai, F., Christensen, B., Dunbabin, M., et al. (2018). Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: How far have they gone? *Environment international*, *116*, 286–299.

Moritz, S., & Bartz-Beielstein, T. (2017a). imputeTS: Time Series Missing Value Imputation in R. *The R Journal*, *9*(1), 207–218. https://doi.org/10.32614/RJ-2017-009

Moritz, S., & Bartz-Beielstein, T. (2017b). imputeTS: time series missing value imputation in R. *The R Journal*, *9*(1), 207.

Moritz, S., Sardá, A., Bartz-Beielstein, T., Zaefferer, M., & Stork, J. (2015). Comparison of different methods for univariate time series imputation in r. *arXiv preprint arXiv:1510.03924*.

Morland, K., Roux, A. V. D., & Wing, S. (2006). Supermarkets, other food stores, and obesity: The atherosclerosis risk in communities study. *American journal of preventive medicine*, *30*(4), 333–339.

Nakagawa, S. (2015). Missing data: Mechanisms, methods and messages. *Ecological statistics: Contemporary theory and application*, 81–105.

Nakagawa, S., & Freckleton, R. P. (2008). Missing inaction: The dangers of ignoring missing data. *Trends in ecology & evolution*, *23*(11), 592–596.

Nakagawa, S., & Freckleton, R. P. (2011). Model averaging, missing data and multiple imputation: A case study for behavioural ecology. *Behavioral Ecology and Sociobiology*, *65*(1), 103–116.

National Environment Protection Council (Australia). (2021). *National environment protection (ambient air quality) measure*. National Environment Protection Council.

Norazian, M. N., Shukri, Y. A., Azam, R. N., & Al Bakri, A. M. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia*, *34*(3), 341–345.

NSW Department of Planning and Environment. (2021). Standards and goals for measuring air pollution [https://www.environment.nsw.gov.au/topics/air/understanding-air-quality-data/standards-and-goals, Last accessed on 2022-09-22].

NSW Department of Planning and Environment. (2022). How and why we monitor air pollution [https://www.environment.nsw.gov.au/topics/air/air-quality-basics/sampling-air-pollution, Last accessed on 2022-09-22].

Orioli, R., Cremona, G., Ciancarella, L., & Solimini, A. G. (2018). Association between PM10, PM2. 5, NO2, O3 and self-reported diabetes in Italy: A cross-sectional, ecological study. *PloS one*, *13*(1), e0191112.

Pearson, J. F., Bachireddy, C., Shyamprasad, S., Goldfine, A. B., & Brownstein, J. S. (2010). Association between fine particulate matter and diabetes prevalence in the US. *Diabetes care*, *33*(10), 2196–2201.

Pedrycz, W., Lu, W., Liu, X., Wang, W., & Wang, L. (2014). Human-centric analysis and interpretation of time series: A perspective of granular computing. *Soft Computing*, *18*, 2397–2411.

Peuquet, D. J., & Duan, N. (1995). An event-based spatiotemporal data model (ESTDM) for temporal analysis of geographical data. *International journal of geographical information systems*, *9*(1), 7–24.

Planning and Environment. (2020). Data download facility [https://www.dpie.nsw.gov.au/air-quality/air-quality-data-services/data-download-facility, Last accessed on 2022-04-30].

Polichetti, G., Cocco, S., Spinali, A., Trimarco, V., & Nunziata, A. (2009). Effects of particulate matter (PM10, PM2. 5 and PM1) on the cardiovascular system. *Toxicology*, *261*(1-2), 1–8.

Pope III, C. A., Dockery, D. W., Spengler, J. D., & Raizenne, M. E. (1991). Respiratory health and PM10 pollution: A daily time series analysis. *American Review of Respiratory Disease*, *144*(3_pt_1), 668–674.

Prasannakumar, V., Vijith, H., Charutha, R., & Geetha, N. (2011). Spatio-temporal clustering of road accidents: GIS based analysis and assessment. *Procedia-social and behavioral sciences*, *21*, 317–325.

Prüss-Üstün, A., Wolf, J., Corvalán, C., Bos, R., & Neira, M. (2016). *Preventing disease through healthy environments: A global assessment of the burden of disease from environmental risks*. World Health Organization.

R Core Team. (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

R Core Team. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.

Rajagopalan, S., & Brook, R. D. (2012). Air pollution and type 2 diabetes: Mechanistic insights. *Diabetes*, *61*(12), 3037–3045.

Rantou, K. (2017). Missing data in time series and imputation methods. *University of the Aegean, Samos*.

Ratcliffe, J. H. (2002). Aoristic signatures and the spatio-temporal analysis of high volume crime patterns. *Journal of quantitative criminology*, *18*, 23–43.

Riley, M., Kirkwood, J., Jiang, N., Ross, G., & Scorgie, Y. (2020). Air quality monitoring in NSW: From long term trend monitoring to integrated urban services. *Air Quality and Climate Change*, *54*(1), 44–51.

Roberts, D. R., Bahn, V., Ciuti, S., Boyce, M. S., Elith, J., Guillera-Arroita, G., Hauenstein, S., Lahoz-Monfort, J. J., Schröder, B., Thuiller, W., et al. (2017). Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. *Ecography*, *40*(8), 913–929.

Roychowdhury, S., & Pedrycz, W. (2002). Modeling temporal functions with granular regression and fuzzy rules. *Fuzzy sets and systems*, *126*(3), 377–387.

Rubin, D. (1976). Inference and missing data. *Biometrika*, *63*(3).

Rubright, J. D., Nandakumar, R., & Gluttin, J. J. (2014). A simulation study of missing data with multiple missing x's. *Practical Assessment, Research, and Evaluation*, *19*(1), 10.

Sainsbury, E., Shi, Y., Flack, J., & Colagiuri, S. (2018). *Burden of diabetes in australia: It's time for more action* (tech. rep.).

Sardá-Espinosa, A. (2017). Comparing time-series clustering algorithms in r using the dtwclust package. *R package vignette*, *12*, 41.

Scibor, M., & Malinowska-Cieslik, M. (2020). The association of exposure to PM10 with the quality of life in adult asthma patients. *International Journal of Occupational Medicine and Environmental Health*, *33*(3).

Shahbazi, H., Karimi, S., Hosseini, V., Yazgi, D., & Torbatian, S. (2018). A novel regression imputation framework for tehran air pollution monitoring network using outputs from WRF and CAMx models. *Atmospheric Environment*, *187*, 24–33.

Sharmin, S., & Rayhan, M. I. (2012). Spatio-temporal modeling of infectious disease dynamics. *Journal of Applied Statistics*, *39*(4), 875–882.

Shin, J., Lee, H., & Kim, H. (2020). Association between exposure to ambient air pollution and age-related cataract: A nationwide population-based retrospective cohort study. *International journal of environmental research and public health*, *17*(24), 9231.

Shmueli, G. (2010). To explain or to predict? *Statistical science*, *25*(3), 289–310.

State of NSW and Department of Planning, Industry and Environment. (2020). *Air quality application programming interface (API) user guide*.

Strachan, D. P. (2000). The role of environmental factors in asthma. *British medical bulletin*, *56*(4), 865–82.

Suris, F. N. A., Bakar, M. A. A., Ariff, N. M., Mohd Nadzir, M. S., & Ibrahim, K. (2022). Malaysia PM10 air quality time series clustering based on dynamic time warping. *Atmosphere*, *13*(4), 503.

Thiering, E., & Heinrich, J. (2015). Epidemiology of air pollution and diabetes. *Trends in Endocrinology and Metabolism*, *26*(7), 384–394. https://doi.org/https://doi.org/10.1016/j.tem.2015.05.002

To, T., Zhu, J., Villeneuve, P. J., Simatovic, J., Feldman, L., Gao, C., Williams, D., Chen, H., Weichenthal, S., Wall, C., et al. (2015). Chronic disease prevalence

in women and air pollution-a 30-year longitudinal cohort study. *Environment international*, *80*, 26–32.

Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, *46*(sup1), 234–240.

Torabi, M. (2013). Spatio–temporal modeling for disease mapping using CAR and B-spline smoothing. *Environmetrics*, *24*(3), 180–188.

Tu, M., Zhang, Y., Xu, J., & Li, Y. (2015). Analysis and modeling of time series based on granular computing. *International Journal of Future Computer and Communication*, *4*(2), 93.

Tyrovolas, S., Chalkias, C., Morena, M., Kalogeropoulos, K., Tsakountakis, N., Zeimbekis, A., Gotsis, E., Metallinos, G., Bountziouka, V., Lionis, C., et al. (2014). High relative environmental humidity is associated with diabetes among elders living in Mediterranean islands. *Journal of Diabetes and Metabolic Disorders*, *13*(1), 1–7.

Waller, L. A., Carlin, B. P., Xia, H., & Gelfand, A. E. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical association*, *92*(438), 607–617.

Waller, L. A., & Gotway, C. A. (2004). *Applied spatial statistics for public health data*. John Wiley & Sons.

Wang, X., & Brown, D. E. (2012). The spatio-temporal modeling for criminal incidents. *Security Informatics*, *1*(1), 1–17.

Wijesekara, L., & Liyanage, L. (2020a). Imputing large gaps of missing values using seasonal decomposition and elastic-net regression in the presence of correlated variables: Algorithm using meteorological data.

Wijesekara, L., & Liyanage, L. (2020b). Modelling environmental impact on public health using machine learning: Case study on asthma. *2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA)*, 1–7.

Wijesekara, L., & Liyanage, L. (2021a). Air quality data pre-processing: A novel algorithm to impute missing values in univariate time series. *2021 IEEE 33rd*

International Conference on Tools with Artificial Intelligence (ICTAI), 996–1001. https://doi.org/https://doi.org/10.1109/ICTAI52525.2021.00159

Wijesekara, L., & Liyanage, L. (2021b). Imputing large gaps of high-resolution environment temperature. *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*, 74–79. https://doi.org/https://doi.org/10.1109/ICIIS53135.2021.9660672

Wijesekara, L., & Liyanage, L. (2023). Mind the large gap: Novel algorithm using seasonal decomposition and elastic net regression to impute large intervals of missing data in air quality data. *Atmosphere*, *14*(2), 355.

Wijesekara, L., Nanthakumaran, P., & Liyanage, L. (2022). Space and time data exploration of air quality based on PM10 sensor data in Greater Sydney 2015–2021. *International Conference on Sensing Technology*, 295–308.

Wijesekara, W. M. L. K. N., & Liyanage, L. (2020c). Comparison of imputation methods for missing values in air pollution data: Case study on Sydney air quality index. *Future of Information and Communication Conference*, 257–269. https://doi.org/https://doi.org/10.1007/978-3-030-39442-4_20

World Health Organization. (2016). *Global report on diabetes* (tech. rep.). World Health Organization.

Wyzga, R. E. (1973). Note on a method to estimate missing air pollution data. *Journal of the Air Pollution Control Association*, *23*(3), 207–208.

Yang, B.-Y., Fan, S., Thiering, E., Seissler, J., Nowak, D., Dong, G.-H., & Heinrich, J. (2020a). Ambient air pollution and diabetes: A systematic review and meta-analysis. *Environmental research*, *180*, 108817.

Yang, J., Zhou, M., Zhang, F., Yin, P., Wang, B., Guo, Y., Tong, S., Wang, H., Zhang, C., Sun, Q., et al. (2020b). Diabetes mortality burden attributable to short-term effect of PM10 in China. *Environmental Science and Pollution Research*, *27*(15), 18784–18792.

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series*, *1168*(2), 022022.

Yuan, H., Xu, G., Yao, Z., Jia, J., & Zhang, Y. (2018). Imputation of missing data in time series for air pollutants using long short-term memory recurrent neural networks. *Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers*, 1293–1300.

Zakaria, N. A., & Noor, N. M. (2018). Imputation methods for filling missing data in urban air pollution data formalaysia. *Urbanism. Arhitectura. Constructii*, *9*(2), 159.

Zanobetti, A., Luttmann-Gibson, H., Horton, E. S., Cohen, A., Coull, B. A., Hoffmann, B., Schwartz, J. D., Mittleman, M. A., Li, Y., Stone, P. H., et al. (2014). Brachial artery responses to ambient pollution, temperature, and humidity in people with type 2 diabetes: A repeated-measures study. *Environmental health perspectives*, *122*(3), 242–248.

Zanolin, M., Pattaro, C., Corsico, A., Bugiani, M., Carrozzi, L., Casali, L., Dallari, R., Ferrari, M., Marinoni, A., Migliore, E., Olivieri, M., Pirina, P., Verlato, G., Villani, S., de Marco, R., Buriani, O., Cavallini, R., Saletti, C., Cellini, M., ... Salomoni, A. (2004). The role of climate on the geographic variability of asthma, allergic rhinitis and respiratory symptoms: Results from the italian study of asthma in young adults. *Allergy: European Journal of Allergy and Clinical Immunology*, *59*(3), 306–314.

Zheng, K., Zhao, S., Yang, Z., Xiong, X., & Xiang, W. (2016). Design and implementation of LPWA-based air quality monitoring system. *IEEE Access*, *4*, 3238–3245.

Zhou, Y., Li, L., & Hu, L. (2017). Correlation analysis of PM10 and the incidence of lung cancer in Nanchang, China. *International Journal of Environmental Research and Public Health*, *14*(10), 1253.

Zhuang, L., & Cressie, N. (2012). Spatio-temporal modeling of sudden infant death syndrome data. *Statistical Methodology*, *9*(1-2), 117–143.