University of Pretoria

Faculty of Health Sciences

School of Health Sciences and Public Health

Unsupervised machine learning in air pollution epidemiology in South Africa: Artificial Intelligence subset application

For the submitted in partial fulfilment of the requirement for the degree

Doctor of Philosophy (Epidemiology)

By

Nandi Sisasenkosi Mwase

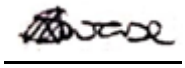
Student number: 17242496

Supervisor: Prof Janine Wichmann

Supervisor: Prof Washington Junger (Rio de Janeiro State University)

DECLARATION

I, Nandi Sisasenkosi Mwase, hereby declare that this thesis which I have submitted for the degree Doctor of Philosophy (Epidemiology) at the School of Health Systems and Public Health at the University of Pretoria, is my own work and has not previously been submitted by me for any other degree or examination at any other tertiary institution.



Nandi Sisasenkosi Mwase

Signed on the 2nd day of _May 2023 in Pretoria

ETHICS STATEMENT

I have obtained ethical approval by the Health Science ethical committee to conduct the research under the above mentioned title of this thesis.

All ethical standards required in terms of the University of Pretoria's Code of ethics for researchers and the policy guidelines for responsible research have been observed.

PUBLICATION/CONFERENCE PRESENTAION

Conference Presentation

Stellenbosch University Global Health (SUGH) conference, poster presentation, “Attitudes and perceptions of postgraduate students towards use of artificial intelligence in public health.”

Fully Online Distance Education Symposium (FODES) presentation “Attitudes and perceptions of postgraduate students towards use of artificial intelligence in public health.”

Manuscripts under review

Attitudes and perceptions of postgraduate students towards the use of artificial intelligence in public health, a cross sectional survey of postgraduate diploma students, South Africa.

DEDICATION

“It takes a village to raise a child.”

This journey has not been mentally or emotionally easy, but I am truly grateful for my family’s support. I appreciate my mother Patricia’s sacrifices to get me to this point, my uncle Malumbo that inspired me to take this journey. I am grateful for the rest of the Mwase clan, my grandfather (Frank Sr), uncle (Frank Jnr), aunt (Thokozile), siblings (Samukhele, Nthato, Frank Jnr Jnr and Akuzike), for supporting me in so many ways to achieve this goal. Although she is not here to see it, I wouldn’t be here without the strong influence of my grandmother, Julia Helen Mwase. She taught me how to work hard for the things I want and perceiver even when the odds were against me. More importantly, she taught me to depend on God because that was the only way to get through life. It is sad that she is not here to see this but I am proud to be a part of her legacy.

ACKNOWLEDGEMENTS

Many thanks to my supervisor Prof Janine Wichmann whom I have had the pleasure of working with throughout my postgraduate. I hope to be just as dedicated to my work as she has shown me to be. In the same vain, I have met a number of inspiring female academics Dr Neo Ledibane, Mrs Lizeka Napoles and Prof Liz Wolvaardt that have all been forces to reckon with and great inspirations in my academic journey at the School of Health Systems and Public Health.

Many thanks to my co-supervisor Prof Washington Junger for taking time to teach me so many technical aspects especially that I came with no prior knowledge. It has been a pleasure being working under his supervision.

Last but not the least, a great thanks to the University of Pretoria that funded my doctoral studies through the UP Postgraduate Bursary from 2020 to 2022. I definitely would not have been able to start or complete my studies without the great financial help from the University.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ARIMA	Autoregressive Integrated Moving Average
CO	Carbon monoxide
CVD	Cardiovascular Disease
DEA	Department of Environmental Affairs
DFFE	Department of Forestry, Fisheries and the Environment
DL	Deep Learning
EPC	Environmental Pollutant Clustering algorithm
HPA	Highveld Priority Area
LOCF	Last observations carried forward
MAR	Missing at Random
MCAR	Missing Completely at Random
mice	Multiple Imputation by Chain Equations
ML	Machine Learning
MNAR	Missing not at Random
mtsdi	multivariate time series data imputation
NO_2	Nitrogen Dioxide
O_3	Ozone
PM	Particulate matter
$PM_{2.5}$	Particulate matter with a mean aerodynamic diameter of 2.5µm or less
PM_{10}	Particulate matter with a mean aerodynamic diameter of 10µm or less
PMF	Positive Matrix Factorization
PMM	Predictive Mean Matching
RD	Respiratory Disease
SO_2	Sulphur dioxide
Tapp	Apparent Temperature
VTAPA	Vaal Triangle Air Pollution Priority Area
WBPA	Waterberg-Bojanala Priority Area
XRF	X-ray Fluorescence

EXECUTIVE SUMMARY

Background:

Clean air is a human right and a condition for healthy living, but air pollution remains a global concern. The World Health Organization (WHO) has stated the detrimental health effects of air pollution, equating the effects to other health risks including an unhealthy diet and smoking tobacco. Air pollution is a complex mixture of droplets, solid particles, and gases, such as particulate matter (PM), nitrogen dioxide (NO_2), ground-level ozone (O_3), and sulphur dioxide (SO_2). Air pollution is globally recognised as the most significant environmental threat to human health. Exposure to air pollution is associated with increased risk of respiratory diseases, cardiovascular diseases, and cancers, as well as increased risk of mortality. The global estimation of the number of deaths from air pollution ranges from 6.7 to 7 million deaths. Low- and middle-income countries (LMIC) are reported to account for a substantial proportion of these fatalities, with Africa accounting for approximately one-million deaths. Long-term exposure to household air pollution has also contributed 4% of global deaths.

There are a number of pollutants that have been associated with negative health effects. As of 2019, in South Africa, the State of Global Air estimated 24 800 premature deaths due to exposure to $PM_{2.5}$. However, this may be an underestimation as there are only a few studies in South Africa sampling $PM_{2.5}$ and associating the pollutant with mortality. Ground-level ozone has contributed to approximately 365 000 deaths, equating to 11% of chronic obstructive pulmonary disease (COPD) deaths globally. However, all air pollutant estimations and the associated number of deaths are reliant on exposure-response functions derived from epidemiological studies that are predominantly conducted in developed countries. Currently, there are limited studies conducted in LMIC, like South Africa that provide a comprehensive understanding of the impact of air pollution. Hence, it is critical for more epidemiological studies on air pollution to be conducted in countries such as South Africa.

The epidemiological evidence on the health effects of air pollution mixtures is lacking globally. This could indicate a current underestimation of the health risks from merely adding air pollutants together in statistical models. There are various traditional statistical methods that have been proposed to investigate the health effects of air

pollution mixtures, such as multi-linear regression, classification and regression tree analysis (CART), cox proportional hazards regression, etc. Recently researchers have also applied Machine Learning (ML) methods, which is a subset of Artificial Intelligence (AI), to address this topic. The majority of studies have applied unsupervised ML, such as k-means clustering, however, such studies are lacking in Africa.

Additionally, there are multiple sources, both man-made and natural, that can lead to different mixtures of air pollutants, such as PM_{10} and $PM_{2.5}$. While many epidemiological studies mainly focus on the mass of PM_{10} and $PM_{2.5}$, few studies investigate the chemical composition and identification of their sources. Positive Matrix Factorization (PMF) is a well-regarded method for source apportionment. Similar to other research areas, ML methods such as k-means and spectral clustering are being used as alternative source apportionment methods. Even fewer studies in South Africa are investigating the use of ML as a source apportionment method.

Therefore, the aim of this PhD thesis was to address some of the research gaps identified above, namely, the lack of studies in Africa on the health effects of air pollution mixtures and $PM_{2.5}$ source apportionment, whilst also assessing the applicability of AI methods, such as unsupervised ML, in air pollution epidemiology in South Africa. The thesis objectives were to:

- Assess the perceptions and attitudes regarding AI in public health among postgraduate students registered for the online Postgraduate Diploma in Public Health at the School of Health Systems and Public Health (SHSPH), University of Pretoria (UP).
- Determine the joint effects of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} on hospital admissions for respiratory disease (RD) and cardiovascular disease (CVD) in Vereeniging and Vanderbijlpark, Gauteng, using traditional statistical analysis, specifically, classification and regression trees. Thereafter, unsupervised Machine Learning methods are utilised to determine the joint effects of the air pollutants on RD and CVD hospital admissions.
- Compare two methods of source apportionment of $PM_{2.5}$ in Pretoria – a traditional method such as Positive Matrix Factorization (PMF) and unsupervised Machine Learning clustering methods.

Method: The PhD project was divided into three parts. The first was a cross-sectional survey among students enrolled in the Postgraduate Diploma in Public Health at UP to assess perceptions and attitudes regarding AI in public health.

The second part of the project was to determine the joint effects of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} on RD and CVD hospital admissions in Vereeniging and Vanderbijlpark, in the Vaal Triangle Airshed Priority Area (VTAPA), South Africa. There was a total of 3 346 observations from 2 January 2011 to 29 February 2020 (before the first recorded COVID-19 case in South Africa). The statistical CART analysis was used to assess the joint effects. Seven air pollution mixtures were created in the analyses, i.e. (mixture 1) PM_{10}, NO_2, and SO_2, (mixture 2) $PM_{2.5}$, NO_2, and SO_2, (mixture 3) PM_{10}, NO_2, and O_3, (mixture 4) $PM_{2.5}$, NO_2, and O_3, (mixture 5) PM_{10}, SO_2, and O_3, (mixture 6) $PM_{2.5}$, SO_2, and O_3, and (mixture 7) O_3, NO_2, and SO_2. Thereafter, unsupervised ML clustering methods – k-means, spectral clustering, and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) – were applied to the air pollution data to determine their joint effects on RD and CVD hospital admissions.

Lastly, source apportionment for $PM_{2.5}$ in Pretoria was performed using PMF analysis and unsupervised ML clustering methods, i.e. k-means, spectral clustering and principle component analysis (PCA). There was a total of 428 observations collected from 18 April 2017 to 12 February 2021. Gravimetric analysis was used to calculate the concentration levels and species identification was done through X-ray Fluorescence (XRF). The following fifteen identified species were used in the PMF model: $PM_{2.5}$, BC, UV-PM, S, Cl, K, Ca, Ti, Fe, Ni, Cu, Zn, Br, U, and Si.

Results: 618 respondents completed an online survey (81.5% response rate). Generally, respondents thought AI would be capable of performing various tasks that did not provide direct care to individuals. Most (69%) agreed that the introduction of AI could reduce job availability in public health fields. Respondents agreed that AI in public health could raise ethical (84%), social (77%), and health equity (77%) challenges. Relatively few respondents (52%) thought they were being adequately trained to work alongside AI tools and the majority (76%) felt training of AI competencies should begin at an undergraduate level.

The air pollution (SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}) and meteorological data (relative humidity and temperature) used was from 1 January 2011 to 29 February 2020 (before the first recorded COVID-19 case in South Africa). Due to the missing air pollution and meteorological data for the VTAPA area, data was imputed using the multiple imputation by chain equations (mice) method.

There were 54 822 respiratory disease (RD) hospital admissions in VTAPA from 2 January 2011 to 29 February 2020 (before the first recorded COVID-19 case). Generally, the risk of RD hospital admissions increased by 1.04 (95% CI 1.01, 1.08) when exposed to mixtures with high levels of NO_2 and varying levels of SO_2, O_3, $PM_{2.5}$, and PM_{10}. There were 22 205 cardiovascular disease (CVD) hospital admissions in VTAPA during the study period. The RRs of CVD hospital admissions increased among those exposed to air pollution mixtures numbered (2), (3), (4), (6), and (7) by 1.11 (95% CI 1.02, 1.20), 1.15 (95% CI 1.04, 1.29), 1.13 (95% CI 1.05, 1.21), 1.11 (95% CI 1.02, 1.20), and 1.14 (95% CI 1.06, 1.22), respectively. Similar to findings for RD, the highest risk for CVD hospitalisation was found when exposed to high levels of NO_2 and varying levels of SO_2, O_3, $PM_{2.5}$, and PM_{10}.

The unsupervised ML clustering methods used – k-means clustering and spectral clustering – showed that the air pollution data SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} were best grouped into two clusters. However a three-cluster spectral clustering model using the normalised Laplacian matrix, showed that the risk of RD hospital admission increased when exposed to SO_2, NO_2, $PM_{2.5}$, and PM_{10} in higher concentration levels, and lower levels of O_3 by 1.04 (95% CI 1.01-1.08). None of the formed cluster mixtures were found to increase the risk of CVD hospital admission. The DBSCAN clustering method did not prove to be an appropriate clustering method, as it greatly reduced the dataset and produced ill-distributed observations within formed clusters.

A seven-factor PMF model was assigned to $PM_{2.5}$ data collected over a 46-month period in Pretoria, South Africa. The seven contributing sources identified included mining (43.2%), biomass/coal burning (14.2%), secondary sulphur (12.1%), road traffic (11.3%), industry/base metal (8.7%), resuspended dust (8.5%), and general exhaust emissions (2.0%). PMF analysis was relatively easy to conduct and analyse, however, the process proved to be computationally taxing for medium to large datasets. Additionally, the modelled $PM_{2.5}$ concentration levels was lower than the

actual $PM_{2.5}$ concentration levels; the correlation between modelled $PM_{2.5}$ and actual $PM_{2.5}$ data was $R^2 = 0.6$.

The seven-cluster spectral clustering model, using the normalized Laplacian matrix, showed feasible sources for the $PM_{2.5}$ data during the 46-month period in Pretoria, South Africa. The possible identified sources of $PM_{2.5}$ were coal burning (42.89%), industry (22.0%), resuspended dust (10.4%), base metal (6.7%), road traffic (6.8%), general exhaust emissions (5.8%), and secondary sulphur (5.5%). Spectral clustering was easy to run, not computationally taxing, and utilised the complete dataset within the clustering. This suggests that it was a good dimension reduction tool that can produce plausible results for source apportionment. However, there was an issue of overlapping clusters and a lack of external validation for the formed clusters. This is a reason of concern when using spectral clustering for source apportionment.

Conclusion: The study contributes to the limited, but growing, knowledge and application of ML and AI in public health and air pollution epidemiology. The survey yielded a variety of views. There was a general assumption that AI in public health could assist in performing particular tasks at different health levels that did not involve direct care. There was also a general consensus that AI had the potential to raise unemployment and ethical challenges in the public health field in South Africa.

SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} mixtures proved to be associated with RD and CVD hospital admission. The mixtures showed that a higher concentration of NO_2 in combination with varying concentrations of SO_2, O_3, $PM_{2.5}$, and PM_{10} can lead to increased risk of both RD and CVD hospitalisation. This result contributes epidemiological evidence that can help policy makers to introduce stricter policies for improving the air quality of national priority areas, such as VTAPA in South Africa.

Unsupervised ML could be useful in determining joint effects of air pollutants on hospital admission and other health outcomes. K-means and spectral clustering were both relatively easy to run and analyse; they were also less time consuming in comparison to the CART analyses. The process also showed promise for analysing more than three air pollutants, in spite of the different interactions. However, it is evident that further study is needed before unsupervised ML can be considered a

reliable and definite tool to study the joint effects of air pollution on different health outcomes.

PMF modelling suggested that mining and industry were the main contributing factors to $PM_{2.5}$ in Pretoria. However, there is a great need for more studies that sample $PM_{2.5}$ in Africa. Source apportionment studies are vital in the evaluation of policies intended to protect communities from the detrimental health effects of $PM_{2.5}$. The PMF software was relatively easy to use and the data produced was relatively easy to analyse for possible sources of $PM_{2.5}$. However, the three model runs only showed 0.4 to 0.6 correlation with the original data.

Unsupervised ML for source apportionment is still a relatively new concept and needs to be further explored. In comparison with PMF, spectral clustering showed potential as a dimension reducing tool for source apportionment. Although the sources identified in the spectral clustering model showed similar sources identified in the PMF model, there were some noticeable limitations. Extensive studies are needed to continue exploring the potential of clustering for source apportionment studies.

Furthermore, there is a need to increase air pollution epidemiology and source apportionment studies in South Africa. This will increase African-based evidence of the detrimental effects of air pollution. Air pollution studies using unsupervised ML has the potential to be used in air pollution and public health studies. This project produces a baseline in the current perceptions of AI in public health and could lead to more in-depth studies on the topic. With hopes to initiate conversation around including AI in public health, this project shows epidemiological evidence that can be used to advocate for stricter, more effectively enforced air quality standards and management plans in VTAPA. Lastly, the project also produces a baseline framework for including the application of ML in epidemiological and source apportionment studies. Spectral clustering provided plausible results in comparison to the results obtained using statistical and traditional models. Although the study used a limited number of unsupervised ML methods, it is highly recommended that other unsupervised ML methods be used in further public health studies to continue investigating the practical implementation of AI in public health.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER 1: BACKGOUND

1.1. DEFINING THE RESEARCH PROBLEM

Artificial Intelligence (AI) and Machine Learning (ML) are often used interchangeably, however, while they are related, they are not necessarily the same. Despite the recent popularity of AI and ML, the concept is not new and has been used since the 1950s.[1-3] In the information technology (IT), computer industry, and some business and finance industries, AI and ML have been adapted and used in multiple procedures.[4-10] There has also been AI and ML application within medical sciences, such as diagnostics and surgeries in the form of neural networks, natural language processing, and even robotics.[11-19] AI and ML have also been utilised in epidemiological research.[15,20-25]

Internationally, the inclusion of AI in medicine, public health, and epidemiology appears to produce more advanced research as compared to research done within Africa. AI and ML research in South Africa is increasing in multiple fields, including the business and financial sectors.[26-28] There is evidence showing South-African-based computer science and technology studies are seemingly at the same standard as their global counterparts, this also includes the progression of AI in diagnostics.[14,29]

Within South African health research AI and ML have been used in diagnostics, HIV research, and in clinical monitoring.[17,30-32] In low-to-middle-income countries (LMIC), the main focus of AI interventions has been on health issues such as tuberculosis,[30,33-34] malaria,[31,35] non-infectious diseases in children and infants, and cervical cancer.[22] Although research concerning AI in public health in Africa is growing, AI applications in public health and medical studies are more prominent in other countries, including China, the United States of America, and Europe.[26,36-37] The use of AI in South Africa is increasing with several studies that show the use and potential benefits of AI application in healthcare and medical studies.[29,38-39] Additionally, ML has also been used as a prediction tool in the provision of healthcare services and placements of healthcare workers.[14] The extent and implementation of AI and ML in epidemiology, in an African context, remains in the early stages, leaving room for exploration.[40]

It would be remiss not to mention that globally the public health and epidemiology sector is in a hype cycle concerning the use of AI and ML in research. Figure 1.1 shows the Gartner hype cycle for AI, the wave of AI involvement and utilisation that is currently occurring.

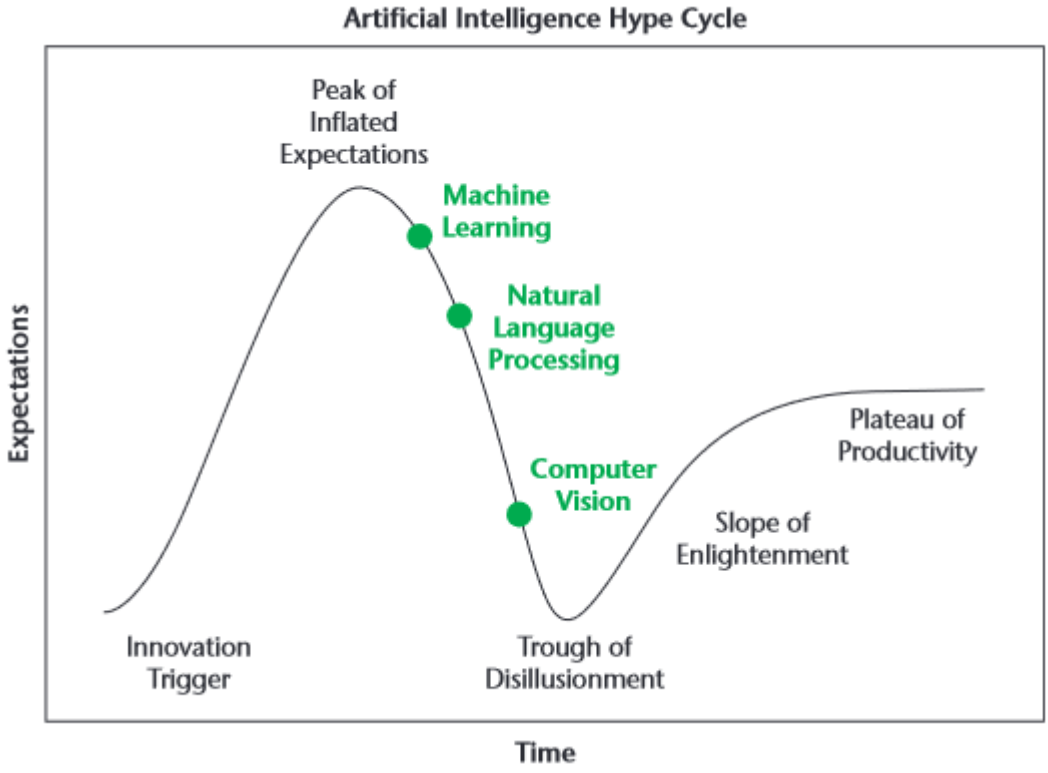


Figure 1.1: "Artificial Intelligence Hype Cycle, Machine learning, Natural Language Processing and Computer Vision on its way down –Adapted from Gartner Hype Cycle for Artificial Intelligence, 2019 gartner.com/smarterwithgartner."[41]

There is a peak of inflated expectation for the use of AI in health research and this is due to new methods being discovered and developed.[42] Although there is a peak in expectation, there seems to be a limited utilisation of AI in public health and environmental epidemiological research. Few studies show how AI and ML are being used to assist in researching causality and association between exposure and health data. The use of ML in research shows more applicability of supervised ML methods, such as prediction modelling,[14,43-44] as opposed to the use of unsupervised ML methods. This suggests an evident knowledge gap in assessing how applicable unsupervised ML methods can be used in air pollution epidemiology studies. More needs to be done to investigate the potential benefits of applying AI and ML methods

in improving environmental epidemiology research. Air pollution and health studies could benefit from AI application, but are difficult to run, since air pollution data are often missing and there is a lack of public digitalised hospital data in South Africa. [45-46]

1.2. MOTIVATION AND RELEVANCE

The main focus of this project was to explore the use of unsupervised ML (a subset of AI) in air pollution epidemiology within a South African context. This project looked to provide insight into what future public health professionals' perceptions and attitudes are on AI. AI is not a common topic taught in the medical or public health curriculum. Although AI seems to be a growing and popular topic, it could be assumed that this popularity only reaches a certain extent and is not as popular within public health in South African tertiary institutions.

Hence, this project provides baseline information to build upon, which will hopefully influence the inclusion of AI in public health education. Understanding perceptions and beliefs surrounding AI can assist in strategizing how information on AI can be better delivered or improved on to present and future public health professionals. This project addresses one of the nine strategic interventions of the National Digital Health Strategy for South Africa (2019-2024)[45] to be achieved by 2024, namely "to develop enhanced digital health technical capacity and skilled workforce for digital technology support and implementation". Currently, there are no studies on the perceptions of AI that have been published in South African literature.

Air pollution and health studies, as a combined discipline, is an emerging research field in South Africa. The Sustainable Development Goals (SDGs) have highlighted the importance of a healthy environment, e.g. clean air, which can help with healthy living.[46] With recognition of the detrimental health effects of air pollution on health by the South African Department of Health,[47] there are more studies to be done on air pollution and its effects on health. Majority of the available data on the global burden of disease (GBD) in Africa, attributed to air pollutants such as $PM_{2.5}$ and O_3, are modelled estimates.[48] With few studies doing first-hand sampling to get accurate figures, modelled estimates could lead to an underestimation of the number of deaths attributed to single air pollutant exposure. LMIC are said to account for a large

proportion of air pollution, but there are fewer air quality monitoring and epidemiological studies in comparison to high-income countries. Hence, it is important that more monitoring and epidemiological studies are conducted to reduce uncertainties and possible underestimated assumptions about the health effects experienced in these LMIC.[49] This project does not only add to this agenda, but also investigates the health effects of air pollution mixtures, as opposed to single pollutant effects on health.

Additionally, this project considers the use of unsupervised ML to run dimension reduction on air pollution data to assess the joint effects of air pollutant mixtures and their association with hospital admissions. Air pollution data and hospital data availability are evident issues in South Africa. Long-term continuous air pollution data are not always readily available in spite of the numerous monitoring stations across the country.[50] Although imputation is a viable solution to address missing data, it can only be implemented to a certain degree. The availability of electronic health data are greatly limited in South Africa, since majority of the public health facilities do not use electronic records and the private health sector only partially represents the country's health.[51-54] The limited availability of some effects of air pollution on health data could also be due to delayed reporting by the national body of statistics, Statistics South Africa (Stats SA). Deaths attributed to ambient air exposure was last reported on in 2018, which reported that only 1.2% of deaths were due to exposure to electric current, radiation, and ambient air.[54] The large portion of missing air pollution and health data could be the reason why studies using AI and ML have not been as readily explored in South Africa, since these applications operate with large datasets.

This project used PMF to identify possible sources of $PM_{2.5}$ sampled in Pretoria. The sampling was conducted over a 46-month period and provides data on $PM_{2.5}$ in the area. Thereafter, unsupervised ML was also used for source apportionment. This project focused on the use of unsupervised ML from the perspective of an epidemiologist/public health scientist. The results are meant to add to the small body of knowledge, based on locally sourced air pollution data, available to South Africa. This project is a public health project that applies existing data science AI methods in air pollution epidemiology, rather than creating novel data science AI methods.

This project also examines the relationship between multi-pollutant exposures and health outcomes by using innovative statistical analysis methods. It draws from traditional epidemiological approaches of investigating individual pollutant association with a health outcome.[55] It explores both non-parametric and statistically-/mathematically-based Unsupervised Dimension Reduction (UDR) statistical methods of analysis. Non-parametric analysis approaches have been used sparingly in the analysis of air pollution health effects, but can provide a wealth of tools for discovering the complex relationship between air pollution and health outcomes.[55] Comparing traditional approaches to source apportionment with unsupervised ML clustering methods, could potentially address the shortcomings of the PMF technique. Although the ML methods suggested in this project are not necessarily "new" techniques in epidemiological studies, they are still very relevant and in need of exploration within the African research context. Even recent European studies have shown that there is still much to discover regarding these methods.[56]

The project will add to the country's knowledge on AI and ML in health studies and hopefully help reduce the associations of sheer 'trendiness' of the use of AI in research. The study could encourage the use of AI and ML as tools to improve the investigation of the adverse health effects of air pollution. These findings can assist in the monitoring, evaluation, and improvement of the South African National Air Quality Standards. Additionally, the findings of adverse health effects of air pollution exposure can emphasise the urgency to provide support to public health action in the country.

The study addresses three of the United Nations SDGs (valid 2016-2030), namely:

- Goal 3.9: "By 2030, substantially reduce the number of deaths and illnesses from hazardous chemicals and air, water and soil pollution and contamination"
- Goal 11.6: "By 2030, reduce the adverse per capita environmental impact of cities, including paying special attention to air quality and municipal and other waste management." Within the same section, Goal 11.6.2 stresses the reduction of "annual mean levels of fine particulate matter, (i.e. $PM_{2.5}$ and PM_{10}) in cities"
- Goal 13: "Take urgent action to combat climate change and its impacts"

1.3. AIM AND OBJECTIVES

1.3.1. AIM

The aim of this study is to assess the applicability of AI methods such as machine learning in the air pollution epidemiology in a South African context**.**

1.3.2. OBJECTIVES

- To assess the perceptions and attitudes regarding AI in public health among postgraduate students registered for the online Postgraduate Diploma in Public Health at the School of Health Systems and Public Health (SHSPH), University of Pretoria (UP).
- To determine the joint effects of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} on hospital admissions for respiratory disease (RD) and cardiovascular disease (CVD) in Vereeniging and Vanderbijlpark, Gauteng, using CART analysis. Thereafter, using the unsupervised Machine Learning methods to determine the joint effects of the air pollutants on RD and CVD hospital admission.
- To compare two methods of source apportionment of $PM_{2.5}$ in Pretoria, namely Positive Matrix Factorization (PMF) and unsupervised Machine Learning clustering methods.

1.4. OUTLINE OF THE THESIS

In the current chapter, the general introduction to the research topic is given, and the problem statement, the significance of the study, the research aims and objectives are addressed. Chapter 2 is a literature review on the topic, which provides an overview on the existing legislation and the current air pollution management in South Africa and highlights the adverse health effects caused by air pollution. Lastly, this chapter gives an overview of the use of AI applications, such as Machine Learning, both in general and specifically in South African and international health and public health sectors.

Chapter 3 explains the methods used in the project, Chapter 4 presents and discusses the results of the survey, and Chapter 5 discusses the imputation done for the missing air pollution and meteorological data.

Chapter 6 discusses the CART analysis done to determine the joint effects of air pollution on RD and CVD hospital admissions, whereas Chapter 7 discusses the use of unsupervised Machine Learning clustering in determining joint effects of air pollution on RD and CVD hospital admissions.

Chapter 8 discusses PMF source apportionment for $PM_{2.5}$ in Pretoria, and Chapter 9 discusses the use of unsupervised Machine Learning clustering methods for source apportionment for $PM_{2.5}$ in Pretoria. Finally, Chapter 10 provides an overall discussion of the study's results.

1.5. REFERENCES

1. Dick S. Artificial intelligence. 2019.

2. Gwagwa A, Kraemer-Mbula E, Rizk N, Rutenberg I, De Beer J. Artificial intelligence (AI) deployments in Africa: Benefits, challenges and policy dimensions. The African Journal of Information and Communication. 2020; 26(1):3-30. doi:10.23962/10539/30361.

3. Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, et al. Machine learning and artificial intelligence: Definitions, applications, and future directions. Current reviews in musculoskeletal medicine. 2020; 13(1):69-76. doi:10.1007/s12178-020-09600-8.

4. Dimitrieska S, Stankovska A, Efremova T. Artificial intelligence and marketing. Entrepreneurship. 2018; 6(2):298-304.

5. Dirican C. The impacts of robotics, artificial intelligence on business and economics. Procedia - Social and Behavioral Sciences. 2015; 195:564-73.

6. Feuerriegel S, Shrestha YR, von Krogh G, Zhang C. Bringing artificial intelligence to business management. Nature Machine Intelligence. 2022; 4(7):611-3.

7. Galasso A, Luo H. Punishing robots: Issues in the economics of tort liability and innovation in artificial intelligence. The economics of artificial intelligence: An agenda. University of Chicago Press. 2019.

8. Langley P. Artificial intelligence and cognitive systems. Artificial intelligence and Cognitive Systems Quarterly. 2011; 2:1-6.

9. Li X, Shi Y, editors. Computer vision imaging based on artificial intelligence. 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS); 2018: IEEE.

10. Loureiro SMC, Guerreiro J, Tussyadiah I. Artificial intelligence in business: State of the art and future research agenda. Journal of Business Research. 2021; 129:911-26.

11. Fitzpatrick F, Doherty A, Lacey G. Using artificial intelligence in infection prevention. Current Treatment Options in Infectious Diseases. 2020:1-10. doi:10.1007/s40506-020-00216-7.

12. Kim MC, Okada K, Ryner AM, Amza A, Tadesse Z, Cotter SY, et al. Sensitivity and specificity of computer vision classification of eyelid photographs for programmatic trachoma assessment. PLoS One. 2019; 14(2):e0210463. doi:10.1371/journal.pone.0210463.

13. Marcus JL, Sewell, W.C., Balzer, L.B., & Krakower, D.S. . Artificial intelligence and machine learning for HIV prevention: Emerging approaches to ending the epidemic. Current HIV/AIDS Reports. 2020; 17 (3):171-9.

14. Moyo S, Doan TN, Yun JA, Tshuma N. Application of machine learning models in predicting length of stay among healthcare workers in underserved communities in South Africa. Human Resources for Health. 2018; 16(1):68. doi:10.1186/s12960-018-0329-1.

15. Pan American Health Organization. Artificial intelligence in public health. . Washington, DC. 2021.

16. Paris CL, Swartout WR, Mann WC. Natural language generation in artificial intelligence and computational linguistics: Springer Science & Business Media. 2013.

17. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. Progress in retinal and eye research. 2018; 67:1-29. doi:10.1016/j.preteyeres.2018.07.004.

18. Stai B, Heller N, McSweeney S, Rickman J, Blake P, Vasdev R, et al. Public perceptions of artificial intelligence and robotics in medicine. Journal of Endourology. 2020; 34(10):1041-8. doi:10.1089/end.2020.0137.

19. Vyborny CJ, Giger ML. Computer vision and artificial intelligence in mammography. American Journal of Roentgenology. 1994; 162(3):699-708.

20. Flouris AD, Duffy J. Applications of artificial intelligence systems in the analysis of epidemiological data. European Journal of Epidemiology. 2006; 21(3):167-70. doi:10.1007/s10654-006-0005-y.

21. Komorowski M, Celi LA. Will artificial intelligence contribute to overuse in healthcare? Critical Care Medicine. 2017; 45(5):912-3. doi:10.1097/ccm.0000000000002351.

22. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet 2020; 395(10236):1579-86. doi:10.1016/S0140-6736(20)30226-9.

23. Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, et al. The role of artificial intelligence in achieving the sustainable development goals. Nature Communications. 2020; 11(1):233. doi:10.1038/s41467-019-14108-y.

24. VoPham T, Hart JE, Laden F, Chiang YY. Emerging trends in geospatial artificial intelligence (GEOAI): Potential applications for environmental epidemiology. Environmental Health. 2018; 17(1):40. doi:10.1186/s12940-018-0386-x.

25. Wong TY, Sabanayagam C. Strategies to tackle the global burden of diabetic retinopathy: From epidemiology to artificial intelligence. Ophthalmologica. 2020; 243(1):9-20. doi:10.1159/000502387.

26. Ferrein A, Meyer T. A brief overview of artificial intelligence in South Africa. AI Magazine. 2012; 33(1):99-103. doi:10.1609/aimag.v33i1.2357.

27. Adisa O, Botai J, Adeola A, Hassen A, Botai C, Darkey D, et al. Application of artificial neural network for predicting maize production in South Africa. Sustainability. 2019; 11(4).

28. Tularam H, Ramsay LF, Muttoo S, Brunekreef B, Meliefste K, de Hoogh K, et al. A hybrid air pollution/land use regression model for predicting air pollution concentrations in Durban, South Africa. Environmental Pollution. 2021; 274:116513. doi:10.1016/j.envpol.2021.116513.

29. Dlamini Z, Francies FZ, Hull R, Marima R. Artificial intelligence (AI) and big data in cancer and precision oncology. Computational and Structural Biotechnology Journal. 2020; 18:2300-11. doi:10.1016/j.csbj.2020.08.019.

30. Aguiar FS, Torres RC, Pinto JV, Kritski AL, Seixas JM, Mello FC. Development of two artificial neural network models to support the diagnosis of pulmonary tuberculosis in hospitalized patients in Rio de Janeiro, Brazil. Medical & Biological Engineering & Computing. 2016; 54(11):1751-9. doi:10.1007/s11517-016-1465-1.

31. Go T, Kim JH, Byeon H, Lee SJ. Machine learning-based in-line holographic sensing of unstained malaria-infected red blood cells. Journal of Biophotonics. 2018; 11(9):e201800101. doi:10.1002/jbio.201800101.

32. Sinisi SE, Polley EC, Petersen ML, Rhee SY, van der Laan MJ. Super learning: An application to the prediction of HIV-1 drug resistance. Statistical applications in genetics and molecular biology. 2007; 6:Article7. doi:10.2202/1544-6115.1240.

33. Jaeger S, Juarez-Espinosa OH, Candemir S, Poostchi M, Yang F, Kim L, et al. Detecting drug-resistant tuberculosis in chest radiographs. International journal of computer assisted radiology and surgery. 2018; 13(12):1915-25. doi:10.1007/s11548-018-1857-9.

34. Lopes UK, Valiati JF. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. Computers in Biology and Medicine. 2017; 89:135-43. doi:10.1016/j.compbiomed.2017.08.001.

35. Andrade BB, Reis-Filho A, Barros AM, Souza-Neto SM, Nogueira LL, Fukutani KF, et al. Towards a precise test for malaria diagnosis in the Brazilian Amazon: Comparison among field microscopy, a rapid diagnostic test, nested pcr, and a computational expert system based on artificial neural networks. Malaria Jounal. 2010; 9:117. doi:10.1186/1475-2875-9-117.

36. Bellinger C, Mohomed Jabbar MS, Zaïane O, Osornio-Vargas A. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health. 2017; 17(1):907. doi:10.1186/s12889-017-4914-3.

37. Cisse M. Look to Africa to advance artificial intelligence. Nature. 2018; 562(7728):461. doi:10.1038/d41586-018-07104-7.

38. Mbunge E, Batani J, Gaobotse G, Muchemwa B. Virtual healthcare services and digital health technologies deployed during coronavirus disease 2019 (COVID-19)

pandemic in South Africa: A systematic review. Global Health Journal. 2022; https://doi.org/10.1016/j.glohj.2022.03.001 doi:10.1016/j.glohj.2022.03.001.

39. van Heerden A, Young S, Park CS. Use of social media big data as a novel HIV surveillance tool in South Africa. PLoS One. 2020; 15(10) doi:10.1371/journal.pone.0239304.

40. Liyanage H, Liaw ST, Jonnagaddala J, Schreiber R, Kuziemsky C, Terry AL, et al. Artificial intelligence in primary health care: Perceptions, issues, and challenges. Yearbook of Medical Informatics. 2019; 28(1):41-6. doi:10.1055/s-0039-1677901.

41. Oosterhoff JH, Doornberg JN. Artificial intelligence in orthopaedics: False hope or not? A narrative review along the line of gartner's hype cycle. EFORT Open Reviews. 2020; 5(10):593-603.

42. Car J, Sheikh A, Wicks P, Williams MS. Beyond the hype of big data and artificial intelligence: Building foundations for knowledge and wisdom. BioMed Central. 2019:1-5.

43. Adams MD, Kanaroglou PS. Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models. Journal of Environmental Management. 2016; 168:133-41. doi:10.1016/j.jenvman.2015.12.012.

44. Dominici F, Peng RD, Barr CD, Bell ML. Protecting human health from air pollution: Shifting from a single-pollutant to a multipollutant approach. Epidemiology 2010; 21(2):187-94. doi:10.1097/EDE.0b013e3181cc86e8.

45. Department of Health. National digital health strategy for South Africa (2019-2024). Pretoria, South Africa. 2019.

46. United Nations. Transforming our world: The 2030 agenda for sustainable development. Sustainable development knowledge platform. 2015. Available from: https://sustainabledevelopment.un.org/post2015/transformingourworld.

47. Department of Health. National climate change response plan white paper. 2011.

48. Health Effects Institute. The state of air quality and health impacts in Africa: A report from the state of global air initiative. 2022. Available from: https://www.stateofglobalair.org/sites/default/files/documents/2022-10/soga-africa-report.pdf.

49. Ostro B, Spadaro JV, Gumy S, Mudu P, Awe Y, Forastiere F, et al. Assessing the recent estimates of the global burden of disease for ambient air pollution: Methodological changes and implications for low- and middle-income countries. Environmental Research. 2018; 166:713-25. doi:10.1016/j.envres.2018.03.001.

50. Tshehla C, Wright CY. 15 years after the national environmental management air quality act: Is legislation failing to reduce air pollution in South Africa? South African Journal of Science. 2019; 115(9-10):1-4.

51. Lokotola C, Wichmann J, Wright C. Effect modification of temperature on air pollution associated with hospital admission for respiratory diseases in Cape Town, South Africa. Environmental Epidemiology. 2019; 3:249.

52. Govender K, Girdwood S, Letswalo D, Long L, Meyer-Rath G, Miot J. Primary healthcare seeking behaviour of low-income patients across the public and private health sectors in South Africa. BMC Public Health. 2021; 21(1):1-10.

53. Olutola B, Wichmann J. Does apparent temperature modify the effects of air pollution on respiratory disease hospital admissions in an industrial area of South Africa? Clean Air Journal. 2021; 31(2):1-11. doi:10.17159/caj/2021/31/2.11366.

54. Department: Statistics South Africa. Mortality and causes of death in South Africa: Findings from death notification 2018. 2021.

55. Davalos AD, Luben TJ, Herring AH, Sacks JD. Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. Annals of Epidemiology. 2017; 27(2):145-53.e1. doi:10.1016/j.annepidem.2016.11.016.

56. Chen J, de Hoogh K, Gulliver J, Hoffmann B, Hertel O, Ketzel M, et al. Development of Europe-wide models for particle elemental composition using supervised linear regression and random forest. Environmental Science & Technology. 2020; 54(24):15698-709. doi:10.1021/acs.est.0c06595.

CHAPTER 2: LITERATURE REVIEW

This chapter presents a literature review on air pollution, its adverse health effects, and the relevant legislation, both internationally and in South Africa. The designated priority areas within South Africa and their causes are also explained. The literature review also summarises evidence of respiratory and cardiovascular disease associated with poor air quality exposure. An introduction to Artificial Intelligence and Machine Learning is given, along with an overview of their current use in general, medical, and public health research.

2.1. AIR QUALITY

Good air quality is essential for good health, but industrialisation and urbanisation have reduced the quality of the air within a population.[1] Air pollution can be defined as the presence of one or more contaminants, such as dust, fumes, gas, mist, odour, smoke, or vapour, that are introduced to the air and is often a health risk.[2] Activities such as rapid urbanization, social-economic problems, and climate change are a few contributing factors to the poor air quality in Africa.[3] Accelerated population growth, mainly in urban areas, has led to the increased use of vehicles, solid fuels for cooking and heating, and poor waste management practices, which have contributed to the rising threat of air pollution in many developing nations.[4-6] The influence of meteorological conditions and seasonality also contribute to poor air quality.[7-8] Dependent on the combination of different meteorological conditions, air pollution concentration levels can increase or decrease within the atmosphere.[9] Air pollution can be categorised into two groups – indoor air pollution and outdoor air pollution.

2.1.1. INDOOR AIR POLLUTION

A healthy indoor air environment is very important, since exposure to air pollutants is higher in areas where people spend most of their time.[10-11] Thus far, the most common pollutants found in indoor residential buildings are carbon monoxide (CO), Ozone (O_3), particulate matter ($PM_{2.5}$), total volatile organic compounds (TVOC) and greenhouse gases such as carbon dioxide (CO_2), however, this might differ dependant on the geographical location of the buildings.[12-14] There are four principles for maintaining good indoor air quality, these include keeping the environment dry, minimising indoor emissions, protection against outdoor air pollution, and good ventilation.[15] Poor

ventilation within a closed area is one of the largest contributing factors to poor indoor air quality, which can include a room located in a residential area or commonly used buildings such as office spaces.[11,16-17]

Additionally close proximity to a source of incomplete combustion of biomass in a poorly ventilated area increases the levels of indoor air pollution.[18] There are numerous sources of indoor air pollution, including environmental tobacco smoke (ETS), cooking using biomass (i.e. coal and wood), the burning of incense, and burning of mosquito coils.[19] Outdoor air pollution sources can impact indoor air quality, e.g. traffic emissions entering nearby residential buildings.[20] Indoor air quality can also be influenced by meteorological conditions that affect the air pollutant concentration levels.[10,12,21] This study does not focus of the adverse health effects associated with indoor air pollution, but rather those adverse health effects associated with outdoor pollution.

2.1.2. OUTDOOR POLLUTION

The sources of outdoor air quality vary according to the region. However, anthropogenic activities, such as the burning of fossil fuels, are common sources of air pollution.[22] The common sources of air pollution within South Africa include traffic emissions, coal burning for power generation, the domestic use of wood, coal, and paraffin, and industrial emissions.[23-25] Due to high poverty levels in African countries, there has been accelerated urbanisation, which has left governments ill-equipped to control increased environmental issues such as air pollution.[26] The surge in migration from rural to urban areas has also increased traffic emissions and industrial activity, which directly contributes to poor outdoor air quality.[27] Areas located near in the designated harbours in South Africa are also at risk of exposure from shipping emissions.[28-30] South Africa also has designated national priority areas that have recorded exceedingly high air pollution concentration levels or are at risk of experiencing exceedingly high levels due to future industrial activity. These priority areas have emission problems that come from primary and secondary metallurgical operations, brick manufacturing, petrochemical industry, biomass burning, domestic fuel burning, iron, steel and ferro alloys, mining, power generation, other small industries, transportation, and waste burning.[31-33]

With the increase of epidemiological evidence, the World Health Organization (WHO) guidelines have become stricter with its continuous goal to promote cleaner ways of living and produce cleaner air.[34-35] The 2005 WHO guidelines provide a list of air quality criteria, which include $PM_{2.5}$ and PM_{10}, NO_2, SO_2, O_3, CO, Pb, and benzene and are based on prior epidemiological evidence[36] and are continuously adapted to reflect current emerging research.[37] The evidence base for adverse health effects related to short- and long-term exposure to the criteria air pollutants has become much larger and broader.[38] As of 2021, the WHO updated to stricter interim air quality guidelines (Table 2.1),[35] reducing the $PM_{2.5}$ guideline to a daily limit of 15 µg/m^3 and a yearly limit of 5 µg/m^3. Although the updated $PM_{2.5}$ guideline is to reduce the health risk associated to exposure, the updated guideline does not seem attainable in many parts of the world. Pai et al., estimated that 90% of the world population are already exposed to $PM_{2.5}$ levels above the guideline.[39] Pai et al., further explains that in a scenario where $PM_{2.5}$ levels decline significantly with no anthropogenic emissions, sources such as natural dust and fires in many parts of the world will still experience annual $PM_{2.5}$ higher than the 5 µg/m^3 guideline.

Table 2.1: WHO guidelines for criteria air pollutants as of 2021.

Pollutant	Averaging Period	Concentration
PM_{10}	24 Hours*	45 µg/m^3
	Annual	15 µg/m^3
$PM_{2.5}$	24 Hours*	15 µg/m^3
	Annual	5 µg/m^3
O_3	Peak season**	60 µg/m^3
	8 Hours	100 µg/m^3
NO_2	Annual	10 µg/m^3
	24 Hours*	25 µg/m^3
SO_2	24 Hours*	40 µg/m^3
CO	24 Hour*	4 mg/m^3

* 99th percentile (i.e., 3 – 4 exceedance days per year)**Average of daily maximum 8 – hour mean O_3 concentration in the six consecutive months with the highest six-month running average O_3 concentration

These guidelines have also been revised to include other forms of particulate matter, such as black and elemental carbon, sand, and dust.[35] In order to manage and control deteriorating urban air quality, multiple countries from different socio-economic statuses have implemented and adapted efficient and effective urban air quality management plans.[40-41] These air quality management plans usually provide a holistic strategy that includes systematic sampling, monitoring and analysis, modelling, and control protocols.[40-43] The recent sustainable development goals (SDGs) established

in 2015 set the ambitious goal to reduce emissions by 2030.[44] The 2021 convention on climate change had developed and developing countries pledge to reduce the use of coal as an energy source and encouraged investment in renewable energy.[45]

The concerns of increasing air pollution can also be attributed to the detrimental effects of climate change. Climate change is a continual global challenge and research has observed a domino effect from climate change to air pollution, and on to health effects.[46] Climate and weather have a strong influence on the spatial and temporal distribution of air pollution concentrations, causing faster formation of ozone when greater sunlight and higher temperatures occur.[47] Climate change threatens to continuously change meteorological conditions and may increase the formation of secondary pollutants, such as surface ozone[48] and $PM_{2.5}$.[49] Conversely, air pollution emissions also have a detrimental effect on climate change, since PM has a warming and cooling effect, which can affect precipitation and regional circulation patterns and, by extent, indirectly affect meteorological conditions.[50] In 2022 the "Integrated Assessment of Air Pollution and Climate Change for Sustainable Development in Africa" was created, which considers strategies, policies, and measures to mitigate air pollutants, while supporting development, human health, and wellbeing in Africa on a warming planet.[51]

The air pollution concentration levels are also influenced by external factors. These external factors include meteorological conditions such as temperature, relative humidity, rainfall, and wind conditions (including seasonal variations throughout the year).[52-53] Countries with dry colder seasons will often experience higher air pollution concentration levels during this time.[54-57] Stronger inversion will occur during this season, because there are often low wind speeds and minimal precipitation.[55,58-60] During the warmer seasons, the increased expected rainfall can assist in reducing the presence of some air pollutants.[61] Warmer conditions also assist in the easier release of emissions out into the troposphere, away from direct contact of the population.[56] Dry winter periods and rainy summers are the seasonal conditions found in parts of South Africa such as Gauteng and other provinces which are away from the coast.[62-63] For provinces found in high priority areas in South Africa higher concentrations of PM and black carbon have been observed in late winter early spring, i.e., July, August

and September, while the opposite has been observed for O_3 where higher concentrations occur in warmer months.[64-67]

Reducing or phasing out of the use of fossil fuels has the potential to limit global warming by up to 2°C.[68] In some studies temperature and apparent temperature have been identified as effect modifiers in air pollution studies, where warmer days increase the presence of air pollution.[69-72]

2.1.3. SOUTH AFRICA AMBIENT AIR QUALITY

The South African National Department of Health has recognised the health risk associated with air pollution, identifying the potential harm air pollution can have on human health.[73] The South African government has designated the Department of Forestry Fisheries and the Environment (DFFE), formally known as the Department of Environmental Affairs (DEA), as the government body that regulates the country's air quality standards. These standards have been adapted from the WHO air quality guidelines.[36] The inclusion of these standards in legislation is to enforce the law and encourage compliance at national, regional, and district levels within the country. Air quality standards are categorized as criteria pollutants which include O_3, CO, SO_2, NO_2, $PM_{2.5}$, and PM_{10}.[74]

The initial efforts in South Africa's air quality management began over 50 years ago with the Atmospheric Pollution Prevention Act (AAPA) in 1965. However, this initiative mainly focused on addressing, but not mitigating, industrial emissions.[75] Due to the insufficient coverage and shortcomings of the AAPA, the 2004 National Environmental Management: Air Quality Act (NEMA AQA) was established.[76] The NEMA AQA established a more comprehensive means of addressing multiple sources of pollution, including industrial and domestic contributors.

The NEMA AQA provided a more wide-ranging legislative plan to manage environmental factors in South Africa. The publication of the Air Quality Act in 2005 significantly addressed environmental management in South Africa, although the set air quality standards are lenient in comparison to the WHO guidelines (Table 2.2).[77-78] Control guidelines for ambient air were included in the Air Quality Act, addressing multiple parties including the polluter, supervisory bodies, and the general public.[74] Implementation of the Act was decentralised through the completion of Air Quality

Management Plans (AQMP); these were also reviewed mid- and long-term.[79] The NEMA AQA also seeks to reduce and manage air quality by enforcing compliance, setting ambient and emission standards, identifying and quantifying all pollutant sources, as well as normalising and standardising air quality monitoring management.[80]

Table 2.2: South African National Ambient Air Quality Standards.

Pollutant	Averaging Period	Concentration	Frequency of Exceedances	Compliance Date
PM_{10}	24 Hours	75 µg/m^3	4	1 January 2015
	1 Year	40 µg/m^3	0	1 January 2015
$PM_{2.5}$ (added in 2012)	24 Hours	40 µg/m^3	4	1 January 2016 - 31 December 2029
		25 µg/m^3	4	1 January 2030
	1 Year	20 µg/m^3	0	1 January 2016 - 31 December 2029
		15 µg/m^3	0	1 January 2030
NO_2	1 Hour	200 µg/m^3	88	Immediate
	1 Year	40 µg/m^3	0	Immediate
SO_2	10 Minutes	500 µg/m^3	526	Immediate
	1 Hour	350 µg/m^3	88	Immediate
	24 Hours	125 µg/m^3	4	Immediate
	1 Year	50 µg/m^3	0	Immediate
Ground-level O_3	8 Hours	120 µg/m^3	11	Immediate
CO	1 Hour	30 mg/m^3	88	Immediate
	8 Hour	10 mg/m^3	11	Immediate
Lead	1 year	0.5 µg/m^3	0	Immediate
Benzene	1 year	5 µg/m^3	0	1 January 2015

In urban areas such as the City of Tshwane, Pretoria, the centralised AQMP was initially implemented during in 2006-2008. [81] The AQMP's objectives are to achieve and sustain acceptable air quality levels as well as minimising health risks and harm

to the environment in Pretoria. The AQMP stipulate the importance of air quality measurement "tools," which include air quality and meteorological monitoring and atmospheric dispersion modelling. The mentioned tools offer a comprehensive emission inventory, to facilitate the effective characterisation of spatial and temporal variations in air pollutant concentrations. [81] The City of Tshwane has nine monitoring stations distributed across the municipality, Bodibeng, Booyseng, Ekandustria, Hammanskraal, Mamelodi, Olivienhoutbosch, Pretoria West, Rosslyn and Tshwane Market. However only three out of the nine stations record $PM_{2.5}$ concentration levels. [82-83] A report by Zunckel et al., state the main air pollutants in the area include, SO_2, oxides of nitrogen (NO_X), PM_{10}, $PM_{2.5}$, CO and VOC. The identified sources of air pollution in the City of Tshwane include industry, mining, domestic coal burning and biomass burning.[83]

In addition to the establishment of the NEMA AQA 39 of 2004,[74] the DFFE identified and stated three National Priority Areas: the Vaal Triangle Airshed Priority Area (VTAPA), the Highveld Priority Area (HPA), and the Waterberg-Bojanala Priority (WBPA) areas in 2006, 2007, and 2012, respectively.[84] The three areas, shown in Figure 2.1, the VTAPA and HPA are demarcated for their history of poor air quality owing to industrial activities that include mining or the risk of extremely poor air quality and require specific air quality management action to rectify their poor air quality situation. The WBPA however is designated a high priority area because of the potential of poor air quality due to emerging industrial activity in the area.[85] These areas are mainly located in the more industrial provinces of the country, namely Gauteng, Mpumalanga, Limpopo, and parts of the Free State and North West.

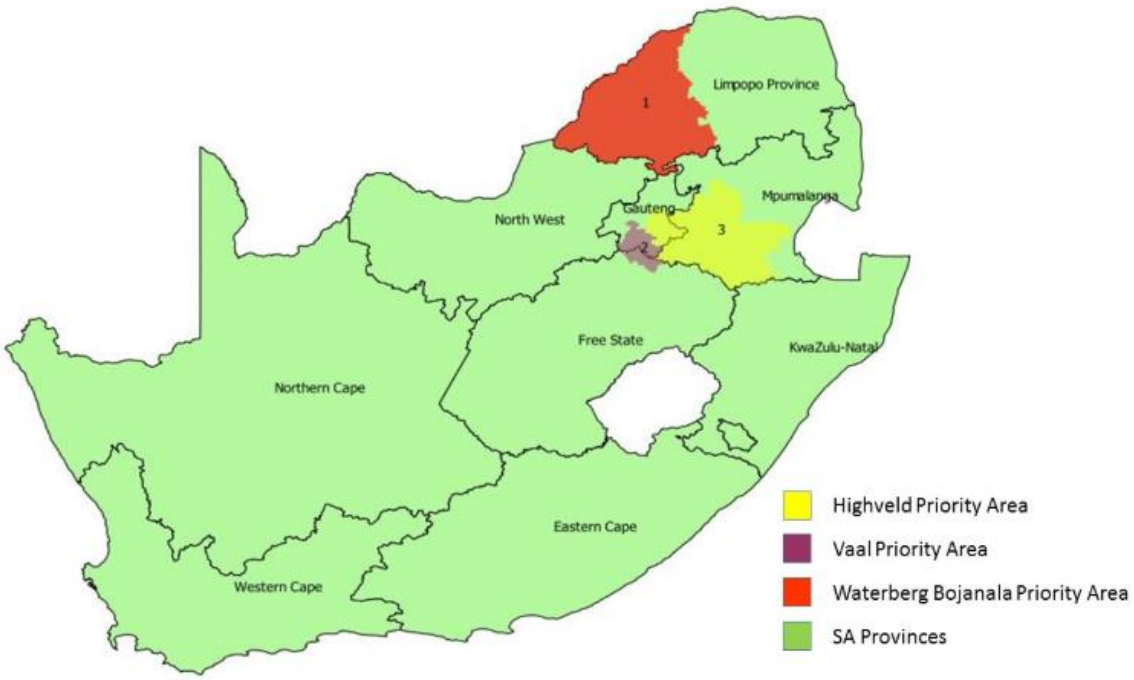


Figure 2.1: Map of the designated priority areas within South Africa.[86]

In 2006 the VTAPA, located in the Gauteng and Free State provinces, was declared the first priority area in the country due to the concern of the rising particulate matter concentration levels.[87] The area is highly industrialised and this heavily contributes to the high levels of air pollutants, such as PM, NO_2, and other noxious and offensive gasses.[31] There are other contributing sources of the high air pollution levels, including domestic and agricultural activities; these high concentration levels have caused concern to the health of the communities found in this area.[31,87] A 2021 source apportionment study in the VTAPA suggested eight sources of $PM_{10-2.5}$ and $PM_{2.5}$; these included industry, coal burning, wood and biomass, waste burning, secondary aerosols, vehicles, dust-related, and, interestingly, aged sea salt.[67] Health studies within the area have also shown the adverse effects on the community from the exposure to $PM_{2.5}$, PM_{10}, NO_2, and SO_2.[71,88] Although there is a decentralised system for air quality monitoring in the area, the lack of funds and efficient monitoring have slowed improvement to monitoring pollution within the VTAPA.[80,89-90]

The HPA was then declared the second national priority area in 2007, and this area is located in Mpumalanga and Gauteng provinces.[32] The HPA has a similar profile to the VTAPA, with its main sources of pollution being mining, coal-fired power plants, industrial and chemical manufacturing, agricultural activity, motor vehicles, and

domestic fuel burning.[84] Health studies have also shown the health risk posed to the community from the high concentration levels.[88,89] The most recently declared priority area in South Africa is the WBPA, which was declared in 2012 and is predominantly located in the Limpopo and North West provinces.[33] The declaration of the priority air quality monitoring network within this area is mainly driven by the expected development and ongoing coal mining and metallurgical activity.[92-93] SO_2, NO_x, CO, and PM_{10} are the main pollutants of concern within the area, with the main causes of pollution being mining and industrial activity, biomass, residential, and motor vehicle emissions.[94] The demand to set it as a priority area was due to the potential of mining and industrial expansion in the area, which has been estimated to increase annual emissions of SO_2, NO_2, and PM_{10} by 370%, 640%, and 530%, respectively, by the year 2030.[95]

Although South Africa has current legislation and AQMPs are to be implemented in the 278 municipalities, there are only 121 government-managed air monitoring stations that report air quality to the national air quality information system.[80,96] Additionally, although there is a set framework to normalise and standardise air quality management across the country, it seems that regulatory authorities are not scrutinising industry pollution reports or demanding follow-up reports of those who are non-compliant.[80] In spite having the most comprehensive air quality legislation in Africa, South Africa is still struggling with the monitoring and enforcement of air pollution legislation. In general, the rapid urbanisation in Africa is not only worsening the quality of air to communities, but is not being preventatively prioritised because of the existing societal challenges on the continent including poverty, the intensifying effects of climate change, and recovery from the COVID-19 pandemic.[97]

2.2. AIR POLLUTION AND HEALTH

There is an increasing Global Burden of Disease (GBD) associated with air pollution.[1,98] Short- and long-term exposure to poor air quality has been associated with harmful health outcomes affecting different age groups.[1,99-101] Air pollution is considered the largest environmental health risk factor and makes a significant contribution to the GBD.[102-103] The quantification of the GBD in relation to air quality has seen great advancements, such as improved chemical transport modelling and concentration-response functions.[104] Such advancements have helped in relating

ambient air pollution levels and the risk of negative health effects. However, low-to-middle-income countries (LMIC) still face challenges in utilising these improvements.[105] In 2019, an estimated 12% of all deaths were attributable to outdoor and household air pollution.[106-107] Approximately 20% of global cardiovascular deaths were attributable to air pollution, and was the fourth-highest-ranking risk factor for mortality. [106-107] Fine particle air pollutants, i.e. PM, was the largest environmental risk factor worldwide, responsible for a substantially higher number of attributable deaths than some behavioural risk factors, such as alcohol use, physical inactivity, and high sodium intake.[108] Air pollution was ranked one of the top-five highest mortality risk factors globally and has further been associated with an estimated 6.7 million deaths, as shown in Figure 2.2.[102] Poor indoor air quality has also been associated with negative health impacts due to the prolonged time individuals spend within closed doors.[17] In 2019, the GBD estimated 2.3 million deaths to exposure from household air pollution, with sub-Saharan Africa accounting for 30% of those deaths.[109]

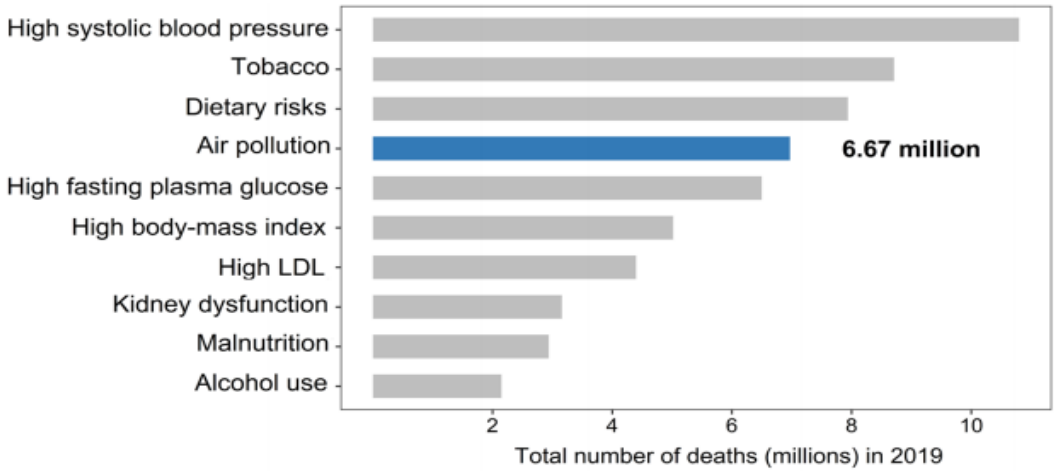


Figure 2.2: Global ranking of risk factors for total deaths from all causes for all ages and sexes in 2019. [102].

The association between mortality and air pollution exposure has been identified in a number of studies. In nineteen European cohorts, it was shown that long-term exposure to $PM_{2.5}$ can increase the risk of all-cause mortality by 6% to 13%.[106] Additionally, long-term exposure to $PM_{2.5}$ at considerably low concentrations, between 10 µg/m^3 and 20 µg/m^3, can increase the risk of mortality.[111] Furthermore, long-term exposure to 'low-level' $PM_{2.5}$ and NO_2 can increase the risk of mortality by 4.4%.[112] Short-term exposure to SO_2, NO_2, and $PM_{2.5}$ have also shown a moderate association with chronic pulmonary obstructive disease (COPD) related mortality.[113] An Australian

cohort also showed that exposure to relatively low levels of $PM_{2.5}$ and NO_2 still showed increased risk of all-cause mortality in the elderly.[112] LMIC are reported to account for a large proportion of deaths attributed to $PM_{2.5}$ exposure.[114] Global reports stated that South Africa had approximately 24 800 and 601 premature deaths due to $PM_{2.5}$[115] and O_3 [116] exposure, respectively. However the deaths attributed to $PM_{2.5}$ and O_3 are based some of these findings on model estimates. Furthermore, the exposure-response functions used are derived from epidemiological studies conducted in developed countries.[105] There has been an increase in African-based research to show the negative associations from exposure to air pollution, but there is still much to be done to explore these associations.[114]

There are a number of pollutants that have been associated with negative health effects. Some of these studies show a closer association of CO_2 and $PM_{2.5}$ with the increased risk of infant mortality.[117-118] Other studies show the exposure to $PM_{2.5}$, PM_{10}, NO_2, SO_2, and black carbon (BC) can increase the risk of cardiovascular and all-cause mortality,[119-121] as well as respiratory and cardiovascular hospital admissions.[70-72]

Respiratory and cardiovascular diseases have the strongest associations with air pollution.[107] Pollutants can enter the respiratory and circulatory system, much easier than they would enter other systems within the body, and cause more evident acute adverse health effects.[122-126] There is growing evidence associating air pollution exposure with a range of conditions. However, for the purpose of this study, the effects on respiratory and cardiovascular diseases are highlighted.

The DFFE awards tenders for the three priority areas and municipalities such as the City of Tshwane, there cross-sectional epidemiological studies are conducted through independent contractors to evaluate AQMP. However, the results of the health studies are not published in peer viewed journals or scientific reports, presented at conferences nor published in the public domain. Thus there are relatively few South African health related studies available on the detrimental health effects of air pollution in priority areas and municipalities such as the City of Tshwane.

2.2.1. AIR POLLUTION AND RESPIRATORY DISEASE

Specific pollutants such as PM, O_3, and NO_2 can cause airway inflammation, airway hyper-responsiveness, and oxidative stress in the lungs, although the mechanism that causes this is still being researched.[127] PM varies in size from coarse to finer PM (Figure 2.3) and these sizes enable it to enter and distribute differently in the respiratory system (Figure 2.4).

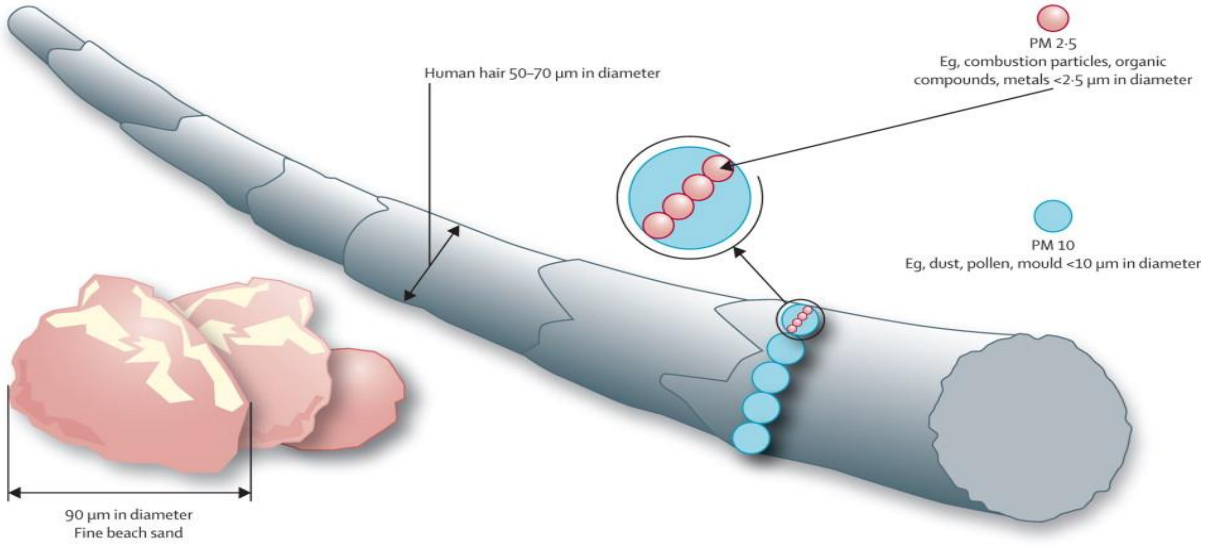


Figure 2.3: Modified image of $PM_{2.5}$ and PM_{10} in comparison to a human hair strand and beach sand.[127]

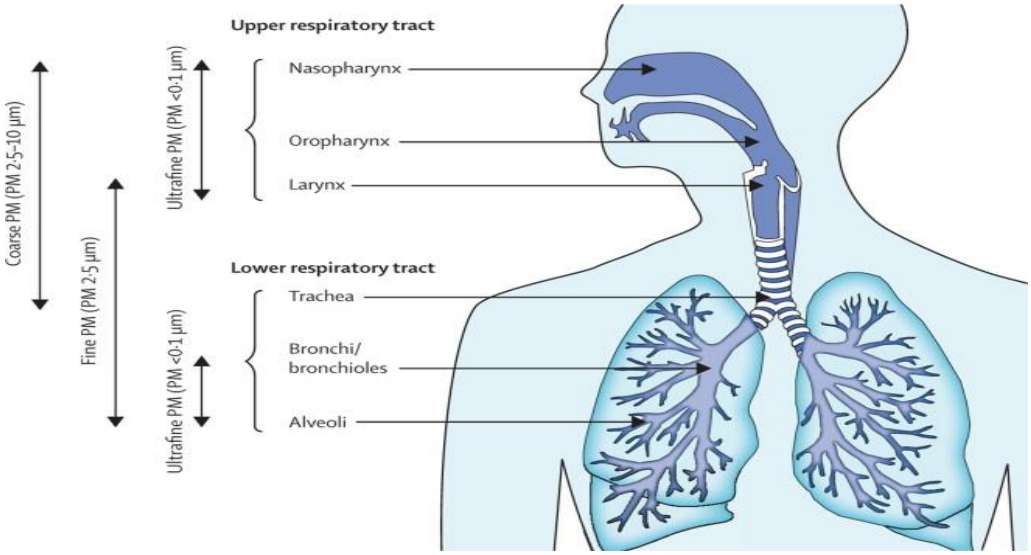


Figure 2.4: The size distribution of the different sizes of PM in the respiratory system.[127]

Thus, the exposure to air pollutants such as PM have been associated with multiple respiratory conditions and the aggravation of respiratory diseases such as COPD, childhood asthma,[128-129] and lung cancer.[130] Short-term air pollution exposure can also cause an irregularity in the respiratory microbiome that can, in turn, affect lung function.[131-135] Exposure to outdoor air pollution not only leads to a decline in lung function, but can also do more harm in susceptible groups like the elderly, children, and expectant mothers.[136] Susceptibility to air pollution exposure increases when an individual has pre-existing or present respiratory conditions, such as COPD, which has been found to make individuals vulnerable to respiratory inflammation when exposed to pollutants such as nitric oxide (FeNO), CO, NO_2, NO, and SO_2.[137]

Evidence shows that short-term exposure to ambient air pollution can increase respiratory hospital admissions.[138-140] Short-term exposure to a 10 $\mu g/m^3$ increase of $PM_{2.5}$, NO_2, and SO_2 was associated with an increased risk of developing COPD-related illness by 2.5% (95% CI: 1.6–3.4%), 4.2% (95% CI: 2.5–6.0%), and 2.1% (95% CI: 0.7–3.5%), respectively.[113] A 2023 case-crossover study conducted in Pretoria, South Africa found that exposure to $PM_{2.5}$ increased respiratory hospital hospitalisation by 2.7% (95% CI: 0.6, 4.9) per 10 $\mu g/m^3$ increase.[141] A study showed short-term exposure to O_3 is associated with increased lengths of hospital stays in children, however, there were inconsistencies found with other pollutants such as NO_2 and SO_2.[142] Exposure to PM and NO_2 can lead to a continued decline of respiratory health.[143-144] A 1992 study in the Transvaal i.e., now includes the VTAPA, health questionnaires were distributed to parents in this area and 65.7% had reported that their children suffered from an upper respiratory infection in that year.[145] Air pollution exposure can aggravate asthma and wheezing in children, which can contribute to inconsistent school attendance.[23, 85, 138]

Short-term exposures to $PM_{2.5}$, NO_2, and O_3 have also been linked to increasing risk of asthma and mortality in the elderly.[146-147] Studies done in Cape Town, South Africa, during 2001-2006, showed an increase in RD mortality of 1.3% (-1.4%; 4.0%) and 2.0% (-1.6%; 5.7%) per inter-quartile range increase in PM_{10} (12 $\mu g.m^{-3}$) and NO_2 (12 $\mu g.m^{-3}$), respectively. In contrast, a decrease of -0.5% (-3.6%, 2.6%) was observed per inter-quartile range increase (8 $\mu g.m^{-3}$) in SO_2.[119] Follow-up studies observed a mortality risk of 0.4% (-0.4%; 1.1%); 1.2% (-0.2%; 2.6%); and -1.9% (-3.7%; 0.0%) for

RD, following a 10 $\mu g.m^{-3}$ increase in the two-day cumulative average of PM_{10}, NO_2, and SO_2 during 2006-2010, respectively.[148]

A systematic review that analysed short-term exposure to O_3, SO_2, NO_2, and the association to emergency visits for asthma pooled relative risks per 10 $\mu g/m^3$ increase of ambient concentrations, showing 1.008 (95% CI: 1.005, 1.011) for maximum eight-hour daily or average twenty-four-hour O_3 exposure; 1.014 (95% CI: 1.008, 1.020) for average twenty-four-hour NO_2 exposure; 1.010 (95% CI: 1.001, 1.020) for twenty-four-hour SO_2 exposure; 1.017 (95% CI: 0.973, 1.063) for maximum one-hour daily O_3 exposure; 0.999 (95% CI: 0.966, 1.033) for one-hour NO_2 exposure; and 1.003 (95% CI: 0.992, 1.014) for one-hour SO_2 exposure. As such, the findings of this systematic review show a correlation between exposure and increased risks of asthma-associated emergency room visits and hospital admissions.[149]

2.2.2. AIR POLLUTION AND CARDIOVASCULAR DISEASE

Detrimental effects to cardiovascular health from exposure to ambient air pollution has been found to affect both high-income countries (HIC) and LMIC[38] There has been a decline in cardiovascular mortality in HIC,[150] but mortality in LMIC has had an incline in cardiovascular related deaths.[151] Short-term exposure to $PM_{2.5}$ and PM_{10} increase the risk of myocardial infarction, where exposure to $PM_{2.5}$ showed to have a higher risk than PM_{10}.[150] Gaseous and particulate air pollutants have shown a close temporal association with hospital admissions due to stroke.[153-154] Although PM has a strong association to both cardiac mortality and hospitalisation, gaseous pollutants such as SO_2 and nitrogen oxides have similar associations.[155-158] Exposure to SO_2 is not only a suggested indicator for cardiovascular hospital admissions, but has also been found to trigger cardiovascular conditions.[159]

Long-term exposure to NO_2 has also been associated with cardiovascular mortality.[160] A study showed increased risk of 3.4% and 2.6% in CVD mortality from long-term exposure to NO_2 and SO_2, respectively, in spite of the relatively low exposure levels that were similar to European measures.[121] A similar study, following a 10 $\mu g/m^3$ increase in the 2-day cumulative average of PM_{10}, NO_2, and SO_2 during 2006-2010 in Cape Town, Johannesburg, and Durban, also showed a risk for CVD mortality at 1.0% (0.3%; 1.7%), 1.0% (-0.3%; 2.3%), and 0.9% (-0.9%; 2.7%), respectively.[148]

Simultaneous exposure to two or more pollutants can influence the onset of disease, for instance, short-term exposure to both $PM_{2.5}$ and NO_2 has been found to increase risk of admissions for arrhythmia.[161] The joint effects of $PM_{2.5}$, SO_2, NO_2, and CO have shown to increase risk of hospitalisation and death caused by congestive heart failure.[151] Research has also shown an associative relationship between stroke and short-term exposure to SO_2, NO_2, CO, and O_3.[163-165]

The elderly and women are two of the commonly identified vulnerable subgroups at risk of CVD.[166-168] Research has shown that acute exposure to SO_2, NO_2, CO, and $PM_{2.5}$ can increase the risk of cardiovascular events in the elderly.[159,164-165,169] Individuals aged above 60 years old are at a higher risk of developing comorbidities, such as hypertension and diabetes, that, in turn, increase their risk of developing CVD.[167] In addition to developing these comorbidities, acute exposure to PM can further increase the risk of CVD in the elderly.[156,168,170-172] Women have also shown to be at a higher risk of developing CVD from air pollution exposure compared to males.[166,168,173-174] Females may share common risk factors with males, however, due to their reproductive systems, females become prone to hypertensive disorders during pregnancy and menopause, and can develop gestational diabetes.[175] Although cardiovascular events are more prevalent in women, men have been found to have higher CVD mortality rates than women.[176]

2.2.3. AIR POLLUTION AND OTHER MORBIDITIES

There is a potential association between the exposure to NO_2 and O_3 and developing cancer.[177] However, long-term exposure to ambient PM is a more common risk factor for lung cancer.[130,178] The risk of developing cancer from exposure to vapours, metallic compounds, and metals has not been found significant.[178] Dependent on the sources of the pollutants, the chemical composition elements found are either carcinogenic or non-carcinogenic.[179] Lung cancer is the most commonly associated cancer from exposure to air pollution.[180-181] There is also available, but limited, research showing the association between bladder cancer and exposure to $PM_{2.5}$ and NO_2.[182-183] However, in a study comprising of fifteen European cohorts, no association was found between $PM_{2.5-10}$, NO, NO_2, or organic carbon exposure and bladder cancer.[184] There have also been linkages between air pollution exposure and breast cancer,[185] although

there is a clearer breast cancer association with NO and NO_x.[186] There has not been a clear association found between exposure to PM and breast cancer.[185,187]

Air pollution exposure has also been associated with different communicable diseases. One study showed a positive association between PM_{10} and type 2 diabetes.[188] With an increased exposure of 10 $\mu g/m^3$ PM_{10}, the odds of developing type 2 diabetes increases by 1.23.[189] Although exposure to PM has been shown to affect the endocrine system, which leads to developing type 2 diabetes, further research is needed as these conclusions are not definitive.[190-192] Studies have also shown that expectant mothers are at a higher risk of developing Gestational Diabetes Mellitus (GDM) upon exposure to PM and O_3.[193-195] Exposure to air pollution for expectant mothers also presents a risk to their unborn children. Exposure to PM_{10}, $PM_{2.5}$, SO_2, and O_3 have been associated with an increased risk of stunted growth and low birthing weights.[196-198] A study in China showed that exposure to PM_{10}, $PM_{2.5}$, SO_2, and NO_2 during pregnancy can increase the risk of congenital heart disease in unborn children.[199] Exposure to traffic related air pollutants, such as CO and NO_2, has also been associated with the onset of depression in the elderly.[200] Although the literature is limited, there have been potential causal associations between air pollution and poor mental health.[201]

More recently, since the start of the global COVID-19 pandemic, there have been studies that link the presence of air pollution and the onset of viral symptoms. Studies tried to suggest that the spread of COVID-19 could be more easily transmitted in areas with higher levels of air pollution.[202] Research also sought to connect the detrimental effects of air pollution exposure and susceptibility to the more severe effects of the COVID-19 virus.[203] However, these theories were disputed, because the incidence data used in such studies were underestimated in most countries and the conclusions made were not decisive.[204] Regardless, studies had supported a positive relationship between COVID-19 restrictive measures and air pollution. Countries such as China, Italy, Spain, and USA, experienced a 30% reduction in air pollution concentration levels as a direct result of lockdown measures that restricted movement.[205] Preliminary results used by the Council for Scientific and Industrial Research (CSIR) also indicated that strict lockdown measures within South Africa reduced NO_2, SO_2, and O_3 concentration levels in the country.[206]

2.3. STATISTICAL ANALYSIS OF MULTI-POLLUTANT EXPOSURE

The evidence summarised in section 2.2 is based off single or dual pollutant models and do not consider the effects of air pollution mixtures. Recently, characterising the relationship between health outcomes and multi-pollutant mixtures has been emphasized in an attempt to better protect public health and inform more sustainable air quality management decisions.[207-208] Research is now aimed at understanding the health effects of multi-pollutant exposures, i.e. the joint effect of two or more pollutants on a health outcome, as opposed to single pollutant exposure that may be a misrepresentation of the effect on a health outcome.[209-211]

Due to the interactive nature of air pollution, observing the health effects of one pollutant does not provide accurate estimates.[212] The presence of more than one air pollutant can cause a worse effect. For example, PM and O_3 have a synergistic effect and can increase acute respiratory inflammation and other respiratory diseases.[208] Air pollution can also have additive and potentiation qualities, that can double the single effect of one air pollutant or make one pollutant more harmful.[213] Previous studies have shown positive correlations among air pollutants such as PM_{10}, NO_2, and SO_2 that are important to consider in analysis.[70,119] In addition, the correlation between air pollutants and external factors, such as temperature, can affect the analysis of the effect on health.[69-72,207,209] High temperatures in the presence of air pollutants like PM can intensify or increase the likelihood of conditions such as migraines.[207] Therefore, more advanced statistical methods are needed to analyse the joint effects of multiple air pollutants on health.

There are five broad classes of statistical approaches identified for examining associations between short-term multi-pollutant exposures and health outcomes; these are (1) Additive Main Effects, (2) Effect Measure Modification, (3) Unsupervised Dimension Reduction, (4) Supervised Dimension Reduction, and (5) Non-parametric methods.[214] These approaches' largest advantage is the ability to examine multi-pollutant exposure however within the different epidemiologic scenarios the number of pollutants in the exposure mixtures are still relatively limited.[214] Modified statistical approaches have been identified to more accurately estimate independent and joint effects of multiple correlated air pollution exposures on human health.[215] The modified

Classification and Regression Tree (CART) analysis method,[216-217] a non-parametric method, has analysed the effects of a mixture of air pollutants on health outcomes.

2.4. AIR POLLUTION SOURCE APPORTIONMENT

Identifying pollution sources is one of the most necessary and noticeable forms of research in environmental pollution studies.[218] Source apportionment is also vital in determining source-specific activities that are most likely to be responsible for the observed adverse health effects.[219] It is vital to influence policy measures to prevent and control environmental pollution, as well as promote sustainable measures for both economical and societal development.[218] It is important to understand the effects air pollutants pose to the wellbeing of the communities they impact, whether it be its effect on general population, age,[220] sex,[221] or socio-economic status, as some of the divisions of populations. It is equally important to understand the sources and source direction of these air pollutants.[222-223] Source apportionment focusses on identifying what the source of air pollution is and how much this contributes to the total ambient pollutant within the area of interest.[224]

Source apportionment studies are often local studies determining the contributing sources to ambient PM, measured at representative monitoring sites.[224] The purpose of source apportionment helps to identify the mass contribution and percentage within a source of air pollutant (mainly PM), which could be long-range traffic emissions, wood burning, and other anthropogenic sources.[223,225] Source apportionment is a widely used tool to provide quantitative information about source contribution of PM to support air quality control and management.[226] A South African review on source apportionment studies showed that source categories in the country do not vary and the major sources come from biomass and coal burning from industry or domestic use.[227] However the studies reviewed were mainly conducted in urban and rural backgrounds with few studies conducted in industrial areas. Mathuthu et al., recommends more research into air particulate source apportionment techniques and tools in South Africa.[227]

Receptor modelling has been widely used in source apportionment studies, as it estimates contributions from emission sources and links contributions to levels of environmental pollution.[218,228] The more widely-used receptor model techniques

include Chemical Mass Balance (CMB),[229] Principal Component Analysis (PCA),[230] UNIMIX,[231] and Positive Matrix Factorisation (PMF).[218] The use of these techniques are highly dependent on the availability of source profiles, ambient pollutant mass concentrations, and available meteorological data.[218] PMF is the more popular receptor model for source apportionment, which has been reported to be a relatively straightforward application.[226,232-234] Additionally, the PMF software is freely available with a detailed manual for the user to work through and has multiple error estimates for each output to validate.[228] Usage of this methodology greatly increased when the United States Environmental Protection Agency (US EPA) released the latest version of PMF in 2014.[228] PMF is classified as a statistically Unsupervised Dimension Reduction statistical method, according to Davalos et al.[214]

The abundance of sample datasets has been increasing in the recent years, due to a rise in pollution monitoring and the increasing importance of environmental pollution research.[235] The increased datasets are reason to obtain more accurate results from source apportionment and include pollution features such as chemical concentrations and pollution sources.[218] Some traditional techniques only classify sample data according to different seasons or limited pollution durations, based on the meteorological conditions and the variability of sources involved.[218] Another shortcoming of these traditional techniques is the issue of directly identifying significant outliers, that are important to take note of, but may affect the overall results of the source apportionment; this difficulty arises as a result of the large data sets and multiple dimensions.[218]

There has been an increase in the demand for classifying the sample data and detecting outliers using data characteristics that support the traditional receptor models.[218] The current source apportionment studies conducted in Africa depend on the implementation of PMF,[62-63,236-239] with a few exceptions that use PCA.[240] One way to address some apparent drawbacks of traditional source apportionment methods is through using unsupervised Machine Learning (ML) methods.[218,241] The different unsupervised ML methods are discussed in section 2.5.3.

2.5. ARTIFICIAL INTELLIGENCE (AI) METHODS.

2.5.1. ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Artificial Intelligence (AI) is the use of software technology to perform multiple tasks, such as automatic knowledge extraction and pattern recognition from data, decision making, and many more.[242] ML is a branch of AI aimed at enabling computers to learn without being programmed, thus, improving the computer's performance when executing tasks.[243-244] The concept of AI was officially established in 1956 by John McCarthy. Later, in 1959, the AI branch referred to as ML was established by Arthur Samuel, who defined it as the learning feature of intelligence by developing algorithms that extract generalized principles from data.[245] Another branch of AI, a further subset of ML, is deep learning (DL).[245] Figure 2.5 defines the differences between AI, ML, and DL. However, ML is the focus of this project as ML focuses on the enhancement of statistical techniques, and statistical analysis which is a large focus of public health research.

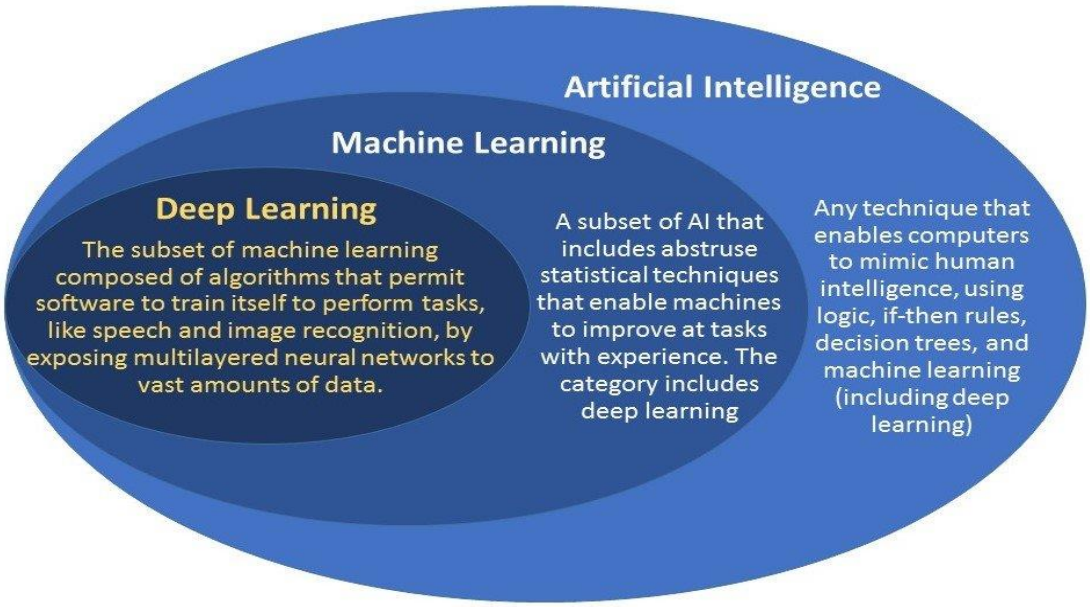


Figure 2.5: Definitions of Artificial Intelligence, machine learning and deep learning.[246]

Machine Learning can perform predictive analytics much faster than human efforts and, as a result, it has the potential to help efficiently increase work productivity.[247] There are two common types of ML, supervised and unsupervised learning. Supervised learning occurs when the dataset is labelled and it involves the individual

labelling of or rebuilding data, which assists in the prediction of outcomes from the provided data.[244,248-249] In contrast, unsupervised learning uses a large amount of unlabelled data and the algorithms establish their own grouping of the data based on available information. [244,248-249]

Although supervised and unsupervised learning are the more popular types of ML, there is another type of ML known as reinforcement learning. Reinforcement ML runs by learning through the recognition of trial and error within its environment.[250-251] An example of this would be building 'intelligent robots', such as self-driving cars.[252] Figure 2.6 illustrates the types of ML, the uses of each type of ML, and method examples for each ML type.

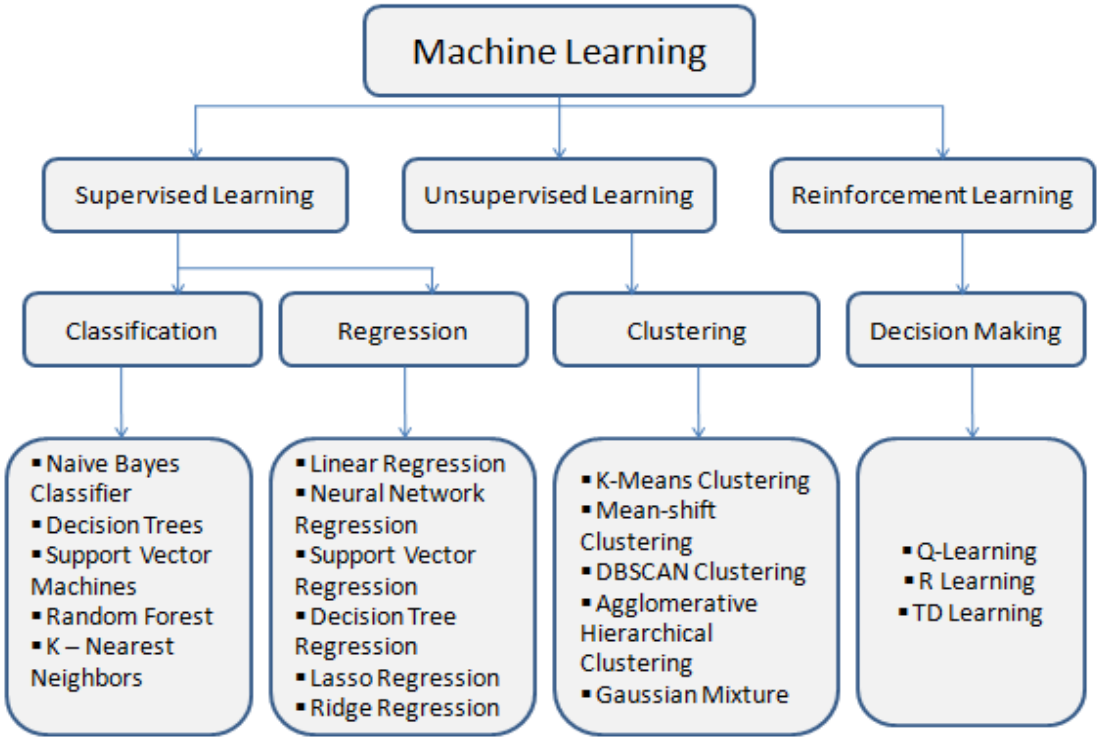


Figure 2.6: Different types of machine learning methods that include supervised, unsupervised and reinforcement machine learning.[253]

2.5.2. MACHINE LEARNING IN HEALTH

Public health data has grown in scale and complexity, and is projected to grow even more in the foreseeable future.[254] There are predictions that advanced big data science approaches, such as AI, can improve health and population outcomes, and this has led to a rise in AI in health care research.[255] However the definition of data

science has been debatable as there are conflicting ideas that it is simply a rebrand of fields such as statistics and computer science.[254,256] Figure 2.7 illustrates where AI was envisioned to elevate the medical industry which shows the pattern of how AI is to assist in improving the process of diagnostics and patient monitoring.

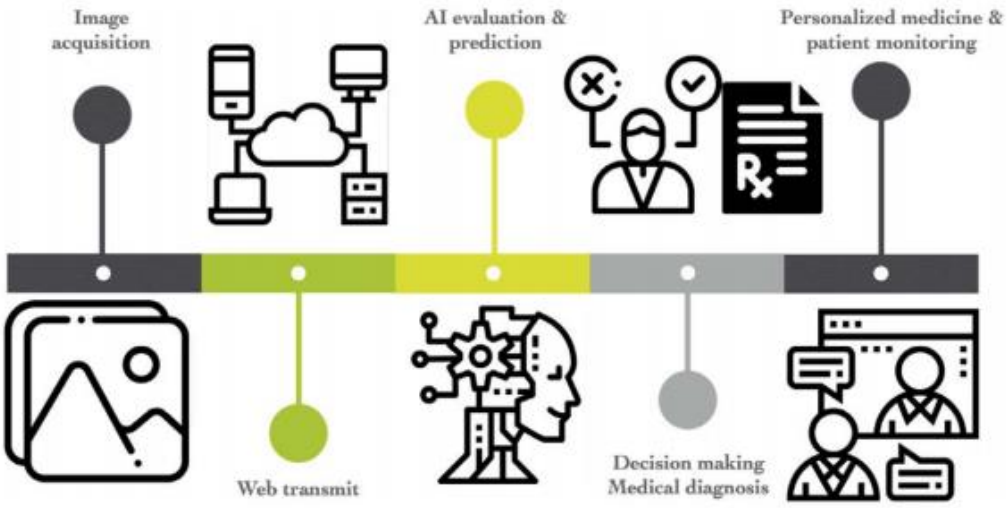


Figure 2.7: An illustration of the envisioned use of AI within diagnostics to improve patient healthcare. [257]

The use of AI applications is not new in the medical and public health context. ML and deep learning have been applied in numerous ways such as AI system-based recommendation of appropriate antibiotics in the 1970s,[258] in diagnostic tools,[245] HIV research,[259] and in clinical monitoring.[260] There are also opportunities to use AI in public health.[261] In LMIC, the main focus of AI interventions has been on health issues such as tuberculosis [262-264] and malaria,[265-266] which used neural networks in diagnostics to detemine positive cases. Other studies included non-infectious diseases in children and infants, preterm birth complications, malnutrition and cervical cancer,[267] by using predictive modelling and natural language processing to assist in the diagnostics and prediction of health complications.

Public health and epidemiology studies have incorporated the use of AI applications. ML has been used is different ways including statistical analysis of large epidemiological data, data mining, screening techniques, infection surveillance and prevention.[268-272] The use of AI applications like ML, has also been considered to assist in achieving some Sustainable Development Goals (SDGs) by predictive modelling, tactile decision making, interactive communication and pattern recognition in the analyses of large-scale interconnected databases to develop environment-

preserving activities.[242] The increased use of ML also has the potential to close the gap for LMIC in achieving the SDGs by using ML as a characterisation tool to estimate risk of disease and implement preventative measures in populations at risk.[267]

Within the context of air pollution epidemiology, ML has used prediction-based or knowledge discovery methods in data mining applications.[268] This often results in the use of classification algorithms that predict categorical outcomes, and regression algorithms for continuous outcomes on single pollutant data.[273] The Pan American Health Organization has summarised subfields of AI that include: [274]

Cognitive Search: The use of AI solutions (such as ML and natural language processing) to incorporate and understand digital content from various sources, such as text, images, video, and machine data. The goal is to improve the relevance of the results generated from a user search.[275-276]

Computer Vision: Training computers to interpret and understand the visual world. Using digital images from cameras and videos and deep-learning methods. Machines can resultantly identify and classify objects more accurately.[277-278]

Deep Learning: A subfield of ML that uses algorithms designed as networks of decisions to learn from data. These networks are often called neural networks and when there are many layers in the network, they are called deep neural networks or deep learning networks. Deep learning can identify diseases based on imaging and can predict health status from electronic health records.[279-282]

Machine Learning: Process of applying training-data to a "*learning algorithm*" The algorithm generates a set of rules, based on identified data patterns. These rules can then be used to classify new data or predict future data. Through using different training-data, the same learning algorithm could be used to generate different models, e.g. pathology prediction and so forth.[283-285]

Natural Language Processing (NLP): NLP automates the ability to read, understand, and derive meaning from human language.[286] Two interesting NLP subfields of particular interest are (i) Natural Language Understanding (NLU) where these algorithms are designed to understand human writings using a coded understanding of grammar, syntax, and semantics;[287] and (ii) Natural Language Generation (NLG),

where the algorithms are designed to automatically transform structured data into plain language. It is considered the opposite of NLU.[286,288]

Robotics: An interdisciplinary research area at the interface of computer science and engineering. The goal of robotics is to design intelligent machines that can assist human activity.[289]

Speech Analytics: The process of analysing live or recorded speech to understand and derive meaning in order to apply the information for various industries such as marketing and customer service.[290-292]

Virtual Agents (chatbots): Also known as '*conversational agents*'. These are software applications that mimic written or spoken human speech to simulate a conversation or interaction with a real person. It performs tasks that include rapid response, information delivery, and producing better service delivery. These are often seen in banking, call-centres, health, and improved customer service delivery.[293]

The above guidelines are based in a United States and global context, however numerous studies have been conducted in Africa under these subthemes (Table 2.3).

Table 2.3. Examples of studies that recently applied AI in public health in Africa.

Component	Examples of uses of AI in Public Health	Reference
Cognitive Search	Use of social media big data as a novel HIV surveillance tool in South Africa	van HeerdenYoung [294]
Computer Vision	Sensitivity and specificity of computer vision classification of eyelid photographs for programmatic trachoma assessment	Kim, Okada, Ryner, Amza, Tadesse, Cotter [295]
Deep Learning	Artificial intelligence (AI) and big data in cancer and precision oncology	Dlamini, Francies, HullMarima [296]
Machine Learning	Application of machine learning models in predicting length of stay among healthcare workers in underserved communities in South Africa	Moyo, Doan, YunTshuma [297]
Natural Language Processing (NLP)	Systematic review of the concept 'male involvement in maternal health' by natural language processing and descriptive analysis	Galle, Plaieser, Van Steenstraeten, Griffin, Osman, Roelens [298]
Robotic	Robotic health assistant (Feverkit) for the rational management of fevers among nomads in Nigeria	Akogun [299]
Speech Analytics	Nigerian innovators create Ubenwa, an app that detects asphyxia in babies	Masso, Chukwu, Calzati,[300]
Virtual Agents (Chatbots)	Virtual healthcare services and digital health technologies deployed during coronavirus disease 2019 (COVID-19) pandemic in South Africa: a systematic review	Mbunge, Batani, GaobotseMuchemwa [301]

In South Africa, an increase is AI research has and is expanding over numerous research fields outside of computer technology. Research fields including human language technologies robotics, and to an extent, in health informatics and biodiversity information management have begun to explore the use of AI in research.[302] Most recently, deep learning has been used in the prediction of maize production.[303] ML has also been used as a prediction tool in the healthcare and placements of healthcare workers.[297]

Although research concerning AI in public health in Africa is growing, AI applications in public health and medical studies are more prominent in countries such as China, USA, and Europe.[268,304] The use of AI in South African is increasing, with several studies showing the use and potential benefits of AI application in healthcare and medical studies,[294-296,301] including its role in the analysis of data on hearing impairment.[305] Additionally, ML has also been used as a prediction tool in healthcare and placements of healthcare workers.[297] The extent and implementation of AI and ML in epidemiology, in an African context, is in the early stages, which leaves room for exploration.[306] Post-COVID-19, there has been an increased interest in the use of AI to derive meaningful information from large amounts of health information that is being produced. [307] Thus, the escalated application and inclusion of AI and ML methods in health have put us into the 'peak of inflated expectations' phase defined by the 'Gartner Artificial Intelligence Hype Cycle'.[308] The inflated expectations are possibly unmatched expectations and favour of the initial capabilities of this technology.[309]

As useful as applications of AI in the public health and medical context are, there is limited knowledge on the preparedness of public health professionals to use these applications. Public health students and experts may need to be better versed in public health data science and learn how to apply these methods to draw distinctions between making inferences from statistical methods and prediction-oriented computational tools.[254] Additionally, using AI methods may assist public health researchers in processing the increasing availability and complexity of health data.[254] With the limited knowledge of AI and ML in health there are misconceptions and concerns about its use.[310-312] Few studies have been done to assess the attitudes and perceptions of those who are studying to work in public health, regarding the use of AI

in their work.[306,311,313] Other concerns about the use of AI include that it could potentially violate ethical values in health research, by creating bias in predictions and potentially making individual information available, reducing the anonymity of individuals.[314-315] Due to the misconceptions of AI in health, there may be a need to restructure health training to cover a basic understanding of AI concepts, limitations, and relevant implications it may have.[311,316-318]

The utilisation of ML can present a wide array of scalable and reliable methods and information.[268] The extent and implementation of AI and ML in epidemiology, in an African context, is in the early stages of exploration. However, the availability of expertise to execute this integration within the African context is limited.[306]

2.5.3. MACHINE LEARNING IN AIR POLLUTION STUDIES

There are a number of ways ML has been used in air pollution studies. A systematic review showed that ML has been used in source apportionment, prediction of air quality or exposure, geo-spatial coverage and generation of hypotheses.[268] Figure 2.8 shows ML in air pollution epidemiology research mainly occurring in North America, Asia, and Europe.[268]

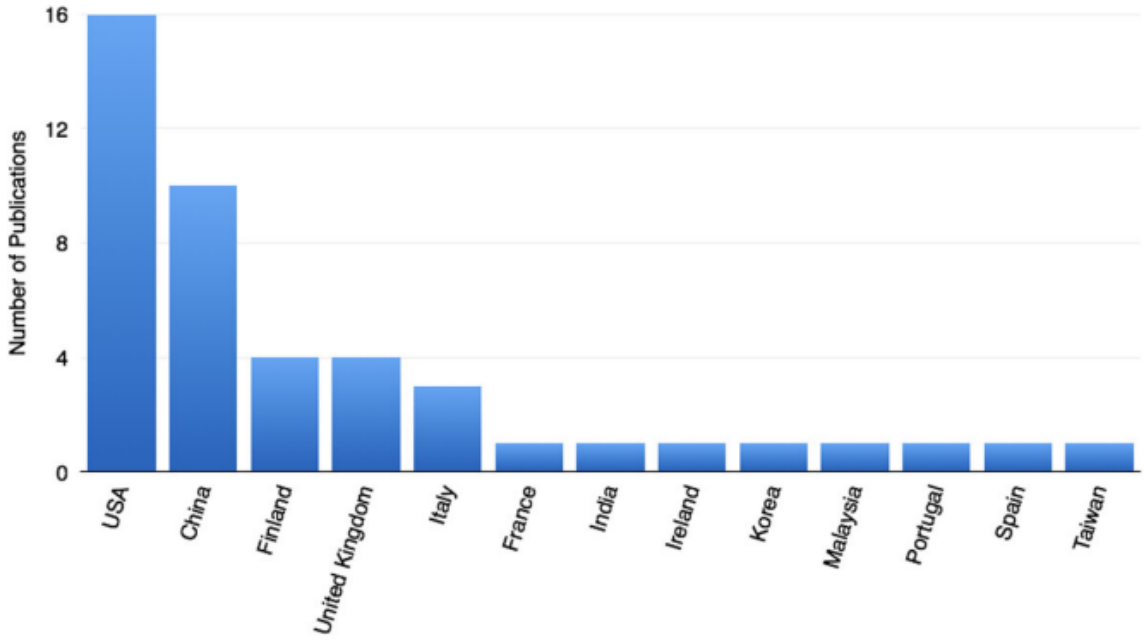


Figure 2.8: Machine learning in air pollution epidemiology studies per country between January 2000 and October 2017.[268]

Source apportionment focuses of grouping data according to the potential sources within a given study area.[62,319-323] Although there are traditional methods of source apportionment, ML methods are trying to be set as alternatives to these traditional methods. Unsupervised clustering ML is more commonly used for source apportionment, by grouping data according to familiarity among the data points as opposed to a general assumption from the researcher on how many groups to set prior.[324-328] The use PCA for source apportionment has also been explored however PCA was found to determine the correlations between pollutants as well as determine the source profiles of the various air pollutants of interest in a study.[218,329]

Predictive modelling has also been a more popular application of ML in air pollution epidemiology studies. There have been multiple ML methods used in air quality and air pollution forecasting and prediction, that also includes the use of support vector machines (SVM),[329-332] artificial neural networks (ANN), and random forest (RF), [330-331] Kohonen self-organising maps,[333] ensemble regression models,[334] and fuzzy evaluation.[332] Some methods have been used to estimate ambient air pollutant concentration levels for pollutants such as PM, CO, NO_2, SO_2, smog, and O_3,[330-331,334] and others have been used to determine air quality indexes (AQI).[329] South African based study had applied predictive modelling to determine healthcare worker allocation in health facilities.[297] Another study used predictive modelling for adverse health outcomes from air pollutants and complex data interactions which included external factors such as gender and race.,[333] Although the predictive modelling was able to produced resonable results the study the limitation of missing data reduced the effecincy of the predicted models.[333]

Moreover, ML methods are being used to monitor, inform, influence, measure, detect, forecast, advise, reduce, and manage air pollution in multiple cities around the world.[335] In addition, this analysis of air pollution is used to see the impact of the environment on population health. Many regions have insufficient air monitoring networks to provide real-time mapping.[336] With the leveraging of air pollution monitoring techniques, meteorological information, and land use information to map real-time pollution, health risk information can be better estimated.[336] The availability of air pollution data in Africa is sorely lacking and this affects the adequate attention needed for policy change and action across the continent.[3,337-338]

It would be important to explore more predictive modelling for air quality monitoring and early warning systems to evaluate the degree of air pollution objectively, which could assist in more accurately predicting pollutant concentrations and climate change.[339-340] Air quality monitoring and management are an important phase of air pollution epidemiological investigations, because the data obtained from quality monitoring and management influences the inferences made in studies. Information on air quality management and developments is available but still in need of better execution and implementation, despite the very evident possibility that it can impact air pollution epidemiology in South Africa.

Unsupervised ML methods such as clustering techniques have also been applied in conducting air of pollution studies.[341-343] Methods such as the Environmental Pollution Clustering (EPC) algorithm uses the k-means algorithm as the foundation algorithm to perform clustering, its main advantage being its ability to handle large data.[218] There are numerous unsupervised ML methods which include but are not limited to hierarchical clustering,[344-347] k-means clustering,[348] spectral clustering[326] and Density-Based Spatial Clustering for application with noise (DBSCAN).[349-350] Even with these new developments in clustering algorithms, an appropriate method for grouping the sample pollution data according to air pollution characteristics has not been established. [218]

K-means clustering is a simpler and common methods of clustering in numerous studies[326,351-352] This method uses an iterative process cluster centre means repeatedly till convergence has been reached. *K* refers to the number of clusters that are determined by the user/researcher that is randomly set. [268] Each data point then assigned itself to the closest mean centre. [353-357] Convergence is then considered to have occurred once the centres are stable and no longer moving. However, the method has some weaknesses, including its susceptibility to outliers.[268,326] Other clustering methods like DBSCAN clustering work on the assumption that highly dense data can be clustered from data that is of lower density.[350] Hierarchical clustering is a type of distance-based clustering that clusters data according to hierarchies formed by the algorithm.[347] Similar to hierarchal clustering in the balanced iterative reducing and clustering using hierarchies (BIRCH) clustering, which is an unsupervised data

mining algorithm that utilises hierarchical clustering over particularly large datasets.[358-359]

Another clustering method is spectral clustering, which is based on k-means clustering, but takes a slightly more advanced approach in its algorithm.[326] Spectral clustering is an algorithm based on a graphic theory.[326] It emphasises the use of either a normalised or unnormalized Laplacian matrix's eigenvalue decomposition to partition or cluster the data.[326,360] This method is thought to outperform some traditional clustering algorithms, such as the k-means clustering and expectation minimization (EM), which assume that the data has smooth spherical or elliptical distribution, whereas spectral clustering makes no such assumptions on the data and is better equipped to adapt to whichever shape the cluster forms.[361-362]

2.6. MISSING DATA

Missing data are a frequent problem encountered in multiple fields of research, more commonly experienced in environmental and occupational health research.[363-364] Although, missing data has also been an issue in longitudinal, clinical, and epidemiological studies.[365-368] In environmental studies, some common causes for missing data include malfunctioning equipment due to extreme weather conditions, as well as the maintenance, repair, and calibration of instruments.[369-372] Data collection at pollution stations located in remote areas have larger periods of missing data, due to faults with power supply.[369]

Health-based studies are often concerned with a daily average concentration associated with adverse health effects.[370] A critical component of exposure sciences and public health involves the monitoring of environmental pollutants.[363,370] Missing data can introduce bias in reporting results and reduce statistical power and precision in a study. In environmental health studies, missing data within a dataset can over- or underestimate the average concentration levels that a population may be exposed to.[371] One early set solution to the issue of missing data was to have the researcher state the magnitude of their missing data, and, more importantly, how the data was handled during statistical analysis.[363] However, this did not correct issues such as biased results, reduction of sample size, undermining of the validity of the study, and reducing the precision and power within the study.[363,373-374] Thus, the method of

imputation was introduced, which is the process of estimating and replacing missing values within a dataset.[378]

2.6.1. CLASSIFICATIONS OF MISSING DATA

To implement imputation methods, the 'mechanism' of missing data must be classified. These mechanisms are defined as Rubin's patterns of missing data.[365,376-378] There are three mechanisms of missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).[378] The MCAR mechanism assumes that the missing data are independent of both observed and unobserved data within a dataset.[379-380] It is the easier and most desirable of the three mechanisms to address.[376,378] The missing data under MCAR is usually under 5% of the dataset,[381] making it ideal to encounter in research. However, it is the least likely to encounter in most fields of research.[377,382]

The MAR mechanism of missing data assumes that the 'missingness' of the data could be dependent on the observed data.[376,378,381] Although the missing data observed within a variable is independent of the value itself; it is dependent on other variables that are included in the dataset.[379-380,381] The MAR assumption is considered preferable to work under.[380] There a many ways to address missing data under this assumption that can be either simple or computationally more complex.[363,369,372,385] Lastly, in cases where neither MCAR nor MAR can be assumed, the data are considered MNAR.[378,381]

The MNAR assumption is the more complicated mechanism of the three and more difficult to address.[382] Data that are MNAR occur when the probability of an observation being missing is related to unobserved values.[376,382] Meaning that the missing values are dependent on itself, and the possibility of predicting these values within the observed data are low.[380,386] There are various ways for dealing with missing data MNAR, which include, but are not limited to, methods similar to those used in handling selection bias.[386] However, these methods require multiple analyses of the missing data mechanism, because the MNAR situation is unique and requires specific approaches to deal with the missing data.[379] Nonetheless, the MNAR assumption can be improved by incorporating more variables in the dataset, which may reduce the severity of the MNAR conditions.[379,384]

In environmental and occupational health studies, the MAR assumption is the most used.[369-370,372,376,387] It is under this assumption that observed variables are used to estimate unobserved variables.[363,369] Buuren,[388] and Allison,[389] identify three general theoretical criteria for methods estimated for handling missing data. Firstly, statistical analysis conducted on the missing data method should ensure that the parameters of interest are close to the values which would have been obtained if the missing values had been observed.[388] The missing data method should minimize parameter bias in the analysis. Secondly, the imputation method should ensure that the majority of the available data are used, disregarding as little data as possible.[384,388] Lastly, the imputation method must produce good estimates of data variability, such as standard errors, p-values, and confidence intervals that are not over- or underestimated, to avoid lowering the efficiency in the analysis.[389]

2.6.2. METHODS TO ADDRESS MISSING DATA

There are simple methods of dealing with missing data in different categories of research, mainly the missing data in the dataset is simply deleted or disregarded, or the missing values within a dataset are replaced by means of imputation.[373,390-391] Imputation of missing values is the preferred method of handling missing data. This method replaces the missing values in a data set according to the assumed mechanism of missing-data. In majority of the imputation methods, the data are presumed to be under the mechanism of MAR. Therefore, the methods of imputation range from simple to more complex.[369-370]

2.6.2.1. UNIVARIATE/ UNIVARIATE TIME SERIES METHODS

The univariate methods are traditional methods of imputation and some are listed in Hadeed et al.[370] These methods include mean, median, random, and last observation carried forward (LOCF). These methods replace the missing data without taking into consideration the time-series nature of the data. R is one of many statistical programmes used to impute data. To perform these simple imputation methods, R packages such as imputeTS can be used. Within the imputeTS package, different imputation algorithm implementations can be performed, including multiple univariate and univariate time-series imputations.[392] Univariate imputation methods are classified as simpler and faster methods of imputation.[374] Mean and median imputations use the mean and median of the one variable to impute throughout the data set.[363,375,392]

Random imputation, however, replaces each missing value by drawing a random sample between two given bounds within the variable.[392] Lastly, LOCF replaces each missing value with the most recent present value before it.[392] Although these methods are simple and quick to conduct, they are highly biased assumptions and may affect the variance of the dataset.[363,369]

Univariate time-series imputations are more complex and consider the times-series characteristics of the partially observed data within the dataset.[370] While most imputation algorithms rely on inter-attribute correlations, univariate time-series imputation methods need to employ time-dependencies found within the partially observed data.[375] In addition, these methods use more advanced algorithms and need more computation time.[375] Kalman imputation, by the Kalman smoothing 'StructTS' option in the imputeTS package, uses a structural model fitted by maximum likelihood and, because KalmanRun is often considered extrapolation, KalmanSmooth is usually the better choice for imputation.[392] Kalman filters are used to fit autoregressive integrated moving average (ARIMA) models to predict missing values based on trends of previously observed measures.[392] This imputation method has been found to better perform under 20-40% missing data in a dataset.[368]

2.6.2.2. MULTIVARIATE METHODS

Multivariate time-series imputation methods are more complex and they use predictor variables between observations to impute the missing values; some methods use regression imputations, such as predictive mean matching (PMM).[370,394-395] Multiple imputation by chained imputation (mice) is a multivariate imputation method used under the assumption that missing data are MAR.[396] The mice package in R creates multiple imputations based on Fully Conditional Specification (FCS), where each incomplete variable is imputed by a separate model.[397] In the mice imputation the algorithm inspects the pattern of missing data, then imputes the data an 'm' number of times, then diagnoses the quality of the imputations before pooling the results of the repeated analysed data.[397] In addition, the mice algorithm can impute mixes of continuous, binary, unordered, and categorical data, so long as the correct regression algorithm is instated.[397] PMM regression is often used for numeric data, but can be used for other forms of data.[370] Other regression modelling can include, but are not

limited to, logistic regression, classification and regression trees, random forest imputations, and unconditional mean imputation.[396-397]

The multivariate time-series data imputation (mtsdi) method, uses an expected maximum likelihood (EM) algorithm-based method to impute missing values for multivariate time-series data.[371] This imputation algorithm takes into account the spatial and temporal correlation structures within the dataset.[371] As a default, a smooth spline is fitted on each time-series iteration in the imputation, and this method can be selected for univariate time-series data.[398] Another regression model, which under mtsdi imputation, is autoregressive integrated moving average (ARIMA). ARIMA is used to pattern temporality in the dataset with the option of seasonal components. However, the algorithm is tailored for missing climate data from several monitors in a given region.[398] Lastly, the mtsdi method can impute data using temporal patterns under the generalised additive model (gam) or generalised linear model (glm). The gam/glm model must be supplied to each covariate and then the covariates are used as part of the imputation model, however, this may result in collinearity among the variates.[394] Imputation using the default EM algorithm with normal multivariate data, under the fitting of a spline, has been found to yield precise estimates.[371]

2.7. REFERENCES

1. Mitchell G, Namdeo A, Kay D. A new disease-burden method for estimating the impact of outdoor air quality on human health. Science of the Total Environment. 2000; 246(2):153-63. doi:10.1016/S0048-9697(99)00455-6.

2. Europe WHO. Glossary on air pollution. Copenhagen WHO Regional Publications; 1980.

3. Abera A, Friberg J, Isaxon C, Jerrett M, Malmqvist E, Sjöström C, et al. Air quality in Africa: Public health implications. Annual Review of Public Health. 2021; 42(1). doi:10.1146/annurev-publhealth-100119-113802.

4. Amegah AK, Agyei-Mensah S. Urban air pollution in Sub-Saharan Africa: Time for action. Environmetal Pollution. 2017; 220(Part A):738-43. doi:10.1016/j.envpol.2016.09.042.

5. Brauer M, Amann M, Burnett RT, Cohen A, Dentener F, Ezzati M, et al. Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. Environmental Science & Technology. 2012; 46(2):652-60. doi:10.1021/es2025752.

6. World Health Organization. Burden of disease from ambient air pollution for 2012. Geneva. 2014.

7. Oji S, Adamu H. Correlation between air pollutants concentration and meteorological factors on seasonal air quality variation. Journal of Air Pollution and Health. 2020. doi:10.18502/japh.v5i1.2856.

8. Yang J, Shao M. Impacts of extreme air pollution meteorology on air quality in China. Journal of Geophysical Research: Atmospheres. 2021; 126(7):e2020JD033210. doi:https://doi.org/10.1029/2020JD033210.

9. Sager L. Estimating the effect of air pollution on road safety using atmospheric temperature inversions. Journal of Environmental Economics and Management. 2019; 98:102250. doi:https://doi.org/10.1016/j.jeem.2019.102250.

10. Morawska L, Afshari A, Bae GN, Buonanno G, Chao CY, Hänninen O, et al. Indoor aerosols: From personal exposure to risk assessment. Indoor Air. 2013; 23(6):462-87. doi:10.1111/ina.12044.

11. Sui X, Tian Z, Liu H, Chen H, Wang D. Field measurements on indoor air quality of a residential building in Xi'an under different ventilation modes in winter. Journal of Building Engineering. 2021; 42:103040.doi:10.1016/j.jobe.2021.103040.

12. Baxter LK, Burke J, Lunden M, Turpin BJ, Rich DQ, Thevenet-Morrison K, et al. Influence of human activity patterns, particle composition, and residential air exchange rates on modeled distributions of PM2.5 exposure compared with central-site monitoring data. Journal of Exposure Science and Environmental Epidemiology. 2013; 23(3):241-7. doi:10.1038/jes.2012.118.

13. Lei L, Chen W, Xue Y, Liu W. A comprehensive evaluation method for indoor air quality of buildings based on rough sets and a wavelet neural network. Building and Environment. 2019; 162:106296. doi:10.1016/j.buildenv.2019.106296.

14. Shaughnessy RJ, McDaniels T, Weschler CJ. Indoor chemistry: Ozone and volatile organic compounds found in tobacco smoke. Environmental Science & Technology. 2001; 35(13):2758-64.

15. Nazaroff WW. Four principles for achieving good indoor air quality. Indoor Air. 2013; 23(5):353-6. doi:10.1111/ina.12062.

16. Aung W, Noguchi M, Pan-Nu Yi E, Thant Z, Uchiyama S, Win-Shwe T, et al. Preliminary assessment of outdoor and indoor air quality in Yangon City, Myanmar. Atmospheric pollution research. 2019; 10(3):722-30. doi:10.1016/j.apr.2018.11.011.

17. Baldelli A. Evaluation of a low-cost multi-channel monitor for indoor air quality through a novel, low-cost, and reproducible platform. Measurement: Sensors. 2021; 17:100059. doi:10.1016/j.measen.2021.100059.

18. US Environmental Protection Agency. Introduction to indoor air quality. Available from: https://www.epa.gov/indoor-air-quality-iaq/introduction-indoor-air-quality.

19. Norbäck D, Lu C, Zhang Y, Li B, Zhao Z, Huang C, et al. Sources of indoor particulate matter (PM) and outdoor air pollution in China in relation to asthma, wheeze, rhinitis and eczema among pre-school children: Synergistic effects between antibiotics use and PM10 and second hand smoke. Environment International. 2019; 125:252-60. doi:10.1016/j.envint.2019.01.036.

20. Laumbach R, Kipen H. Respiratory health effects of air pollution: Update on biomass smoke and traffic pollution. Journal of Allergy and Clinical Immunology. 2012; 129(1):3-11. doi:10.1016/j.jaci.2011.11.021.

21. Mohamed S, Rodrigues L, Omer S, Calautit J. Overheating and indoor air quality in primary schools in the UK. Energy and Buildings. 2021; 250:111291. doi:10.1016/j.enbuild.2021.111291.

22. Charlesworth SM, Booth CA. Urban pollution : Science and management. Newark, United Kingdom: John Wiley & Sons, Incorporated; 2019.

23. Guarnieri M, Balmes JR. Outdoor air pollution and asthma. Lancet. 2014; 383(9928):1581-92. doi:10.1016/s0140-6736(14)60617-6.

24. Anderson JO, Thundiyil JG, Stolbach A. Clearing the air: A review of the effects of particulate matter air pollution on human health. Journal of Medical Toxicology. 2012; 8(2):166-75. doi:10.1007/s13181-011-0203-1.

25. Hime NJ, Marks GB, Cowie CT. A comparison of the health effects of ambient particulate matter air pollution from five emission sources. International Journal of Environmental Research and Public Health. 2018; 15(6):1-24. doi:10.3390/ijerph15061206.

26. Abera A, Friberg J, Isaxon C, Jerrett M, Malmqvist E, Sjöström C, et al. Air quality in Africa: Public health implications. Annual Review of Public Health. 2021; 42(1):193-210.

27. US Environmental Protection Agency. Criteria air pollutants. 2016. Available from: https://www.epa.gov/criteria-air-pollutants.

28. Barregard L, Molnàr P, Jonson JE, Stockfelt L. Impact on population health of baltic shipping emissions. International Journal of Environmental Research and Public Health. 2019; 16(11):1954 doi:10.3390/ijerph16111954.

29. Mwase NS, Ekström A, Jonson JE, Svensson E, Jalkanen JP, Wichmann J, et al. Health impact of air pollution from shipping in the baltic sea: Effects of different spatial resolutions in Sweden. International Journal of Environmental Health Research. 2020; 17(21):7963 doi:10.3390/ijerph17217963.

30. Tularam H, Ramsay LF, Muttoo S, Brunekreef B, Meliefste K, de Hoogh K, et al. A hybrid air pollution / land use regression model for predicting air pollution concentrations in Durban, South Africa. Environmental Pollution. 2021; 274:116513. doi:10.1016/j.envpol.2021.116513.

31. Department of Environment Forestry and Fisheries. Draft second generation air quality management plan for Vaal Triangle Airshed Priority Area. Pretoria,: Government Gazette; 2020.

32. Department of Environmental Affairs. Highveld Priority Area air quality management plan. Pretoria. 2012.

33. Department of Environmental Affairs. The Waterberg-Bojanala Priority Area air quality management plan: Baseline characterisation. 2014.

34. World Health Organization. Ambient (outdoor) air pollution in cities database 2014. 2018 Available from: http://www.who.int/phe/health_topics/outdoorair/databases/cities-2014/en/.

35. World Health Organization. WHO global air quality guidelines. Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. 2021 978-92-4-003422-8.

36. World Health Organization. WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide - global update 2005 - summary of risk assessment. Geneva; 2006.

37. Europe WHO. Update of WHO global air quality guidelines. Copenhagen. 2019. Available from: http://www.euro.who.int/en/health-topics/environment-and-health/air-quality/activities/update-of-who-global-air-quality-guidelines.

38. World Health Organization Europe. Review of evidence on health aspects of air pollution – REVIHAAP project: Final technical report. Copenhagen; 2013.

39. Pai SJ, Carter TS, Heald CL, Kroll JH. Updated world health organization air quality guidelines highlight the importance of non-anthropogenic PM2.5. Environmental science & technology letters. 2022; 9(6):501-6.

40. Brunt H, Jones SJ. A pragmatic public health-driven approach to enhance local air quality management risk assessment in wales, UK. Environmental Science & Policy. 2019; 96:18-26. doi:https://doi.org/10.1016/j.envsci.2019.02.008.

41. Gulia S, Khanna I, Shukla K, Khare M. Ambient air pollutant monitoring and analysis protocol for low and middle income countries: An element of comprehensive urban air quality management framework. Atmospheric Environment. 2020; 222. doi:10.1016/j.atmosenv.2019.117120.

42. Franco JF, Gidhagen L, Morales R, Behrentz E. Towards a better understanding of urban air quality management capabilities in Latin America. Environmental Science & Policy. 2019; 102:43-53. doi:10.1016/j.envsci.2019.09.011.

43. Schweizer D, Cisneros R, Traina S, Ghezzehei TA, Shaw G. Using national ambient air quality standards for fine particulate matter to assess regional wildland fire smoke and air quality management. Journal of environmental management. 2017; 201:345-56. doi:10.1016/j.jenvman.2017.07.004.

44. World Health Organization. Air pollution and health. 2021. Available from: https://www.who.int/teams/environment-climate-change-and-health/air-quality-and-health/policy-progress/sustainable-development-goals-air-pollution.

45. UN climate change conference UK 2021. COP26 goals. 2021. Available from: https://ukcop26.org/cop26-goals/.

46. Orru H, Ebi KL, Forsberg B. The interplay of climate change and air pollution on health. Current Environmental Health Reports. 2017; 4(4):504-13. doi:10.1007/s40572-017-0168-6.

47. Kinney PL. Interactions of climate change, air pollution, and human health. Current Environmental Health Reports. 2018; 5(1):179-86. doi:10.1007/s40572-018-0188-x.

48. Jacob DJ, Winner DA. Effect of climate change on air quality. Atmospheric Environment. 2009; 43(1):51-63.

49. Sanguineti PB, Lanzaco BL, López ML, Achad M, Palancar GG, Olcese LE, et al. PM2.5 monitoring during a 10-year period: Relation between elemental concentration and meteorological conditions. Environmental Monitoring and Assessment. 2020; 192(5):313. doi:10.1007/s10661-020-08288-0.

50. Fiore AM, Naik V, Leibensperger EM. Air quality and climate connections. Journal of the Air & Waste Management Association. 2015; 65(6):645-85. doi:10.1080/10962247.2015.1040526.

51. Kaudia A, Sokona Y, Mantlana B, Mbandi A, Osano P, Kehbila AG, et al. The launch of the first-ever integrated assessment of air pollution and climate change for

sustainable development in Africa. Clean Air Journal. 2022; 32(2) doi:10.17159/caj/2022/32/2.15320.

52. Chamseddine A, Alameddine I, Hatzopoulou M, El-Fadel M. Seasonal variation of air quality in hospitals with indoor-outdoor correlations. Building and Environment. 2019; 148:689-700. doi:10.1016/j.buildenv.2018.11.034.

53. Poole JA, Barnes CS, Demain JG, Bernstein JA, Padukudru MA, Sheehan WJ, et al. Impact of weather and climate change with indoor and outdoor air quality in asthma: A work group report of the AAAAI environmental exposure and respiratory health committee. Journal of Allergy and Clinical Immunology. 2019; 143(5):1702-10. doi:10.1016/j.jaci.2019.02.018.

54. Boffetta P, La Vecchia C, Moolgavkar S. Chronic effects of air pollution are probably overestimated. Risk Analysis. 2015; 35(5):766-9. doi:10.1111/risa.12320.

55. Hsu W, Hwang S, Kinney PL, Lin S. Seasonal and temperature modifications of the association between fine particulate air pollution and cardiovascular hospitalization in new york state. Science of the Total Environment. 2017; 578:626-36. .

56. Jacob DJ, Winner DA. Effect of climate change on air quality. Atmospheric Environment. 2009; 43(1):51-63.
.
57. Zhang R, Jing J, Tao J, Hsu S-C, Wang G, Cao J, et al. Chemical characterization and source apportionment of PM 2.5 in Beijing: Seasonal perspective. Atmospheric Chemistry and Physics. 2013; 13(14):7053-74.

58. Khare P, Baruah BP. Elemental characterization and source identification of PM2.5 using multivariate analysis at the suburban site of North-East India. Atmospheric Environment. 2010; 98(1):148-62. doi:10.1016/j.atmosres.2010.07.001.

59. Wang J, Ogawa S. Effects of meteorological conditions on PM2.5 concentrations in Nagasaki, Japan. International Journal of Environmental Research and Public Health. 2015; 12(8):9089-101. doi:10.3390/ijerph120809089.

60. Wang J, Wang Y, Liu H, Yang Y, Zhang X, Li Y, et al. Diagnostic identification of the impact of meteorological conditions on PM2.5 concentrations in Beijing. Atmospheric Environment. 2013; 81:158-65. doi:10.1016/j.atmosenv.2013.08.033.

61. Chen Y, Schleicher N, Fricker M, Cen K, Liu X, Kaminski U, et al. Long-term variation of black carbon and PM2.5 in Beijing, China with respect to meteorological conditions and governmental measures. Environmental Pollution. 2016; 212:269-78. doi:10.1016/j.envpol.2016.01.008.

62. Adeyemi A, Molnar P, Boman J, Wichmann J. Source apportionment of fine atmospheric particles using positive matrix factorization in Pretoria, South Africa. Environmental Monitoring and Assessment. 2021; 193(11):716. doi:10.1007/s10661-021-09483-3.

63. Howlett-Downing C, Boman J, Molnár P, Shirinde J, Wichmann J. PM2.5 chemical composition and geographical origin of air masses in Pretoria, South Africa. Water, Air, & Soil Pollution. 2022; 233(7) doi:10.1007/s11270-022-05746-y.

64. Feig G, Ncipha X, Vertue B, Naidoo S, Mabaso D, Ngcukana N, et al. Analysis of a period of elevated ozone concentration reported over the Vaal Triangle on 2 June 2013. Clean Air Journal= Tydskrif vir Skoon Lug. 2014; 24(1):10-6.

65. Feig GT, Vertue B, Naidoo S, Ncgukana N, Mabaso D. Measurement of atmospheric black carbon in the Vaal Triangle and Highveld priority areas. Clean Air Journal. 2015; 25(1):46-.

66. Govender K, Sivakumar V. A decadal analysis of particulate matter (PM2.5) and surface ozone (O3) over Vaal Priority Area, South Africa. Clean Air Journal. 2019; 29(2)

67. Muyemeki L, Burger R, Piketh SJ, Beukes JP, Van Zyl PG. Source apportionment of ambient PM10-25 and PM2.5 for the Vaal Triangle, South Africa. South African Journal of Science. 2021; 117(5-6):1-11.

68. Lelieveld J, Klingmüller K, Pozzer A, Burnett RT, Haines A, Ramanathan V. Effects of fossil fuel and total anthropogenic emission removal on public health and climate. Proceedings of the National Academy of Sciences. 2019; 116(15):7192-7. doi:10.1073/pnas.1819989116.

69. Lokotola C, Wichmann J, Wright C. Effect modification of temperature on air pollution associated with hospital admission for respiratory diseases in Cape Town, South Africa. Environmental Epidemiology. 2019; 3:249.

70. Lokotola CL, Wright CY, Wichmann J. Temperature as a modifier of the effects of air pollution on cardiovascular disease hospital admissions in Cape Town, South Africa. Environmental Science and Pollution Research. 2020; 27:16677-85. doi:10.1007/s11356-020-07938-7.

71. Mwase N, Olutola B, Wichmann J. Temperature modifies the association between air pollution and respiratory disease hospital admissions in an industrial area of South Africa: The Vaal Triangle Air Pollution Priority Area. Clean Air Journal. 2022; 32(2) doi:10.17159.caj.2022.32.2.14588.

72. Olutola B, Wichmann J. Does apparent temperature modify the effects of air pollution on respiratory disease hospital admissions in an industrial area of South Africa? Clean Air Journal. 2021; 31(2):1-11. doi:10.17159/caj/2021/31/2.11366.

73. Department of Health. National climate change & health adaptation plan 2014-2019. 2014.

74. Department of Environmental Affairs. National environmental management: Air quality act, 2004 (act no. 39 of 2004) national ambient air quality standards Government Gazette; 2009.

75. Naiker Y, Diab RD, Zunckel M, ET. H. Introduction of local air quality management in South Africa. Environ Science & Policy. 2012; 17:62-71.

76. Department of Environmental Affairs and Tourism. National environmental management: Air quality bill. 2003.

77. Garland RM, Wernecke B, Feig G, Langerman K. The new WHO global air quality guidelines: What do they mean for South Africa? Clean Air Journal. 2021; 31(2):1-2.

78. Howlett-Downing C. The association between sources of air pollution and respiratory health in Pretoria, South Africa: University of Pretoria; 2022.

79. Department of Environmental Affairs. Manual for air quality managemant planning - April 2012. 2012:12-6.

80. Tshehla C, Wright CY. 15 years after the national environmental management air quality act: Is legislation failing to reduce air pollution in South Africa? South African Journal of Science. 2019; 115(9-10):1-4.

81. Department of Environmental Affairs. Air quality management plan for the City of Tshwane metropolitan municipality 2006-2008 app/05/ctmm-02a. 2009.

82. C & M consulting engineers. City of Tshwane air quality monitoring network monthly activity report:November 2014. 2014.

83. Zunckel M, Raghunandan A, Moodley Y, Pillay B, Frank S, du Sart M. Review and compile the air quality management plan for the City of Tshwane – baseline assessment report-draft. 2019. uMN041-19.

84. Department of Environmental Affairs. National environmental management: Air quality act (act no. 39 of 2004) - Highveld Priority Area air quality management plan. 2011.

85. Albers PN, Mathee A, Voyi KVV, Wright CY. Household fuel use and child respiratory ill health in two towns in Mpumalanga, South Africa. SAMJ: South African Medical Journal. 2015; 105(7):573-7. doi:10.7196/SAMJNEW.7934.

86. Conradie EH, Pienaar JJ, Beukes JP, Van Zyl PG. Spatial and temporal deposition of selected biogeochemical important trace species in South Africa: North-West University . Potchefstroom Campus; 2018.

87. Deparment of Environmental Affairs and Tourism. Executive summary of The Vaal Triangle Airshed Proirity Area air quality managenet plan. 2008.

88. Olutola B. Influence of air pollution on the respiratory health of residents in the Vaal and Highveld Air Pollution Priority Areas of South Africa Pretoria: University of Pretoria. 2019.

89. Moreoane L. Evaluating the quality of air quality management plans (AQMP) in South Africa–the case of the Vaal Triangle: North-West University (South Africa); 2021.

90. Muyemeki L. An integrated systems approach towards air quality management in the Vaal Triangle Priority Area: North-West University (South Africa); 2021.

91. Wright CY, Oosthuizen R, John J, Garland RM, Albers P, Pauw C. Air quality and human health among a low income community in the Highveld Priority Area. Clean Air Journal: Tydskrif vir Skoon Lug. 2011; 20(1):12-20.

92. Feig GT, Naidoo S, Ncgukana N. Assessment of ambient air pollution in the Waterberg Priority Area 2012-2015. 2016.

93. Venter AD, Vakkar V, Beukes JP, van Zyl PG, Laakso H, Mabaso D, et al. An air quality assessment in the industrialised western Bushveld Igneous Complex, South Africa. South African Journal of Science. 108(9/10) doi:10.4102/sajs.v108i9/10.1059.

94. Department of Environmental Affairs. The Waterberg-Bojanala Priority Area air quality management plan: Baseline characterisation. 2014.

95. Deparment of Environmental Affairs. Waterberg-Bojanala Priority Area air quality management plan: Threat assessment. 2014.

96. South African Air Quality Information System. 2022. Available from: http://saaqis.environment.gov.za/.

97. Okello G, Nantanda R, Awokola B, Thondoo M, Okure D, Tatah L, et al. Air quality management strategies in Africa: A scoping review of the content, context, co-benefits and unintended consequences. Environment International. 2022:107709.

98. Spiru P, Simona PL. A review on interactions between energy performance of the buildings, outdoor air pollution and the indoor air quality. Energy Procedia. 2017; 128:179-86. doi:10.1016/j.egypro.2017.09.039.

99. Madaniyazi L, Xerxes S. Outdoor air pollution and the onset and exacerbation of asthma. Chronic Diseases and Translational Medicine. 2021; 7(2):100-6. doi:10.1016/j.cdtm.2021.04.003.

100. Si Q, Cardinal BJ. The health impact of air pollution and outdoor physical activity on children and adolescents in mainland China. The Journal of Pediatrics. 2017; 180:251-5. doi:10.1016/j.jpeds.2016.10.016.

101. Tofful L, Canepari S, Sargolini T, Perrino C. Indoor air quality in a domestic environment: Combined contribution of indoor and outdoor PM sources. Building and Environment. 2021; 202: 108050. doi:10.1016/j.buildenv.2021.108050.

102. Health Effects Institute. State of global air 2020 2020. Available from: www.stateofglobalair.org.

103. Rajak R, Chattopadhyay A. Short and long term exposure to ambient air pollution and impact on health in India: A systematic review. International Journal of Environmental Research and Public Health. 2020; 30(6):593-617. doi:10.1080/09603123.2019.1612042.

104. McDuffie E, Martin R, Yin H, Brauer M. Global burden of disease from major air pollution sources (GBD maps): A global approach. Research Reports: Health Effects Institute. 2021; 2021.

105. Ostro B, Spadaro JV, Gumy S, Mudu P, Awe Y, Forastiere F, et al. Assessing the recent estimates of the global burden of disease for ambient air pollution: Methodological changes and implications for low- and middle-income countries. Environmental Research. 2018; 166:713-25. doi:10.1016/j.envres.2018.03.001.

106. Murray CJ, Aravkin AY, Zheng P, Abbafati C, Abbas KM, Abbasi-Kangevari M, et al. Global burden of 87 risk factors in 204 countries and territories, 1990-2019: A systematic analysis for the global burden of disease study 2019. Lancet. 2020; 396(10258):1223-49. doi:10.1016/s0140-6736(20)30752-2.

107. Brauer M, Casadei B, Harrington RA, Kovacs R, Sliwa K. Taking a stand against air pollution - the impact on cardiovascular disease. European Heart Journal. 2021; 143(14):e800-4. doi:10.1093/eurheartj/ehaa1025.

108. Health Effects Institute. State of global air 2018. 2018. Available from: https://www.stateofglobalair.org/sites/default/files/soga-2018-report.pdf.

109. State of Global Air. Health impacts of household air pollution 2023. Available from: https://www.stateofglobalair.org/health/hap#regional-burden.

110. Beelen R, Hoek G, Raaschou-Nielsen O, Stafoggia M, Andersen Z, Weinmayr G, et al. Natural-cause mortality and long-term exposure to particle components: An analysis of 19 European cohorts within the multi-center ESCAPE project. Environmental Health Perspectives. 2015; 123(6):525-33. doi:10.1289/ehp.1408095.

111. Vodonos A, Awad YA, Schwartz J. The concentration-response between long-term PM2.5 exposure and mortality: a meta-regression approach. Environmental Research. 2018; 166:677-89. doi:10.1016/j.envres.2018.06.021

112. Hanigan IC, Rolfe MI, Knibbs LD, Salimi F, Cowie CT, Heyworth J, et al. All-cause mortality and long-term exposure to low level air pollution in the '45 and up study' cohort, sydney, australia, 2006-2015. Environment International. 2019; 126:762-70. doi:10.1016/j.envint.2019.02.044.

113. DeVries R, Kriebel D, Sama S. Outdoor air pollution and COPD-related emergency Department visits, hospital admissions, and mortality: A meta-analysis. International Journal of Chronic Obstructive Pulmonary Disease. 2017; 14(1):113-21. doi:10.1080/15412555.2016.1216956.

114. Health Effects Institute. The state of air quality and health impacts in Africa: A report from the state of global air initiative. 2022. Available from:

https://www.stateofglobalair.org/sites/default/files/documents/2022-10/soga-africa-report.pdf.

115. State of Global Air. Health impacts of PM2.5. 2023. Available from: https://www.stateofglobalair.org/health/pm.

116. State of Global Air. Health impacts of ozone 2023. Available from: https://www.stateofglobalair.org/health/ozone.

117. Heft-Neal S, Burney J, Bendavid E, Burke M. Robust relationship between air quality and infant mortality in Africa. Nature. 2018; 559(7713):254-8.

118. Osakede UA, Ajayi PI. Air pollution and health status in Sub-Sahara Africa (SSA). Journal of Economics and Sustainable Development. 2019; 10(22):2019.

119. Thabethe NDL, Voyi K, Wichmann J. Association between ambient air pollution and cause-specific mortality in Cape Town, Durban, and Johannesburg, South Africa: Any susceptible groups? Environmental Science and Pollution Research. 2021; 28:42868-76. doi:10.1007/s11356-021-13778-w.

120. Wichmann J. Heat effects of ambient apparent temperature on all-cause mortality in Cape Town, Durban and Johannesburg, South Africa: 2006-2010. Science of the Total Environment. 2017; 587-588:266-72. doi:10.1016/j.scitotenv.2017.02.135.

121. Wichmann J, Voyi K. Ambient air pollution exposure and respiratory, cardiovascular and cerebrovascular mortality in Cape Town, South Africa: 2001-2006. International Journal of Environmental Research and Public Health. 2012; 9(11):3978-4016. doi:10.3390/ijerph9113978.

122. Brugha R, Grigg J. Urban air pollution and respiratory infections. Paediatric Respiratory Reviews. 2014; 15(2):194-9. doi:10.1016/j.prrv.2014.03.001.

123. Cosselman KE, Navas-Acien A, Kaufman JD. Environmental factors in cardiovascular disease. Nature Reviews Cardiology. 2015; 12(11):627-42. doi:10.1038/nrcardio.2015.152.

124. Gent JF, Bell ML. Air pollution, population vulnerability, and standards for ambient air quality. American Journal of Respiratory and Critical Care Medicine. 2010; 182(3):296-7. doi:10.1164/rccm.201004-0538ED.

125. Lu F, Xu D, Cheng Y, Dong S, Guo C, Jiang X, et al. Systematic review and meta-analysis of the adverse health effects of ambient PM2.5 and PM10 pollution in the Chinese population. Environmental Research. 2015; 136:196-204. doi:10.1016/j.envres.2014.06.029.

126. Pope CA, Dockery DW. Health effects of fine particulate air pollution: Lines that connect. Journal of the Air & Waste Management Association. 2006; 56(6):709-42. doi:10.1080/10473289.2006.10464485.

127. Guarnieri M, Balmes JR. Outdoor air pollution and asthma. Lancet (London, England). 2014; 383(9928):1581-92. doi:10.1016/S0140-6736(14)60617-6.

128. Deng Q, Lu C, Norbäck D, Bornehag CG, Zhang Y, Liu W, et al. Early life exposure to ambient air pollution and childhood asthma in China. Environmental Research. 2015; 143(Pt A):83-92. doi:10.1016/j.envres.2015.09.032.

129. Ding L, Zhu D, Peng D, Zhao Y. Air pollution and asthma attacks in children: A case-crossover analysis in the city of Chongqing, China. Environmental Pollution. 2017; 220(Part A):348-53. doi:10.1016/j.envpol.2016.09.070.

130. Raaschou-Nielsen O, Andersen ZJ, Beelen R, Samoli E, Stafoggia M, Weinmayr G, et al. Air pollution and lung cancer incidence in 17 European cohorts: Prospective analyses from the European study of cohorts for air pollution effects (ESCAPE). The Lancet Oncology. 2013; 14(9):813-22. doi:10.1016/S1470-2045(13)70279-1.

131. Dima E, Kyriakoudi A, Kaponi M, Vasileiadis I, Stamou P, Koutsoukou A, et al. The lung microbiome dynamics between stability and exacerbation in chronic obstructive pulmonary disease (COPD): Current perspectives. Respiratory Medicine. 2019; 157:1-6.

132. Li X, Sun Y, An Y, Wang R, Lin H, Liu M, et al. Air pollution during the winter period and respiratory tract microbial imbalance in a healthy young population in northeastern China. Environmental Pollution. 2019; 246:972-9. doi:10.1016/j.envpol.2018.12.083.

133. Mariani J, Favero C, Carugno M, Pergoli L, Ferrari L, Bonzini M, et al. Nasal microbiota modifies the effects of particulate air pollution on plasma extracellular vesicles. International Journal of Environmental Health Research. 2020; 17(2):611.

134. Wang L, Cheng H, Wang D, Zhao B, Zhang J, Cheng L, et al. Airway microbiome is associated with respiratory functions and responses to ambient particulate matter exposure. Ecotoxicology and Environmental Safety. 2019; 167:269-77. doi:https://doi.org/10.1016/j.ecoenv.2018.09.079.

135. Xue Y, Chu J, Li Y, Kong X. The influence of air pollution on respiratory microbiome: A link to respiratory disease. Toxicology Letters. 2020; 334:14-20. doi:10.1016/j.toxlet.2020.09.007.

136. Santos UdP, Arbex MA, Braga ALF, Mizutani RF, Cançado JED, Terra-Filho M, et al. Environmental air pollution: Respiratory effects. Jornal Brasileiro de Pneumologia. 2021; 47(1) doi:10.36416/1806-3756/e20200267.

137. Yao Y, Chen X, Chen W, Wang Q, Fan Y, Han Y, et al. Susceptibility of individuals with chronic obstructive pulmonary disease to respiratory inflammation associated with short-term exposure to ambient air pollution: A panel study in Beijing. Science of the Total Environment. 2021; 766: 142639. doi:10.1016/j.scitotenv.2020.142639.

138. Capraz O, Deniz A, Dogan N. Effects of air pollution on respiratory hospital admissions in Istanbul, Turkey, 2013 to 2015. Chemosphere. 2017; 181:544-50. doi:10.1016/j.chemosphere.2017.04.105.

139. Chen K, Glonek G, Hansen A, Williams S, Tuke J, Salter A, et al. The effects of air pollution on asthma hospital admissions in Adelaide, South Australia, 2003-2013: Time-series and case-crossover analyses. Clinical & Experimental Allergy. 2016; 46(11):1416-30. doi:10.1111/cea.12795.

140. Nhung NTT, Schindler C, Dien TM, Probst-Hensch N, Perez L, Künzli N. Acute effects of ambient air pollution on lower respiratory infections in Hanoi children: An eight-year time series study. Environment international. 2018; 110:139-48. doi:10.1016/j.envint.2017.10.024.

141. Howlett-Downing C, Boman J, Molnár P, Shirinde J, Wichmann J. Case-crossover study for the association between increased hospital admissions for respiratory diseases and the increase in atmospheric PM2.5 and PM2.5-bound trace elements in Pretoria, South Africa. International Journal of Environmental Health Research. 2023:1-15

142. Nhung NTT, Schindler C, Dien TM, Probst-Hensch N, Künzli N. Association of ambient air pollution with lengths of hospital stay for Hanoi children with acute lower-respiratory infection, 2007-2016. Environmental Pollution. 2019; 247:752-62. doi:10.1016/j.envpol.2019.01.115.

143. Carugno M, Consonni D, Randi G, Catelan D, Grisotto L, Bertazzi PA, et al. Air pollution exposure, cause-specific deaths and hospitalizations in a highly polluted Italian region. Environmental Research. 2016; 147:415-24. doi:10.1016/j.envres.2016.03.003.

144. Pannullo F, Lee D, Neal L, Dalvi M, Agnew P, O'Connor FM, et al. Quantifying the impact of current and future concentrations of air pollutants on respiratory disease risk in England. Environmental Health. 2017; 16(1):29. doi:10.1186/s12940-017-0237-1.

145. Terblanche AP, Opperman L, Nel CM, Reinach SG, Tosen G, Cadman A. Preliminary results of exposure measurements and health effects of the Vaal Triangle air pollution health study. South African Medical Journal. 1992; 81(11):550-6.

146. Liu Y, Pan J, Zhang H, Shi C, Li G, Peng Z, et al. Short-term exposure to ambient air pollution and asthma mortality. American Journal of Respiratory and Critical Care Medicine. 2019; 200(1):24-32. doi:10.1164/rccm.201810-1823OC.

147. Zhou M, He G, Liu Y, Yin P, Li Y, Kan H, et al. The associations between ambient air pollution and adult respiratory mortality in 32 major Chinese cities, 2006-2010. Environmental Research. 2015; 137:278-86. doi:10.1016/j.envres.2014.12.016.

148. Thabethe N, Wichmann J, Voyi K. The association between the daily number of deaths due to respiratory, cardiovascular and cerebrovascular diseases and ambient

air pollution levels in Cape Town, Durban and Johannesburg during 1 January 2006 to 31 December 2010 University of Pretoria; 2017.

149. Zheng X-Y, Orellano P, Lin H-L, Jiang M, Guan W-J. Short-term exposure to ozone, nitrogen dioxide, and sulphur dioxide and emergency department visits and hospital admissions due to asthma: A systematic review and meta-analysis. Environment International. 2021; 150:106435. doi:10.1016/j.envint.2021.106435.

150. Lopez AD, Adair T. Is the long-term decline in cardiovascular-disease mortality in high-income countries over? Evidence from national vital statistics. International Journal of Epidemiology. 2019; 48(6):1815-23.

151. Baddour LM, Addolorato G, Ammirati E, Mensah G, Johnson C, Rwegerera GM, et al. Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study; 2020.

152. Luo C, Zhu X, Yao C, Hou L, Zhang J, Cao J, et al. Short-term exposure to particulate air pollution and risk of myocardial infarction: A systematic review and meta-analysis. Environmental Science and Pollution Research. 2015; 22(19):14651-62. doi:10.1007/s11356-015-5188-x.

153. Shah AS, Lee KK, McAllister DA, Hunter A, Nair H, Whiteley W, et al. Short term exposure to air pollution and stroke: Systematic review and meta-analysis. BMJ (Clinical research ed.). 2015; 350:h1295. doi:10.1136/bmj.h1295.

154. Vidale S, Arnaboldi M, Bosio V, Corrado G, Guidotti M, Sterzi R, et al. Short-term air pollution exposure and cardiovascular events: A 10-year study in the urban area of Como, italy. International journal of cardiology. 2017; 248:389-93. doi:10.1016/j.ijcard.2017.06.037.

155. Abdolahnejad A, Jafari N, Mohammadi A, Miri M, Hajizadeh Y, Nikoonahad A. Cardiovascular, respiratory, and total mortality ascribed to PM10 and PM2.5 exposure in Isfahan, Iran. Journal of Education and Health Promotion. 2017; 6:109. doi:10.4103/jehp.jehp_166_16.

156. Kaufman JD, Adar SD, Barr RG, Budoff M, Burke GL, Curl CL, et al. Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the multi-ethnic study of atherosclerosis and air pollution): A longitudinal cohort study. Lancet. 2016; 388(10045):696-704. doi:10.1016/s0140-6736(16)00378-0.

157. Robertson S, Miller MR. Ambient air pollution and thrombosis. Particle and Fibre Toxicology. 2018; 15(1):1-16. doi:10.1186/s12989-017-0237-x.

158. Zhang Z, Guo C, Lau AKH, Chan TC, Chuang YC, Lin C, et al. Long-term exposure to fine particulate matter, blood pressure, and incident hypertension in Taiwanese adults. Environmental Health Perspectives. 2018; 126(1):017008. doi:10.1289/ehp2466.

159. Nhung NTT, Schindler C, Chau NQ, Hanh PT, Hoang LT, Dien TM, et al. Exposure to air pollution and risk of hospitalization for cardiovascular diseases amongst Vietnamese adults: Case-crossover study. Science of the Total Environment. 2020; 703:134637. doi:10.1016/j.scitotenv.2019.13463.7

160. Faustini A, Rapp R, Forastiere F. Nitrogen dioxide and mortality: Review and meta-analysis of long-term studies. European Respiratory Journal. 2014; 44:744-53.

161. Mordukhovich I, Coull B, Kloog I, Koutrakis P, Vokonas P, Schwartz J. Exposure to sub-chronic and long-tern particulate air pollution and heart rate variability in an elderly cohort: The normative aging study. Environmental Health 2015; 14:87.

162. Shah A, Langrish J, Nair H, al e. Global association of air pollution and heart failure: A systematic review and meta-analysis. Lancet. 2013; 382:1039-48.

163. Henrotin J, Zeller M, Lorgis L, Cottin Y, Giroud M, Bejot Y. Evidence of the role of short-term exposure to ozone on ischaemic cerebral and cardiac events: The Dijon Vascular Project(DIVA). Heart. 2010; 96:1990-6.

164. Ljungman PL, Mittleman MA. Ambient air pollution and stroke. Stroke. 2014; 45:3734-41.

165. Shin H, Burnett R, Cohen A, Hubbell BJ. Outdoor fine particles and nonfatal strokes: Systematic review and meta-analysis. Epidemiology. 2014; 25:835-42.

166. Dastoorpoor M, Sekhavatpour Z, Masoumi K, Mohammadi MJ, Aghababaeian H, Khanjani N, et al. Air pollution and hospital admissions for cardiovascular diseases in Ahvaz, Iran. Science of the Total Environment. 2019; 652:1318-30. doi:10.1016/j.scitotenv.2018.10.285.

167. Tibuakuu M, Michos ED, Navas-Acien A, Jones MR. Air pollution and cardiovascular disease: A focus on vulnerable populations worldwide. Current Epidemiology Reports. 2018; 5(4):370-8. doi:10.1007/s40471-018-0166-8.

168. Yitshak-Sade M, Nethery R, Abu Awad Y, Mealli F, Dominici F, Kloog I, et al. Lowering air pollution levels in massachusetts may prevent cardiovascular hospital admissions. Journal of the American College of Cardiology. 2020; 75(20):2642-4. doi:10.1016/j.jacc.2020.03.056.

169. Shah ASV, Langrish JP, Hunter AL, Donaldson K, Newby DE, Mills NL, et al. Global association of air pollution and heart failure: A systematic review and meta-analysis. The Lancet. 2013; 382(9897):1039-48. doi:10.1016/S0140-6736(13)60898-3.

170. Cesaroni G, Forastiere F, Stafoggia M, Andersen ZJ, Badaloni C, Beelen R, et al. Long term exposure to ambient air pollution and incidence of acute coronary events: Prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE project. BMJ. 2014; 348(jan21 3):f7412. doi:10.1136/bmj.f7412.

171. Goldberg MS, Burnett RT, Stieb DM, Brophy JM, Daskalopoulou SS, Valois M-F, et al. Associations between ambient air pollution and daily mortality among elderly persons in Montreal, Quebec. The Science of the Total Environment. 2013; 463-464:931-42. doi:10.1016/j.scitotenv.2013.06.095.

172. Pun VC, Kazemiparkouhi F, Manjourides J, Suh HH. Long-term pm2. 5 exposure and respiratory, cancer, and cardiovascular mortality in older US adults. American Journal of Epidemiology. 2017; 186(8):961-9.

173. Hart JE, Puett RC, Rexrode KM, Albert CM, Laden F. Effect modification of long-term air pollution exposures and the risk of incident cardiovascular disease in US women. Journal of the American Heart Association. 2015; 4(12).

174. Stockfelt L, Andersson EM, Molnár P, Gidhagen L, Segersson D, Rosengren A, et al. Long-term effects of total and source-specific particulate air pollution on incident cardiovascular disease in Gothenburg, Sweden. Environmental Research. 2017; 158:61-71. doi:10.1016/j.envres.2017.05.036.

175. Humphries KH, Izadnegahdar M, Sedlak T, Saw J, Johnston N, Schenck-Gustafsson K, et al. Sex differences in cardiovascular disease - impact on care and outcomes. Frontiers in Neuroendocrinology. 2017; 46:46-70. doi:10.1016/j.yfrne.2017.04.001.

176. Bots SH, Peters SAE, Woodward M. Sex differences in coronary heart disease and stroke mortality: A global assessment of the effect of ageing between 1980 and 2010. BMJ Global Health. 2017; 2(2):e000298. doi:10.1136/bmjgh-2017-000298.

177. Hystad P, Demers PA, Johnson KC, Carpiano RM, Brauer M. Long-term residential exposure to air pollution and lung cancer risk. Epidemiology. 2013; 24(5):762-72.

178. Vallero DA Fundamentals of air pollution. Waltham, MA: Elsevier Science; 2014.

179. Morakinyo O, Mokgobu M, Mukhola M, Hunter R. Health risk assessment of airborne pollutants in fine particulate matter in an industrial area in Pretoria West, South Africa: Tshwane University of Technology; 2018.

180. Myers R, Brauer M, Ladhar S, Atkar-Khattra S, Yee J, Ho C, et al. Oa09.07 association between outdoor air pollution and lung cancer in female never smokers. Journal of Thoracic Oncology. 2018; 13(10):S342. doi:10.1016/j.jtho.2018.08.286.

181. Xing DF, Xu CD, Liao XY, Xing TY, Cheng SP, Hu MG, et al. Spatial association between outdoor air pollution and lung cancer incidence in China. BMC Public Health. 2019; 19 doi:10.1186/s12889-019-7740-y.

182. Castaño-Vinyals G, Cantor KP, Malats N, Tardon A, Garcia-Closas R, Serra C, et al. Air pollution and risk of urinary bladder cancer in a case-control study in Spain. Occupational and Environmental Medicine. 2008; 65(1):56-60.

183. Coleman NC, Burnett RT, Higbee JD, Lefler JS, Merrill RM, Ezzati M, et al. Cancer mortality risk, fine particulate air pollution, and smoking in a large, representative cohort of US adults. Cancer Causes & Control. 2020; 31(8):767-76.

184. Pedersen M, Stafoggia M, Weinmayr G, Andersen ZJ, Galassi C, Sommar J, et al. Is there an association between ambient air pollution and bladder cancer incidence? Analysis of 15 European cohorts. European urology focus. 2018; 4(1):113-20.

185. Turner MC, Andersen ZJ, Baccarelli A, Diver WR, Gapstur SM, Pope CA, et al. Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. CA: A Cancer Journal for Clinicians. 2020; 70(6):460-79. doi:10.3322/caac.21632.

186. White AJ, Bradshaw PT, Hamra GB. Air pollution and breast cancer: A review. Current Epidemiology Reports. 2018; 5(2):92-100.

187. White AJ, Keller JP, Zhao S, Carroll R, Kaufman JD, Sandler DP. Air pollution, clustering of particulate matter components, and breast cancer in the sister study: A US-wide cohort. Environmental Health Perspectives. 2019; 127(10):107002.

188. Andersen ZJ, Raaschou-Nielsen O, Ketzel M, Jensen SS, Hvidberg M, Loft S, et al. Diabetes incidence and long-term exposure to air pollution: A cohort study. Diabetes Care. 2012; 35(1):92-8. doi:10.2337/dc11-1155.

189. Eze IC, Imboden M, Kumar A, von Eckardstein A, Stolz D, Gerbase MW, et al. Air pollution and diabetes association: Modification by type 2 diabetes genetic risk score. Environment International. 2016; 94:263-71. doi:10.1016/j.envint.2016.04.032.

190. Kawada T. Air pollution and diabetes mellitus. Diabetes Research and Clinical Practice. 2015; 108(1):e7. doi:10.1016/j.diabres.2015.01.002.

191. Thiering E, Heinrich J. Epidemiology of air pollution and diabetes. Trends in Endocrinology and Metabolism. 2015; 26(7):384-94. doi:10.1016/j.tem.2015.05.002.

192. Yang B-Y, Fan S, Thiering E, Seissler J, Nowak D, Dong G-H, et al. Ambient air pollution and diabetes: A systematic review and meta-analysis. Environmental Research. 2020; 180:108817. doi:10.1016/j.envres.2019.108817.

193. Hu H, Ha S, Henderson BH, Warner TD, Roth J, Kan H, et al. Association of atmospheric particulate matter and ozone with gestational diabetes mellitus. Environmental Health Perspectives. 2015; 123(9):853-9. doi:10.1289/ehp.1408456.

194. Lu MC, Wang P, Cheng TJ, Yang CP, Yan YH. Association of temporal distribution of fine particulate matter with glucose homeostasis during pregnancy in women of Chiayi City, Taiwan. Environmental Research. 2017; 152:81-7. doi:10.1016/j.envres.2016.09.023.

195. Robledo CA, Mendola P, Yeung E, Mannisto T, Sundaram R, Liu D, et al. Preconception and early pregnancy air pollution exposures and risk of gestational

diabetes mellitus. Environmental Research. 2015; 137:316-22. doi:10.1016/j.envres.2014.12.020.

196. Goyal N, Canning D. Exposure to ambient fine particulate air pollution in utero as a risk factor for child stunting in Bangladesh. International Journal of Environmental Research and Public Health. 2017; 15(1) doi:10.3390/ijerph15010022.

197. Lee PC, Roberts JM, Catov JM, Talbott EO, Ritz B. First trimester exposure to ambient air pollution, pregnancy complications and adverse birth outcomes in Allegheny County, PA. Maternal and Child Health Journal. 2013; 17(3):545-55. doi:10.1007/s10995-012-1028-5.

198. Vinikoor-Imler LC, Davis JA, Meyer RE, Messer LC, Luben TJ. Associations between prenatal exposure to air pollution, small for gestational age, and term low birthweight in a state-wide birth cohort. Environmental Research. 2014; 132:132-9. doi:10.1016/j.envres.2014.03.040.

199. Zhang Q, Sun S, Sui X, Ding L, Yang M, Li C, et al. Associations between weekly air pollution exposure and congenital heart disease. Science of the Total Environment. 2021; 757:143821. doi:10.1016/j.scitotenv.2020.143821.

200. Wei F, Wu M, Qian S, Li D, Jin M, Wang J, et al. Association between short-term exposure to ambient air pollution and hospital visits for depression in China. Science of the Total Environment. 2020; 724 doi:10.1016/j.scitotenv.2020.138207.

201. Braithwaite I, Zhang S, Kirkbride JB, Osborn DPJ, Hayes JF. Air pollution (particulate matter) exposure and associations with depression, anxiety, bipolar, psychosis and suicide risk: A systematic review and meta-analysis. Environmental Health Perspectives. 2019; 127(12):126002. doi:10.1289/ehp4595.

202. Comunian S, Dongo D, Milani C, Palestini P. Air pollution and COVID -19: The role of particulate matter in the spread and increase of COVID -19's morbidity and mortality. International Journal of Environmental Research and Public Health. 2020; 17(12) doi:10.3390/ijerph17124487.

203. Wang B, Chen H, Chan YL, Oliver BG. Is there an association between the level of ambient air pollution and COVID-19? American Journal of Physiology-Lung Cellular and Molecular Physiology. 2020; 319(3):L416-l21. doi:10.1152/ajplung.00244.2020.

204. Copat C, Cristaldi A, Fiore M, Grasso A, Zuccarello P, Signorelli SS, et al. The role of air pollution (PM and NO2) in COVID-19 spread and lethality: A systematic review. Environmental Research. 2020; 191:110129. doi:10.1016/j.envres.2020.110129.

205. Urrutia-Pereira M, Mello-da-Silva CA, Solé D. COVID-19 and air pollution: A dangerous association? Allergologia et immunopathologia. 2020; 48(5):496-9. doi:10.1016/j.aller.2020.05.004.

206. Sokhi RS, Singh V, Querol X, Finardi S, Targino AC, de Fatima Andrade M, et al. A global observational analysis to understand changes in air quality during

exceptionally low anthropogenic emission conditions. Environment international. 2021; 157:106818.

207. Lee H, Myung W, Cheong HK, Yi SM, Hong YC, Cho SI, et al. Ambient air pollution exposure and risk of migraine: Synergistic effect with high temperature. Environment International. 2018; 121(Pt 1):383-91. doi:10.1016/j.envint.2018.09.022.

208. Valavanidis A, Loridas S, Vlahogianni T, Fiotakis K. Influence of ozone on traffic-related particulate matter on the generation of hydroxyl radicals through a heterogeneous synergistic effect. Journal of Hazardous Materials. 2009; 162(2):886-92. doi:10.1016/j.jhazmat.2008.05.124.

209. Bergmann S, Li B, Pilot E, Chen R, Wang B, Yang J. Effect modification of the short-term effects of air pollution on morbidity by season: A systematic review and meta-analysis. Science of the Total Environment. 2020; 716:136985. doi:10.1016/j.scitotenv.2020.136985.

210. Rodríguez-Villamizar LA, Rojas-Roa NY, Fernández-Niño JA. Short-term joint effects of ambient air pollutants on emergency department visits for respiratory and circulatory diseases in Colombia, 2011-2014. Environmental Pollution. 2019; 248:380-7. doi:10.1016/j.envpol.2019.02.028.

211. Urman R, McConnell R, Islam T, Avol EL, Lurmann FW, Vora H, et al. Associations of children's lung function with ambient air pollution: Joint effects of regional and near-roadway pollutants. Thorax. 2014; 69(6):540-7. doi:10.1136/thoraxjnl-2012-203159.

212. Stafoggia M, Breitner S, Hampel R, Basagaña X. Statistical approaches to address multi-pollutant mixtures and multiple exposures: The state of the science. Current Environmental Health Reports. 2017; 4(4):481-90. doi:10.1007/s40572-017-0162-z.

213. Seinfeld JH, Pandis SN. Atmospheric chemistry and physics : From air pollution to climate change. Third edition. ed. Hoboken, New Jersey: John Wiley & Sons; 2016.

214. Davalos AD, Luben TJ, Herring AH, Sacks JD. Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. Annals of Epidemiology. 2017; 27(2):145-53.e1. doi:10.1016/j.annepidem.2016.11.016.

215. Billionnet C, Sherrill D, Annesi-Maesano I. Estimating the health effects of exposure to multi-pollutant mixture. Annals of Epidemiology. 2012; 22(2):126-41. doi:10.1016/j.annepidem.2011.11.004.

216. Gass K, Klein M, Chang HH, Flanders WD, Strickland MJ. Classification and regression trees for epidemiologic research: An air pollution example. Environmental Health. 2014; 13(1):1-10. doi:10.1186/1476-069X-13-17.

217. Gass K, Klein M, Sarnat SE, Winquist A, Darrow LA, Flanders WD, et al. Associations between ambient air pollutant mixtures and pediatric asthma emergency

department visits in three cities: A classification and regression tree approach. Environmental Health. 2015; 14:58. doi:10.1186/s12940-015-0044-5.

218. Chen M, Wang P, Chen Q, Wu J, Chen X. A clustering algorithm for sample data based on environmental pollution characteristics. Atmospheric Environment. 2015; 107:194-203. doi:10.1016/j.atmosenv.2015.02.042.

219. Hopke PK, Dai Q, Li L, Feng Y. Global review of recent source apportionments for airborne particulate matter. Science of The Total Environment. 2020; 740:140091.

220. Usemann J, Decrue F, Korten I, Proietti E, Gorlanova O, Vienneau D, et al. Exposure to moderate air pollution and associations with lung function at school-age: A birth cohort study. Environment international. 2019; 126:682-9. doi:10.1016/j.envint.2018.12.019.

221. Shin HH, Parajuli RP, Gogna P, Maquiling A, Dehghani P. Pollutant-sex specific differences in respiratory hospitalization and mortality risk attributable to short-term exposure to ambient air pollution. Science of the Total Environment. 2021; 755(Pt 2):143135. doi:10.1016/j.scitotenv.2020.143135.

222. Molnár P, Sallsten G. Contribution to PM(2.5) from domestic wood burning in a small community in Sweden. Environmental Science: Processes & Impacts. 2013; 15(4):833-8. doi:10.1039/c3em30864b.

223. Molnar P, Tang L, Sjoberg K, Wichmann J. Long-range transport clusters and positive matrix factorization source apportionment for investigating transboundary PM2.5 in Gothenburg, Sweden. Environmental Science: Processes & Impacts. 2017; 19(10):1270-7. doi:10.1039/c7em00122c.

224. Belis C, Bo L, Amato F, Haddad I, et.al. European guide on air pollution source apportionment with receptor models. Luxembourg: European Commission, Joint Research Centre, Institute for Environment and Sustainability, 2014.

225. Murillo j, Roman S, Marín J, Cardenas B. Source apportionment of PM2.5 in the metropolitan area of Costa Rica using receptor models. Atmospheric and Climate Sciences. 2015; 3(4):562-75. doi:10.4236/acs.2013.34059.

226. Hopke PK. Review of receptor modeling methods for source apportionment. Journal of the Air & Waste Management Association. 2016; 66(3):237-59. doi:10.1080/10962247.2016.1140693.

227. Mathuthu M, Dudu VP, Manjoro M. Source apportionment of air particulates in South Africa: A review. Atmospheric and Climate Sciences. 2018; 9(1):100-13.

228. U.S. Environmental Protection Agency. Receptor modeling, positive matrix factorization. 2014. Available from: http://www.epa.gov/heasd/research/pmf.html.

229. Watson JG, Chow JC, Fujita EM. Review of volatile organic compound source apportionment by chemical mass balance. Atmospheric Environment. 2001; 35(9):1567-84. doi:10.1016/S1352-2310(00)00461-1.

230. Thurston GD, Spengler JD. A quantitative assessment of source contributions to inhalable particulate matter pollution in metropolitan Boston. Atmospheric Environment. 1985; 19(1):9-25. doi:10.1016/0004-6981(85)90132-5.

231. Henry RC. Unmix version 2 manual Prepared for the US Environmental Protection Agency; 2000.

232. Tiwari S, Hopke PK, Thimmaiah D, Dumka UC, Srivastava AK, Bisht DS, et al. Nature and sources of ionic species in precipitation across the Indo-Gangetic Plains, India. Aerosol and Air Quality Research. 2016; 16(4):943-57. doi:10.4209/aaqr.2015.06.0423.

233. Ulbrich IM, Canagaratna MR, Zhang Q, Worsnop DR, Jimenez JL. Interpretation of organic components from positive matrix factorization of aerosol mass spectrometric data. Atmospheric Chemistry and Physics. 2009; 9:2891-918.

234. VanCuren RT, Gustin MS. Identification of sources contributing to PM2.5 and ozone at elevated sites in the western U.S. by receptor analysis: Lassen Volcanic National Park, California, and Great Basin National Park, Nevada. Science of the Total Environment. 2015; 530-531:505-18. doi:10.1016/j.scitotenv.2015.03.091.

235. Shang J, Zheng Y, Tong W, Chang E, Yu Y. Inferring gas consumption and pollution emission of vehicles throughout a city. Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. 2014: 1027-36.

236. Nicola MW, Stuart JP, Pieter van Z, Willy M, Roelof B, Brigitte L, et al. Source apportionment of ambient fine and coarse aerosols in Embalenhle and Kinross, South Africa. Clean Air Journal. 31(2):1-13.

237. Novela RJ, Gitari WM, Chikoore H, Molnar P, Mudzielwana R, Wichmann J. Chemical characterization of fine particulate matter, source apportionment and long-range transport clusters in Thohoyandou, South Africa. Clean Air J.; 30(2):1-2 doi:10.17159/caj/2020/30/2.8735.

238. Williams J, Petrik L, Wichmann J. PM2.5 chemical composition and geographical origin of air masses in Cape Town, South Africa. Air Quality, Atmosphere & Health. 2020; 14(3):431-42. doi:10.1007/s11869-020-00947-y.

239. Zhou Z, Dionisio KL, Verissimo TG, Kerr AS, Coull B, Howie S, et al. Chemical characterization and source apportionment of household fine particulate matter in rural, peri-urban, and urban West Africa. Environmental science & technology. 2014; 48(2):1343-51. doi:10.1021/es404185m.

240. Madikizela LM, Chimuka L, Ncube S. Metal pollution source apportionment in two important rivers of Eastern Cape province, South Africa: A case study of Bizana and Mthatha rivers. Environmental Forensics. 2021:1-14. doi:10.1080/15275922.2021.1940382.

241. Witten IH, Frank E, Hall MA, Pal CJ. Data mining : Practical machine learning tools and techniques. 4th ed. Saint Louis: Elsevier Science; 2016.

242. Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, et al. The role of artificial intelligence in achieving the sustainable development goals. Nature Communications. 2020; 11(1):233. doi:10.1038/s41467-019-14108-y.

243. Samuel A. Studies in machine learning using the game of checkers. IBM Journal of Research and Development. 1959; 3(3):210-29.

244. Mitchell T. Machine learning 1st edition. Education M-H, editor. New York, NY; 1997.

245. Schmidt-Erfurth U, Sadeghipour A, Gerendas BS, Waldstein SM, Bogunović H. Artificial intelligence in retina. Progress in Retinal and Eye Research. 2018; 67:1-29. doi:10.1016/j.preteyeres.2018.07.004.

246. Dhande M. What is the difference between AI, machine learning and deep learning? 2020.

247. Mueller J, Massaron L. Machine learning for dummies. Hoboken, NJ: John Wiley & Sons, Inc; 2016.

248. Bian J, Buchan I, Guo Y, Prosperi M. Statistical thinking, machine learning. Journal of Clinical Epidemiology. 2019; 116:136-7. doi:10.1016/j.jclinepi.2019.08.003.

249. Shai SS, Shai BD. Understanding machine learning - from theory to algorithms: Cambridge University Press; 2014.

250. Oliver M, Ralph N. Risk-sensitive reinforcement learning. Machine Learning. 2002; 49(2-3):267-90. doi:10.1023/A:1017940631555.

251. Shivaram K, Peter S. Characterizing reinforcement learning methods through parameterized learning problems. Machine Learning. 2011; 84(1-2):205-47. doi:10.1007/s10994-011-5251-x.

252. Martínez-Tenor An, Cruz-Martín A, Fernández-Madrigal J-A. Teaching machine learning in robotics interactively: The case of reinforcement learning with lego® mindstorms. Interactive Learning Environments. 2019; 27(3):293-306. doi:10.1080/10494820.2018.1525411.

253. Rajbanshi S. Everything you need to know about machine learning. 2021. Available from: https://www.analyticsvidhya.com/blog/2021/03/everything-you-need-to-know-about-machine-learning/.

254. Goldsmith J, Sun Y, Fried LP, Wing J, Miller GW, Berhane K. The emergence and future of public health data science. Public Health Reviews. 2021; 42:1604023. doi:10.3389/phrs.2021.1604023.

255. Car J, Sheikh A, Wicks P, Williams MS. Beyond the hype of big data and artificial intelligence: Building foundations for knowledge and wisdom. BioMed Central. 2019: 1-5.

256. Donoho D. 50 years of data science. Journal of Computational and Graphical Statistics. 2017; 26(4):745-66.

257. Jheng YC, Kao CL, Yarmishyn AA, Chou YB, Hsu CC, Lin TC, et al. The era of artificial intelligence-based individualized telemedicine is coming. Journal of the Chinese Medical Association. 2020; 83(11):981-3. doi:10.1097/jcma.0000000000000374.

258. Yu VL, Fagan LM, Wraith SM, Clancey WJ, Scott AC, Hannigan J, et al. Antimicrobial selection by a computer. A blinded evaluation by infectious diseases experts. JAMA. 1979; 242(12):1279-82.

259. Sinisi SE, Polley EC, Petersen ML, Rhee SY, van der Laan MJ. Super learning: An application to the prediction of HIV-1 drug resistance. Statistical Applications in Genetics and Molecular Biology. 2007; 6:Article 7. doi:10.2202/1544-6115.1240.

260. Komorowski M, Celi LA. Will artificial intelligence contribute to overuse in healthcare? Critical Care Medicine. 2017; 45(5):912-3. doi:10.1097/ccm.0000000000002351.

261. Panch T, Pearson-Stuttard J, Greaves F, Atun R. Artificial intelligence: Opportunities and risks for public health. Lancet. 2019; 1(1):e13-4. doi:10.1016/S2589-7500(19)30002-0.

262. Aguiar FS, Torres RC, Pinto JV, Kritski AL, Seixas JM, Mello FC. Development of two artificial neural network models to support the diagnosis of pulmonary tuberculosis in hospitalized patients in Rio de Janeiro, Brazil. Medical & Biological Engineering & Computing. 2016; 54(11):1751-9. doi:10.1007/s11517-016-1465-1.

263. Jaeger S, Juarez-Espinosa OH, Candemir S, Poostchi M, Yang F, Kim L, et al. Detecting drug-resistant tuberculosis in chest radiographs. International Journal of Computer Assisted Radiology and Surgery. 2018; 13(12):1915-25. doi:10.1007/s11548-018-1857-9.

264. Lopes UK, Valiati JF. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. Computers in Biology and Medicine. 2017; 89:135-43. doi:10.1016/j.compbiomed.2017.08.001.

265. Andrade BB, Reis-Filho A, Barros AM, Souza-Neto SM, Nogueira LL, Fukutani KF, et al. Towards a precise test for malaria diagnosis in the Brazilian Amazon: Comparison among field microscopy, a rapid diagnostic test, nested PCR, and a computational expert system based on artificial neural networks. Malaria Jounal. 2010; 9:117. doi:10.1186/1475-2875-9-117.

266. Go T, Kim JH, Byeon H, Lee SJ. Machine learning-based in-line holographic sensing of unstained malaria-infected red blood cells. Journal of Biophotonics. 2018; 11(9):e201800101. doi:10.1002/jbio.201800101.

267. Schwalbe N, Wahl B. Artificial intelligence and the future of global health. Lancet 2020; 395(10236):1579-86. doi:10.1016/S0140-6736(20)30226-9.

268. Bellinger C, Mohomed Jabbar MS, Zaïane O, Osornio-Vargas A. A systematic review of data mining and machine learning for air pollution epidemiology. BMC Public Health. 2017; 17(1):907. doi:10.1186/s12889-017-4914-3.

269. Fitzpatrick F, Doherty A, Lacey G. Using artificial intelligence in infection prevention. Current Treatment Options in Infectious Diseases. 2020:1-10. doi:10.1007/s40506-020-00216-7.

270. Flouris AD, Duffy J. Applications of artificial intelligence systems in the analysis of epidemiological data. European Journal of Epidemiology. 2006; 21(3):167-70. doi:10.1007/s10654-006-0005-y.

271. VoPham T, Hart JE, Laden F, Chiang YY. Emerging trends in geospatial artificial intelligence (geoai): Potential applications for environmental epidemiology. Environmental Health. 2018; 17(1):40. doi:10.1186/s12940-018-0386-x.

272. Wong TY, Sabanayagam C. Strategies to tackle the global burden of diabetic retinopathy: From epidemiology to artificial intelligence. Ophthalmologica. 2020; 243(1):9-20. doi:10.1159/000502387.

273. Bi Q, Goodman KE, Kaminsky J, Lessler J. What is machine learning: A primer for the epidemiologist. American Journal of Epidemiology. 2019. doi:10.1093/aje/kwz189.

274. Pan American Health Organization. Artificial intelligence in public health. . Washington, DC; 2021.

275. Kliestik T, Novak A, Lăzăroiu G. Live shopping in the metaverse: Visual and spatial analytics, cognitive artificial intelligence techniques and algorithms, and immersive digital simulations. Linguistic & Philosophical Investigations. 2022; 21:187-202.

276. Langley P. Artificial intelligence and cognitive systems. Artificial Intelligence And Cognitive Systems Quarterly. 2011; 2:1-6.

277. Li X, Shi Y, editors. Computer vision imaging based on artificial intelligence. 2018 International Conference on Virtual Reality and Intelligent Systems (ICVRIS); 2018: IEEE.

278. Vyborny CJ, Giger ML. Computer vision and artificial intelligence in mammography. American Journal of Roentgenology. 1994; 162(3):699-708.

279. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nature Medicine. 2019; 25(1):24-9.

280. Kaul D, Raju H, Tripathy B. Deep learning in healthcare. Deep learning in data analytics: Springer. 2022: 97-115.

281. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: Review, opportunities and challenges. Briefings in Bioinformatics. 2018; 19(6):1236-46.

282. Mittal S, Hasija Y. Applications of deep learning in healthcare and biomedicine. Deep learning techniques for biomedical and health informatics: Springer. 2020: 57-77.

283. Dahiwade D, Patle G, Meshram E, editors. Designing disease prediction model using machine learning approach. 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC); 2019: IEEE.

284. Das TK, editor. A customer classification prediction model based on machine learning techniques. 2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT); 2015: IEEE.

285. Zhang L, Wen J, Li Y, Chen J, Ye Y, Fu Y, et al. A review of machine learning in building load prediction. Applied Energy. 2021; 285:116452.

286. Mishra BK, Kumar R. Natural language processing in artificial intelligence: CRC Press; 2020.

287. Sarikaya R, Hinton GE, Deoras A. Application of deep belief networks for natural language understanding. IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2014; 22(4):778-84.

288. Paris CL, Swartout WR, Mann WC. Natural language generation in artificial intelligence and computational linguistics: Springer Science & Business Media; 2013.

289. Ingrand F, Ghallab M. Robotics and artificial intelligence: A perspective on deliberation functions. AI communications. 2014; 27(1):63-80.

290. Korvel G, Kurowski A, Kostek B, Czyzewski A. Speech analytics based on machine learning. Machine learning paradigms: Springer. 2019: 129-57.

291. Melamed ID, Gilbert M. Speech analytics. Spoken Language Understanding: Systems for Extracting Semantic Information from Speech. 2011:397-416.

292. Scheidt S, Chung QB. Making a case for speech analytics to improve customer service quality: Vision, implementation, and evaluation. International Journal of Information Management. 2019; 45:223-32. doi:10.1016/j.ijinfomgt.2018.01.002.

293. Albayrak N, Özdemir A, Zeydan E, editors. An overview of artificial intelligence based chatbots and an example chatbot application. 2018 26th signal processing and communications applications conference (SIU); 2018: IEEE.

294. van Heerden A, Young S. Use of social media big data as a novel HIV surveillance tool in South Africa. PLoS One. 2020; 15(10):e0239304. doi:10.1371/journal.pone.0239304.

295. Kim MC, Okada K, Ryner AM, Amza A, Tadesse Z, Cotter SY, et al. Sensitivity and specificity of computer vision classification of eyelid photographs for programmatic trachoma assessment. Plos one. 2019; 14(2):e0210463. doi:10.1371/journal.pone.0210463.

296. Dlamini Z, Francies FZ, Hull R, Marima R. Artificial intelligence (AI) and big data in cancer and precision oncology. Computational and Structural Biotechnology Journal. 2020; 18:2300-11. doi:10.1016/j.csbj.2020.08.019.

297. Moyo S, Doan TN, Yun JA, Tshuma N. Application of machine learning models in predicting length of stay among healthcare workers in underserved communities in South Africa. Human Resources for Health. 2018; 16(1):68. doi:10.1186/s12960-018-0329-1.

298. Galle A, Plaieser G, Van Steenstraeten T, Griffin S, Osman NB, Roelens K, et al. Systematic review of the concept 'male involvement in maternal health' by natural language processing and descriptive analysis. BMJ Global Health. 2021; 6(4):e004909. doi:10.1136/bmjgh-2020-004909.

299. Akogun O. Robotic health assistant (feverkit) for he rational management of fevers among nomads in Nigeria. Nurs Leadersh. 2011; 24(2):58-67. doi:10.12927/cjnl.2011.22465.

300. Masso A, Chukwu M, Calzati S. (non) negotiable spaces of algorithmic governance: Perceptions on the ubenwa health app as a ‘relocated’solution. New Media & Society. 2022; 24(4):845-65.

301. Mbunge E, Batani J, Gaobotse G, Muchemwa B. Virtual healthcare services and digital health technologies deployed during coronavirus disease 2019 (COVID-19) pandemic in South Africa: A systematic review. Global Health Journal. 2022; https://doi.org/10.1016/j.glohj.2022.03.001.

302. Ferrein A, Meyer T. A brief overview of artificial intelligence in South Africa. AI Magazine. 2012; 33(1).

303. Adisa O, Botai J, Adeola A, Hassen A, Botai C, Darkey D, et al. Application of artificial neural network for predicting maize production in South Africa. Sustainability. 2019; 11(4) doi:10.3390/su11041145.

304. Cisse M. Look to Africa to advance artificial intelligence. Nature. 2018; 562(7728):461. doi:10.1038/d41586-018-07104-7.

305. Madahana M, Khoza-Shangase K, Moroe N, Mayombo D, Nyandoro O, Ekoru J. A proposed artificial intelligence-based real-time speech-to-text to sign language translator for South African official languages for the COVID-19 era and beyond: In pursuit of solutions for the hearing impaired. South African Journal of Communication Disorders. 2022; 69(2). doi:10.4102/sajcd.v69i2.915.

306. Liyanage H, Liaw ST, Jonnagaddala J, Schreiber R, Kuziemsky C, Terry AL, et al. Artificial intelligence in primary health care: Perceptions, issues, and challenges. Yearbook of Medical Informatics. 2019; 28(1):41-6. doi:10.1055/s-0039-1677901.

307. Manteghinejad A, Javanmard S. Challenges and opportunities of digital health in a post-covid19 world. Journal of Research in Medical Sciences. 2021; 26(1):11. doi:0.4103/jrms.JRMS_1255_20.

308. Oosterhoff JH, Doornberg JN. Artificial intelligence in orthopaedics: False hope or not? A narrative review along the line of gartner's hype cycle. EFORT Open Reviews. 2020; 5(10):593-603.

309. Bini SAMD. Artificial intelligence, machine learning, deep learning, and cognitive computing: What do these terms mean and how will they impact health care? The Journal of Arthroplasty. 2018; 33(8):2358-61. doi:10.1016/j.arth.2018.02.067.

310. Albarrán Lozano I, Molina JM, Gijón C. Perception of artificial intelligence in Spain. Telematics and Informatics. 2021; 63 doi:10.1016/j.tele.2021.101672.

311. Mehta N, Harish V, Bilimoria K, Morgado F, Ginsburg S, Law M, et al. Knowledge and attitudes on artificial intelligence in healthcare: A provincial survey study of medical students. medRxiv. 2021; 10(1) doi:10.15694/mep.2021.000075.1.

312. Yüzbaşıoğlu E. Attitudes and perceptions of dental students towards artificial intelligence. Journal of Dental Education. 2021; 85(1):60-8. doi:10.1002/jdd.12385.

313. Emmert-Streib F, Yli-Harja O, Dehmer M. Artificial intelligence: A clarification of misconceptions, myths and desired status. Frontiers in Artificial Intelligence. 2020; 3:524339. doi:10.3389/frai.2020.524339.

314. Bostrom N, Yudkowsky E. The ethics of artificial intelligence. The Cambridge handbook of artificial intelligence. 2014; 1:316-34.

315. Keskinbora KH. Medical ethics considerations on artificial intelligence. Journal of Clinical Neuroscience. 2019; 64:277-82. doi:10.1016/j.jocn.2019.03.001.

316. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthcare Journal. 2019; 6(2):94-8. doi:10.7861/futurehosp.6-2-94.

317. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digital Medicine. 2018; 1 doi:10.1038/s41746-018-0061-1.

318. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digital Medicine. 2020; 3(1) doi:10.1038/s41746-020-0294-7.

319. Ashrafi K, Fallah R, Hadei M, Yarahmadi M, Shahsavani A. Source apportionment of total suspended particles (TSP) by positive matrix factorization (PMF) and chemical mass balance (CMB) modeling in Ahvaz, Iran. Archives of Environmental Contamination and Toxicology. 2018; 75(2):278-94. doi:10.1007/s00244-017-0500-z.

320. Cheng YH, Yang LS. Characteristics of ambient black carbon mass and size-resolved particle number concentrations during corn straw open-field burning episode observations at a rural site in southern Taiwan. International Journal of Environmental Research and Public Health. 2016; 13(7):688.

321. Das A, Kumar R, Patel SS, Saha MC, Guha D. Source apportionment of potentially toxic elements in street dust of a coal mining area in Chhattisgarh, India, using multivariate and lead isotopic ratio analysis. Environmental Monitoring and Assessment. 2020; 192(6):1-14.

322. Han F, Kota SH, Wang Y, Zhang H. Source apportionment of PM2.5 in Baton Rouge, Louisiana during 2009–2014. Science of The Total Environment. 2017; 586:115-26. doi:https://doi.org/10.1016/j.scitotenv.2017.01.189.

323. Molnár P, Tang L, Sjöberg K, Wichmann J. Long-range transport clusters and positive matrix factorization source apportionment for investigating transboundary PM2.5 in Gothenburg, Sweden. Environmental Science: Processes & Impacts. 2017; 19(10):1270-7. doi:10.1039/c7em00122c.

324. Guo G, Li K, Zhang D, Lei M. Quantitative source apportionment and associated driving factor identification for soil potential toxicity elements via combining receptor models, SOM, and geo-detector method. Science of The Total Environment. 2022; 830:154721.

325. Jing D, Cheng N, Zhang C, Chen Z, Cai X, Li S, et al. A novel approach for VOC source apportionment combining characteristic factor and pattern recognition technology in a Chinese industrial area. Journal of Environmental Sciences. 2022; 121:25-37.

326. Kumar V, Sahu M, Biswas P. Source apportionment of particulate matter by application of machine learning clustering algorithms. Aerosol and Air Quality Research. 2022; 22(3):210240. doi:10.4209/aaqr.210240.

327. Liu X, Lu D, Zhang A, Liu Q, Jiang G. Data-driven machine learning in environmental pollution: Gains and problems. Environmental Science & Technology. 2022; 56(4):2124-33.

328. Shan Y, Shi J. Data mining for source apportionment of trace elements in water and solid matrix. Trace metals in the environment-new approaches and recent advances: IntechOpen; 2019.

329. Singh KP, Gupta S, Rai P. Identifying pollution sources and predicting urban air quality using ensemble learning methods. Atmospheric Environment. 2013; 80:426-37.

330. Chen J, Chen H, Zheng G, Pan JZ, Wu H, Zhang N, editors. Big smog meets web science: Smog disaster analysis based on social media and device data on the web. Proceedings of the 23rd international conference on world wide web; 2014.

331. Lary DJ, Malakar N, Moore A, Roscoe B, Adams ZL, Faruque FS, et al. Estimating the global abundance of ground level presence of particulate matter (PM2.5) Geospatial Health. 2014; 8(3):S611-S30. doi:10.4081/gh.2014.292.

332. Xu Y, Yang W, Wang J. Air quality early-warning system for cities in China. Atmospheric Environment. 2017; 148:239-57. doi:10.1016/j.atmosenv.2016.10.046.

333. Kebalepile MM, Dzikiti LN, Voyi K. Supervised kohonen self-organizing maps of acute asthma from air pollution exposure. International Journal of Environmental Health Research. 2021; 18(21):11071. doi:doi:10.3390/ijerph182111071

334. Lary DJ, Lary T, Sattler B. Using machine learning to estimate global PM2.5 for environmental health studies. Environmental Health Insights. 2015; 9s1 doi:10.4137/EHI.S15664.

335. Oprea M, Iliadis L, editors. An artificial intelligence-based environment quality analysis system. Engineering Applications of Neural Networks; 2011. Berlin, Heidelberg: Springer Berlin Heidelberg.

336. Adams MD, Kanaroglou PS. Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models. Journal of Environmental Management. 2016; 168:133-41. doi:10.1016/j.jenvman.2015.12.012.

337. Watts N, Adger WN, Agnolucci P, Blackstock J, Byass P, Cai W, et al. Health and climate change: Policy responses to protect public health. Lancet. 2015; 386(10006):1861-914. doi:10.1016/s0140-6736(15)60854-6.

338. Wetsman N. Air-pollution trackers seek to fill Africa's data gap. Nature. 2018; 556(7701):284. doi:10.1038/d41586-018-04330-x.

339. Freeman BS, Taylor G, Gharabaghi B, Thé J. Forecasting air quality time series using deep learning. Journal of the Air & Waste Management Association. 2018; 68(8):866-86. doi:10.1080/10962247.2018.1459956.

340. Yang Z, Wang J. A new air quality monitoring and early warning system: Air quality assessment and air pollutant concentration prediction. Environmental Research 2017; 158:105-17. doi:10.1016/j.envres.2017.06.002.

341. Austin E, Coull BA, Zanobetti A, Koutrakis P. A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition. Environment International. 2013; 59:244-54. doi:10.1016/j.envint.2013.06.003.

342. Barrón-Adame JM, Cortina-Januchs MG, Vega-Corona A, Andina D. Unsupervised system to classify SO2 pollutant concentrations in Salamanca, Mexico. Expert Systems with Applications. 2012; 39(1):107-16. doi:10.1016/j.eswa.2011.05.083.

343. Pires JCM, Sousa SIV, Pereira MC, Alvim-Ferraz MCM, Martins FG. Management of air quality monitoring using principal component and cluster analysis—part ii: CO, NO2 and O3. Atmospheric Environment. 2008; 42(6):1261-74. doi:10.1016/j.atmosenv.2007.10.041.

344. Ghasemi JB, Rofouei MK, Amiri N, Maghsoudi A, Zolfonoun E. A chemometric study on the stream sediments of meshkinshahr, NW Iran, using supervised and unsupervised classification methods. Arabian Journal of Geosciences. 2015; 8(5):2853-61. doi:10.1007/s12517-014-1302-5.

345. Jehan S, Khan S, Khattak SA, Muhammad S, Rashid A, Muhammad N. Hydrochemical properties of drinking water and their sources apportionment of pollution in Bajaur agency, Pakistan. Measurement. 2019; 139:249-57.

346. Shi R, Zhao J, Shi W, Song S, Wang C. Comprehensive assessment of water quality and pollution source apportionment in Wuliangsuhai Lake, Inner Mongolia, China. International Journal of Environmental Health Research. 2020; 17(14):5054.

347. Govender P, Sivakumar V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980-2019). Atmospheric Pollution Research. 2020; 11(1):40-56. doi:10.1016/j.apr.2019.09.009.

348. Chakraborty S, Nagwani NK, Dey L. Performance comparison of incremental k-means and incremental dbscan algorithms. arXiv preprint arXiv:1406.4751. 2014;

349. Khan K, Rehman SU, Aziz K, Fong S, Sarasvady S, editors. Dbscan: Past, present and future. The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014); 2014: IEEE.

350. Bousiotis D, Beddows D, Singh A, Haugen M, Diez S, Edwards PM, et al. A study on the performance of low-cost sensors for source apportionment at an urban background site. Atmospheric Measurement Techniques. 2022; 15(13):4047-61.

351. Khorshidi N, Parsa M, Lentz DR, Sobhanverdi J. Identification of heavy metal pollution sources and its associated risk assessment in an industrial town using the k-means clustering technique. Applied Geochemistry. 2021; 135:105113.

352. Liang C-S, Wu H, Li H-Y, Zhang Q, Li Z, He K-B. Efficient data preprocessing, episode classification, and source apportionment of particle number concentrations. Science of the Total Environment. 2020; 744:140923.

353. Chan AK, Wozny TA, Bisson EF, Pennicooke BH, Bydon M, Glassman SD, et al. Classifying patients operated for spondylolisthesis: A k-means clustering analysis of clinical presentation phenotypes. Neurosurgery. 2021; 89(6):1033-41. doi:10.1093/neuros/nyab355.

354. Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. Journal of intelligent information systems. 2001; 17(2-3):107-45. doi:10.1023/A:1012801612483.

355. Jain AK, Murty MN, Flynn PJ. Data clustering a review. ACM Computing Surveys. 1999; 31(3):264-323. doi:10.1145/331499.331504.

356. Mittal M, Goyal LM, Hemanth DJ, Sethi JK. Clustering approaches for high-dimensional databases: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019; 9(3) doi:10.1002/widm.1300.

357. Pollard D. Strong consistency of k-means clustering. The Annals of Statistics. 1981; 9(1):135-40.

358. Lorbeer B, Kosareva A, Deva B, Softić D, Ruppel P, Küpper A. Variations on the clustering algorithm BIRCH. Big Data Research. 2018; 11:44-53.

359. Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large databases. ACM Sigmod Record. 1996; 25(2):103-14.

360. Goubko M, Veremyev A. Bilinear matrix equation characterizes laplacian and distance matrices of weighted trees. Discrete Applied Mathematics. 2021; 305:1-9. doi:10.1016/j.dam.2021.08.025.

361. Aggarwal CC, Reddy CK. Data clustering: algorithms and applications.Ser. Chapman & hall/crc/.Data Mining and Knowledge Discovery. CRC Press. 2013.

362. von Luxburg U, Belkin M, Bousquet O. Consistency of spectral clustering. The Annals of Statistics. 2008; 36(2):555-86.

363. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC Medical Research Methodology. 2012; 12(1):1-10. doi:10.1186/1471-2288-12-96.

364. Lee KJ, Tilling KM, Cornish RP, Little RJA, Bell ML, Goetghebeur E, et al. Framework for the treatment and reporting of missing data in observational studies: The treatment and reporting of missing data in observational studies framework. Journal of Clinical Epidemiology. 2021; 134:79-88. doi:10.1016/j.jclinepi.2021.01.008.

365. Chen J, Hunter S, Kisfalvi K, Lirio RA. A hybrid approach of handling missing data under different missing data mechanisms: Visible 1 and varsity trials for ulcerative colitis. Contemporary Clinical Trials. 2021; 100 doi:10.1016/j.cct.2020.106226.

366. Jeong CW, Washington SL, Herlemann A, Gomez SL, Carroll PR, Cooperberg MR. The new surveillance, epidemiology, and end results prostate with watchful waiting database: Opportunities and limitations. European Urology. 2020; 78(3) doi:10.1016/j.eururo.2020.01.009.

367. Lupattelli AP, Wood MEP, Nordeng HP. Analyzing missing data in perinatal pharmacoepidemiology research: Methodological considerations to limit the risk of bias. Clinical Therapeutics. 2019; 41(12):2477-87. doi:10.1016/j.clinthera.2019.11.003.

368. Mulla ZD, Seo B, Kalamegham R, Nuwayhid BS. Multiple imputation for missing laboratory data: An example from infectious disease epidemiology. Annals of Epidemiology. 2009; 19(12):908-14. doi:10.1016/j.annepidem.2009.08.002.

369. Gómez-Carracedo MP, Andrade JM, López-Mahía P, Muniategui S, Prada D. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemometrics and Intelligent Laboratory Systems. 2014; 134:23-33. doi:10.1016/j.chemolab.2014.02.007.

370. Hadeed SJ, O'Rourke MK, Burgess JL, Harris RB, Canales RA. Imputation methods for addressing missing data in short-term monitoring of air pollutants. Science of the Total Environment. 2020; 730 doi:10.1016/j.scitotenv.2020.139140.

371. Junger WL, De Leon AP. Imputation of missing data in time series for air pollutants. Atmospheric Environment. 2015; 102:96-104. doi:10.1016/j.atmosenv.2014.11.049.

372. Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality data sets. Atmospheric Environment. 2004; 38(18):2895-907. doi:10.1016/j.atmosenv.2004.02.026.

373. Alamoodi AH, Zaidan BB, Zaidan AA, Albahri OS, Chen J, Chyad MA, et al. Machine learning-based imputation soft computing approach for large missing scale and non-reference data imputation. Chaos, Solitons & Fractals. 2021; 151 doi:10.1016/j.chaos.2021.111236.

374. Awan SE, Bennamoun M, Sohel F, Sanfilippo F, Dwivedi G. Imputation of missing data with class imbalance using conditional generative adversarial networks. Neurocomputing. 2021; 453:164-71. doi:10.1016/j.neucom.2021.04.010.

375. Moritz S, Bartz-Beielstein T. Imputets: Time series missing value imputation in R. RJ. 2017; 9(1):207. doi:10.32614/RJ-2017-009.

376. Little RJA, Rubin DB. Statistical analysis with missing data. New York, United States: John Wiley & Sons, Incorporated; 2002.

377. Mirzaei A, Carter SR, Patanwala AE, Schneider CR. Missing data in surveys: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy. 2021; 18(2):2308-16. doi:10.1016/j.sapharm.2021.03.009.

378. Rubin DB. Inference and missing data. Biometrika. 1976; 63(3) doi:10.2307/2335739.

379. Randahl D. Raoul: An R-package for handling missing data. Uppsala universitet, Statistiska institutionen; 2016.

380. Allison P. Missing data. In the sage handbook of quantitative methods in psychology. Maydeu-Olivares REMaA, editor: Sage Publications; 2009.

381. Li C. Little's test of missing completely at random. The Stata Journal. 2013; 13(4):795-809. doi:10.1177/1536867X1301300407.

382. Donders ART, Van Der Heijden GJ, Stijnen T, Moons KG. A gentle introduction to imputation of missing values Journal of Clinical Epidemiology 2006; 59 (10):1087–91.

383. Schafer JL. Analysis of incomplete multivariate data. London ;: Chapman & Hall; 1997.

384. Allison PD, editor. Handling missing data by maximum likelihood. SAS global forum; 2012.

385. Liu Y, Brown SD. Comparison of five iterative imputation methods for multivariate classification. Chemometrics and Intelligent Laboratory Systems. 2013; 120:106-15. doi:10.1016/j.chemolab.2012.11.010.

386. King G, Honaker J, Joseph A, Scheve K. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. American political science review. 2001; 95(1):49-69.

387. Quinteros ME, Lu S, Blazquez C, Cárdenas-R JP, Ossa X, Delgado-Saborit JM, et al. Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile. Atmospheric Environment. 2019 200:40–9.

388. Van Buuren S. Flexible imputation of missing data. Boca Raton, FL: CRC Press; 2012.

389. Allison P. Why you probably need more imputations than you think. 2012. Available from: https://statisticalhorizons.com/more-imputations.

390. Paradis AD, Fitzmaurice GM, Koenen KC, . BS. A prospective investigation of neurodevelopmental risk factors for adult antisocial behavior combining official arrest records and self-reports. Journal of Psychiatric Research. 2015; 68:363–70.

391. Bethlehem J. Applied survey methods: A statistical perspective John Wiley & Sons; 2009.

392. Moritz S. Package imputets,. 2017. Available from: http://cran.r-project.org/web/packages/imputeTS/imputeTS.pdf.

393. Moritz S, Sardá A, Bartz-Beielstein T, Zaefferer M, Stork Jr. Comparison of different methods for univariate time series imputation in R.;2015.

394. Little RJ. Missing-data adjustments in large surveys Journal of Business & Economic Statistics. 1988; 6 (3):287–96.

395. Engels JM, Diehr P. Imputation of missing longitudinal data: A comparison of methods. Journal of Clinical Epidemiology. 2003; 56(10):968–76.

396. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? International Journal of Methods in Psychiatric Research. 2011; 20(1):40-9. doi:10.1002/mpr.329.

397. Van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2011; 45(1):1-67.

398. Junger WL, de Leon AP. Package 'mtsdi'. 2018. Available from: https://cran.rstudio.com/web/packages/mtsdi/mtsdi.pdf.

CHAPTER 3: METHODOLOGY

This chapter explains the methods used to achieve the three PhD thesis objectives. The methods used are the survey among postgraduate students, the classification and regression trees analysis and unsupervised ML methods to investigate the joint effects of air pollutant mixtures. Lastly, the methods of the PMF for source apportionment and unsupervised ML cluster algorithms for source apportionment, are explained.

3.1. METHODS

3.1. KNOWLEDGE, ATTITUDES AND PERCEPTIONS OF THE USE OF AI APPLICATIONS IN PUBLIC HEALTH RESEARCH AMONG POSTGRADUATE STUDENTS

3.1.1. STUDY SETTING AND STUDY DESIGN

The study was conducted in the School of Health Systems and Public Health (SHSPH) at the University of Pretoria, South Africa. A cross-sectional study was done within SHSPH's Postgraduate Diploma in Public Health students. In this study, an online cross-sectional survey was sent to 758 enrolled diploma students. The questionnaire was conducted between 13 June and 19 June 2022. The cross-sectional study was used to assess the knowledge, attitudes, and perceptions of the use of AI applications in public health. This study provided an overview of the relationship between the use of AI applications in public health and how it is perceived by the students.

Ethical clearance was obtained prior to distributing the questionnaire (Appendix 1). An online questionnaire (Appendix 2) was prepared using Qualtrics. This questionnaire was accompanied with a consent form (Appendix 3) that all students acknowledged before choosing to participate in the study. The methodology used in this study was adapted from a Canadian study conducted among medical students.[1]

3.1.2. SURVEY INSTRUMENT (QUESTIONNAIRE) AND HEALTH OUTCOME MEASUREMENTS

The questionnaire was created in Qualtrics, after which an online invitation was sent to students, enrolled for an online module. Consent was obtained from the students on the first page of the survey.

The questionnaire had three designated sections that focused on: (A) demographics of the respondents that included their gender, age, residence, and interaction with computer science and AI; (B) basic knowledge of AI terminology including AI, ML, DL, neural networks (NN), and 'algorithm'; and (C) perceptions of AI use in public health. This last section was separated into four subsections addressing the perceptions of AI and task performance at individual primary care, health systems, and population; perceptions of AI and impact on public health careers; perceptions of AI and ethics; and perceptions of AI and public health education.

3.1.3. STATISTICAL ANALYSIS

The data from Qualtrics was exported to a Microsoft Excel spreadsheet and then imported to STATA 15, where all demographic and statistical analysis was performed. The scaled questions regarding perceptions and attitudes were reported in percentages. Content analysis was used for the open-ended responses obtained from the questionnaires. Chi-square tests were applied to assess the correlations between sections B, C, and some of the demographic information captured in section A. The significance level was $p<0.05$.

3.2. JOINT EFFECT OF SO_2, NO_2, O_3 $PM_{2.5}$ AND PM_{10}, ON RESPIRATORY AND CARDIOVASCULAR DISEASE HOSPTIAL ADMISSIONS.

3.2.1. STUDY DESIGN

The association between daily air pollution mixtures and daily counts of RD and CVD hospital admissions were investigated using a case-crossover epidemiology study design. A case-crossover study design is often used in air pollution epidemiological studies to establish a causal relationship between air pollutants and acute health outcomes.[2-3] This design is a variant of the time-series study design. It shows the

effects of transient acute exposures on acute emergency events, while comparing each person's exposure in a time period just prior to a case-defining event with the person's exposure at other times (i.e. unit of analysis is on the individual level).[4]

A case-crossover study design is an analytical epidemiological approach where each case as its own controls and is often used to investigate short-term effects of intermittent pollutant exposure on the onset of acute outcomes.[5] The design is recommended to study the effects of short-term air pollution exposure on health outcomes with an abrupt onset, such as myocardial infarction or asthma attacks.[6-7] Since the cases serve as their own controls confusion caused by individual characteristics is eliminated. Additionally, it allows for the use of routinely monitored air pollution information while simultaneously allowing the study of individuals (rather than days) as the unit of observation.[8] Having the cases as their own controls also limits selection bias and increases efficiency.[9] For each RD and CVD hospital admission, the exposure variables (SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}) for that day was compared to a referent group, where all the exposure mixture variables were at their lowest concentrations. This study design controls for seasonal trends.[10-11]

3.2.2. STUDY SETTING AND POPULATION

The study area included Vereeniging and Vanderbijlpark. They are located in Sedibeng district municipality, Gauteng Province, South Africa, found in the VTAPA.

3.2.3. HOSPITAL ADMISSION DATA

The individual-level cause-specific hospital admission data for Vereeniging and Vanderbijlpark (1 January 2011-29 February 2020) was requested from a private hospital group. Some of this data (January 2011-October 2016) has also been used in a PhD thesis completed in 2019, where Prof Wichmann was a co-supervisor, within the same area.[12]

Cause-specific hospital admission data was indicated by the International Classification of Diseases 11th Revision codes. The specific codes of interest were CD11 RD codes J00-J99 and CVD codes I00-I99.[13]

3.2.4. AIR POLLUTION AND METEOROLOGICAL DATA

SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} were investigated for the study period 1 Jan 2011 to 29 February 2020 (before the first COVID-19 case in South Africa) and downloaded as 24-hour averages, from the SAAQIS website.[14] This air pollution source is the same used in local studies.[15-17] Six air pollution monitoring stations (Diepkloof, Klipriver, Sebokeng, Sharpeville, Three Rivers, and Zamdela) maintained by the South African Weather Services (SAWS) were assigned to the VTAPA.

Although there are thirteen air monitoring stations within the area, the six air pollution monitoring stations continuously assess real-time concentrations of the criteria air pollutants using equivalent methods of the United States Environmental Protection Agency and in accordance with ISO 17025 guidelines (Figure 3.1).[18] Additionally, the six stations collected meteorological data such as ambient temperature, relative humidity, rainfall, wind speed, and wind direction. Temperature and relative humidity are by default included as confounders in the air pollution epidemiology studies that apply the case-crossover epidemiology design.[2-3,19] Eight other air pollution monitoring stations in the VTAPA area were considered, but these stations had large proportions of missing data and could not be applied to this PhD project.

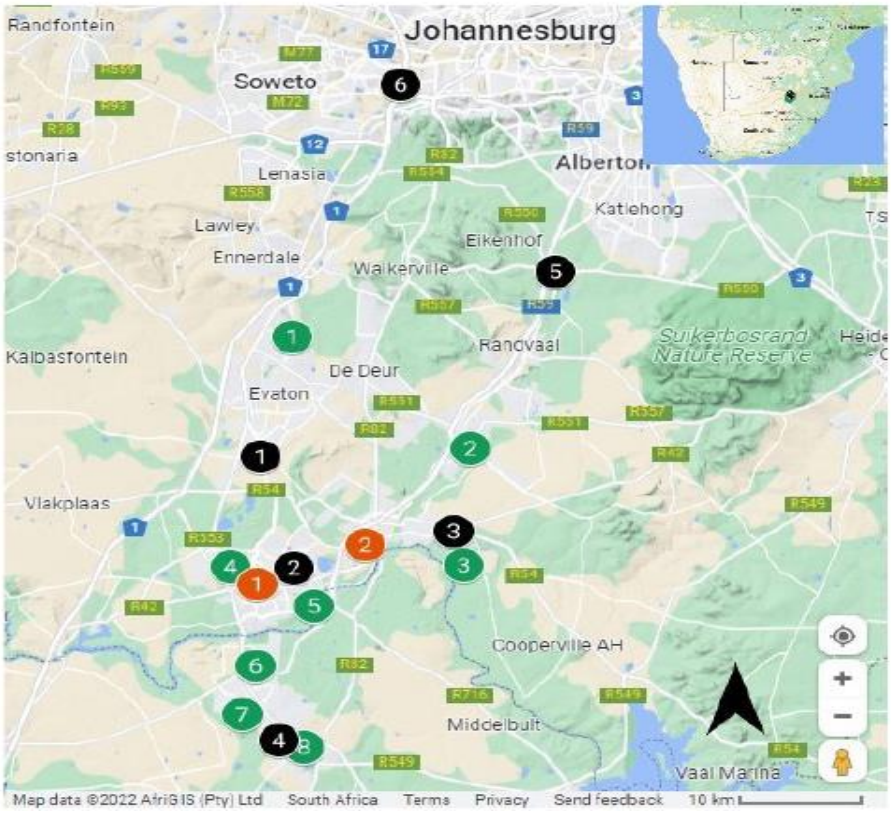


Figure 3.1: Map of the location of the air pollution stations in the Vaal Triangle Air Pollution Priority Area and the two hospitals. Red symbols are the hospitals (1) Vanderbijlpark and (2) Vereeniging. Black symbols are the pollution monitoring stations of the VTAPA: (1) Sebokeng, (2) Sharpeville, (3) Three Rivers, (4) Zamdela, (5) Kliprivier, (6) Diepkloof. Green symbols are the pollution monitoring stations of the City of Johannesburg:(1) Orange Farm, (2) Meyerton, (3) Randwater, (4) Vanderbijlpark, (5) North West University, (6) Bongani Mabaso Eco Park, (7) AJ Jacobs, (8) Leitrim.[20]

Daily 24-hour averages (midnight-to-midnight), as well as hourly data for the pollutants measured at the selected sites, were downloaded from the SAAQIS website.[14] Afterwards, combined 24-hour averages for the pollutants (SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}) were calculated across the site of interest. Although the air pollution data are readily accessible and available, there were many gaps of missing data, requiring the use of imputation for the missing data.

3.2.5. DATA ANALYSIS

3.2.5.1. IMPUTATION OF EXPOSURE DATA

Before imputation could be implemented the data was carefully cleaned. SAWS implements a detailed data validation and quality assurance process, according to the New Zealand “Good Practice Guide for Air Quality Monitoring and Data Management” (2009).[18] During this process, the data are checked for any zero drift, and if identified the figure is replaced with the values recoded at last calibration. Outliers are also

identified if they are outside the range of three standard deviations of the calculated mean. If these outliers represent a legitimate spike in the concentration (3 standard deviations of the mean of values within an hour of either side of the data point), they are retained. Lastly, negative values less than -1 were removed from the validated data.[18]

The datasets went through further cleaning of any overlooked outliers. The daily average data for the identified days were compared with the hourly values recorded for that day. Additionally, three to five days of hourly data from before and after the identified outlier was compared to see whether the figure had to be deleted, or if there was an error in recording the value that simply needed correction.

3.2.5.1.1. MISSING DATA ASSUMPTION

It cannot be fully determined what the cause of missing data was. However, the reason for the incomplete data may have been due to premature shutdown of the monitors, battery power loss, or equipment failure occurring at some point during a 24-hour monitoring period. Other reasons for interruptions in the monitoring of outdoor air quality could be extreme weather conditions beyond the range of the manufacturers' recommended conditions.[21] However, exploring the conditions proximal to the shutdown of these monitors indicates the ambient conditions were within range of the manufacturers' operating environment, suggesting MNAR (missingness not at random) was not present. Since MNAR was unlikely, the MAR (missingness at random) assumption remained the main mechanism of 'missingness' for the air pollution and meteorological data.

3.2.5.1.2. METHODS OF HANDLING MISSING DATA

The imputation methods used in the study were derived from recent studies by Hadeed et al,[21] and Junger and De Leon, 2015[22]. R statistical program was used for all the imputations. The R packages used were naniar, imputeTS, mice, and mtsdi. Mice and mtsdi imputations were also run using meteorological conditions i.e., relative humidity, temperature and wind speed, as part of the comparison. Mice imputation was the overall imputed dataset used in this section of the study. This method has also been implemented in a study conducted in the same area.[20]

3.2.6.2. DESCRIPTIVE STATISTICS

Descriptive statistics were reported for the health outcomes, air pollutants, and weather conditions. Non-parametric tests were applied on the variables because there was no Gaussian distribution. All descriptive statistics and non-parametric tests were performed in R.

3.2.6.3. CLASSIFICATION AND REGRESSION TREE STATSTICAL ANALYSIS (CART)

A modified CART method was applied,[23] which has been used in other studies.[24-25] The number of sets of all possible joint effects are grouped into quartiles and this is based on the daily concentrations of each pollutant (e.g. $PM_{2.5}$, NO_2, and SO_2), i.e. each pollutant had one reference level (lowest quartile) and three higher levels (three other quartiles). This simplification yields four to the power of three (64) different types of days, each of which can be viewed as a unique mixture. Days when SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} were all in the lowest quartile, were placed as the referent group and the remaining days were used to estimate the regression tree.[23]

The CART analysis took all possible joint effects and collapsed them into groups that have similar predicted values for the outcome through a recursive partitioning process. The end product of a typical CART analysis is a dendrogram illustrating the paths of dichotomous splits.[23,25] Every tree started with a 'root node' that contains the observations from which the tree had grown. The observations were then partitioned into two 'child nodes' based on the value of an independent predictor variable. The resulting child nodes each contain a subset of the original observations. Each child node could be further partitioned, again, based on the value of an independent predictor variable. The process continued until a set of partitioning criteria was no longer met, resulting in 'terminal nodes'. Terminal nodes, by definition, cannot have offspring. The collection of terminal nodes forms a complete partition of the observations in the root node.[23,25] The criteria for a terminal node is that it lies above the stated alpha, i.e. $\alpha=0.15$, and the number of days are below 60.[23]

Pollutants within a mixture, e.g. PM_{10}, NO_2, and SO_2, were parameterised as dichotomous variables representing each ordinal split of the quartiles (e.g. comparing SO_2 quartile one vs. SO_2 quartiles two to four) and were then introduced one at a time

in a regression model. The pollutant-split resulting in the smallest p-value was selected as the first split, and the dataset was partitioned accordingly. This process was then repeated for each subset of the data until the p-values for the remaining splits were below a given alpha, resulting in the formation of a terminal node. Likelihood ratio tests were also applied for the inclusion of the terminal nodes in the final model.[23,25]

Due to the nature of the method used and the high correlation between PM_{10} and $PM_{2.5}$, the pollutants were divided into seven air pollution mixtures:

Mixture 1: PM_{10}, NO_2, and SO_2

Mixture 2: $PM_{2.5}$, NO_2, and SO_2

Mixture 3: PM_{10}, NO_2, and O_3

Mixture 4: $PM_{2.5}$, NO_2, and O_3

Mixture 5: PM_{10}, SO_2, and O_3

Mixture 6: $PM_{2.5}$, SO_2, and O_3

Mixture 7: O_3, NO_2, and SO_2

Descriptive statistics for the air pollution clusters were formed using unsupervised Machine Learning (ML) methods. Thereafter, Wilcoxon rank-sum and Kruskal Wallis tests were applied to determine the differences between and among the formed clusters.

3.2.6.4. REGRESSION MODELLING

Similar to other time-stratified case-crossover studies, standard time-series quasi-Poisson regression models were applied with daily aggregated health data.[26] The daily number of hospital admissions (frequency) usually has a quasi-Poisson distribution, i.e. over-dispersed, meaning very few days with a small number of hospital admissions. The distribution has a long tail as the frequency declines, meaning very few days had a large number of hospital admissions.

Where:

$Y_t \sim \text{Poisson}(\mu_t)$ Equation (1)

$\log(\mu_t) = \alpha + \beta POLQ_t + \gamma Tapp_t + \eta pubhol_t + \lambda Strata_t$ Equation (2)

Where:

t is the day of the hospital admission

Y_t is the hospital admission count on day t

α is the intercept

$POLQ_t$ is the ordinal variable of the air pollutant quartiles on day t

β is a vector of coefficients for POLQ

$Tapp_t$ is the linear term of apparent temperature

γ is a vector of coefficients for Tapp

$pubhol_t$ is a binary variable for public holidays.

η indicates the vector of coefficients for DOW and pubhol

$Strata_t$, the model which is conditioned on a categorical variable of the year and calendar month used to control for long-term trend and seasonality (e.g. from 12011 (January 2011) to 22020 (February 2020), and the day of the week on day, and λ is a vector of coefficients for Strata.

Temperature alone has been considered as a confounder and did not consider relative humidity despite relative humidity playing a significant role in health effects.[27] Apparent temperature (Tapp) is a well-established confounder of the air pollution hospital admissions or air pollution morbidity relationship.[28] Extreme levels of heat, weather high or low can increase morbidity and hospital admissions. Tapp is a better indicator of health effects than including temperature and relative humidity individually.[3,29] This is because it incorporates temperature, humidity, and sometimes wind speed.

Tapp was calculated using the following equations:[20,30]

Saturation vapour pressure

$= 6.112 \times 10^{(7.5 \times \text{temperature °C}/ (237.7 + \text{temperature °C})}$ Equation (3)

Actual vapour pressure

= (relative humidity (%) × saturation vapour pressure)/100 Equation (4)

Dew point temperature °C

= (−430.22 + 237.7 × ln (actual vapour pressure))/−ln (actual vapour pressure) + 19.08) Equation (5)

Apparent temperature °C

−2.653 + (0.994 × temperature °C) + 0.0153 × (dew point temperature °C) Equation (6)

The two-day cumulative average (i.e. average of lag0 and lag1) of the air pollution quartiles and Tapp was used in the models. Lag0 refers to the air pollution or Tapp level on the day of hospitalisation and lag1 refers to the levels the day before hospitalisation. There is no default method to include lags of air pollutants and Tapp in models.[31-34] The air pollution quartiles at lag 0-1 was used in the models. Lag 0-1 is usually the most significant and strongest.

The shape (i.e. linear or non-linear) of the association between the Tapp and hospital admissions was studied using SAS and R. The models with a non-linear term for Tapp and a linear term for Tapp were compared using log likelihood ratio tests. The non-linear term for Tapp will be included in the regression models as a natural spline with three degrees of freedom. Separate regression models were run for RD and CVD hospital admissions for combined ages and sexes.

The joint effect of a SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} mixture (i.e. various day types included in a terminal node) were estimated using the base model that included a binary variable of that mixture. The base model is the model specified in Equation (2), but without including the $POLQ_t$ variable. Prior to running these models, the previously

withheld referent group (day type 111) was included again in the dataset. The reference level of the binary mixture variable was the SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} mixtures when the levels of all air pollutants within the different mixtures were in the lowest quartile within the stated pollutant mixture, i.e. day type 111. The joint effects were then reported as rate ratios (RRs) along with the 95% confidence intervals (CI), as done in other studies.[14,15]

3.2.6.5. UNSUPERVISED MACHINE LEARNING METHODS

Unsupervised ML clustering methods were then used to determine the effects of mixtures that include all five air pollutants and how these mixtures would affect RD and CVD hospital admissions. The methods used in this section were run in R using tidyverse, cluster, factoextra, NbClust, fpc, caret, and mlbench packages. Before setting the number of clusters, all data was scaled using standardised scaling, thereafter the optimal number of clusters was determined using the 'elbow', silhouette, and gap statistics methods.[35]

K-means clustering is a simple and commonly used[36-37] unsupervised ML clustering algorithm. It is a centroid based algorithm that tries to minimise the variance of data points within a cluster.[38] Although there can be limitation in converging during the clustering process,[39] there are a number of clustering methods that reduce the possible error by using non-parametric settings converge.[40] K-means clustering executes dimension reduction of a dataset by distributing the data points to *k* number of clusters.[41] The k-means clustering method has been used to assess the joint effects, as in a study by Riches et al.[42]

When a dataset D = {x1, x2, x3,.., xn} is assigned *k* number of clusters, where C = {c1, c2, c3, …, ck} is the set of cluster centroids. The sum of squares estimated errors (SSE) can be calculated using:[43-44]

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} |x_i - c_k|^2$$

Equation (7)

where *xi* is a point and *ck* is the centroid of cluster Ck.[43]

The other clustering method used was spectral clustering, under the Laplacian and normalised Laplacian matrices. First a similarity matrix was made using a Gaussian kernel, or the Radial Basis Function (RBF) kernel to group the dataset. Thereafter the Laplacian and normalised Laplacian matrices functions were created. The Laplacian matrix is derived by:[43,45]

$$L(G) = D(G) - W(G) \qquad \text{Equation (8)}$$

Where D(G) is the diagonal matrix constructed with the degrees of node or vertex.[43]

W(G) the similarity matrix for G. w (i, j), the elements of W(G) represent the weight of the edge connected by the nodes i and j, where:[43]

$$w(i,j) = \begin{cases} 1, if\ there\ is\ an\ edge\ joining\ vertices\ i\ and\ j \\ 0, otherwise \end{cases} \qquad \text{Equation (9)}$$

Each node or vertex represents a data point.

Using the normalised Laplacian matrix, there are two variants similar to that in the Laplacian matrix, and this is defined by: [43]

$$L_{sym}(G) = D(G)^{\frac{-1}{2}} L(G) D(G)^{\frac{-1}{2}} = I - D(G)^{\frac{-1}{2}} W(G) D(G)^{\frac{-1}{2}}$$

$$L_{rm}(G) = D(G)^{-1} L(G) = I - D(G)^{-1} W(G) \qquad \text{Equation (10)}$$

Where $L_{sym}(G)$ and $L_{rm}(G)$ refer to a symmetric matrix and the random walk perspective Laplacian matrix, respectively, and D(G) and L(G) have usual meanings.[43]

Lastly, DBSCAN clustering was applied and this method clustered the data according to the arbiter shapes that were dependent on the radius set, which can be set by the researcher; thereafter, a minimum number of points for a distance within the radius were grouped together.[46-47] The clusters obtained through the different methods were then inserted into Equation (6), replacing the $POLQ_t$, in order to calculate the risk. All unsupervised ML clustering was run in R using the tidyverse, broom, and gnm packages.

3.3. SOURCE APPORTIONMENT ANALYSIS

A commonly used method of source apportionment was compared with different unsupervised ML clustering algorithms for source apportionment. This was done to assess the differences between the two methods and to investigate whether the unsupervised ML clustering could address some of the shortcomings of PMF (as stated in Chapter 2).

The PMF data was available from studies previously conducted at the SHSPH, University of Pretoria.[48-49] The two studies had available data from 18 April 2017 to 28 February 2020. The sampling site was located at the SHSPH, University of Pretoria, South Africa. Samples were collected on the roof of the HW Snyman South Building, Prinshof campus (S 25°43'53" E 28°12'01") (Figure 3.2). During this time period, $PM_{2.5}$ samples were collected every third day for twenty-four hours, with duplicates every fifth measurement. An additional year of sampling was done, where samples were collected every sixth day, due to COVID-19 lockdown regulations in South Africa. There were five lockdown levels: lockdown level five was implemented between 27 March and 30 April 2020, followed by level four (1-31 May 2020), level three (1 June to 17 August 2020), level two (18 August to 20 September 2020) and level one (21 September to 28 December 2020). The country was on an adjusted level three lockdown from 29 December 2020 until the end of the PMF study period (12 February 2021).

Figure 3.2: Image of sampling site, derived from Google Earth.

The sample area located in Gezina, Pretoria, is an urban area far from industrial manufacturing companies, freeways, and highways. However, it is situated near the major Steve Biko Gezina Road, and within five to ten kilometres of the central business district (Pretoria CBD). It is also in close proximity to the Tshwane District hospital incinerator.[50] The $PM_{2.5}$ measurements were collected over approximately 46 months (17 April 2017 to 12 February 2021), totalling 428 samples. Of the 428 $PM_{2.5}$ samples, 124 samples were collected from 19 April 2018 to 23 April 2019 for my MSc (Epidemiology) project.[50]

Gravimetric analysis was used to determine the mass of the $PM_{2.5}$ samples, using a Mettler Toledo microbalance located at the Air Quality Laboratory, SHSPH. The weighing followed a standard operating procedure (SOP), where three field blanks were used before and after each batch of twenty filters.[50] The SOP used for the weighing procedure was a modified version of the SOP used in the ULTRA study.[51] The filters were conditioned for at least twenty-four hours before weighing in the weighing room of the Air Quality Laboratory, SHSPH. The temperature and relative humidity in the weighing room are maintained at 21.0 ± 1.0°C and 50 ± 5%, respectively.[51]

Soot analyses were also performed on the collected $PM_{2.5}$ samples using a modified SOP (i.e. reflectance analyses).[52-53] The black soot index analyses were done using the M43D smoke stain reflectometer (Diffusion Systems Ltd., London, UK) at the Air Quality Laboratory of the SHSPH. Black carbon (BC) and UVPM (a proxy for organic carbonaceous particulate matter absorbing UV light at 370 nm) analyses were performed using a Model OT21 Optical Transmissometer (Magee Scientific Corp., Berkeley, CA USA) at the Department of Occupational and Environmental Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Sweden. The additional absorption of UV light, at 370 nm, due to the organics, indicate the presence of biomass burning.

For the PMF analysis the species of $PM_{2.5}$ samples were determined through the non-destructive process of X-ray Fluorescence (XRF). The XRF analyses were performed using a XEPOS 5 energy-dispersive X-ray fluorescence (EDXRF) spectrometer (Spectro analytical instruments GmbH, Germany) at the Department of Chemistry and Molecular Biology, Atmospheric Science division, University of Gothenburg, Sweden. The concentrations of the following elements were analysed: S, Cl, Si, K, Ca, Ti, V, Fe, Ni, Cu, Zn, As, Se, Br, Sb, Ba, Pb, and U. All concentrations that were detected above the limit of detection (LoD) were subtracted from the concentrations of the same elements in the samples.[48-49,54-56] There were fifteen species used in the PMF analysis from 18 April 2017 to 12 February 2021: $PM_{2.5}$, BC, UV-PM, S, Cl, Zn, Si, Fe, K, Ca, Ni, Ti, Br, Cu, and U.

3.3.1. POSITIVE MATRIX FACTORISATION (PMF) ANALYSIS

The Positive Matrix Factorization (PMF) technique was applied to identify the possible sources of pollution that contributed to $PM_{2.5}$ in Pretoria from 18 April 2017 to 12 February 2021. This method was carried out by using the U.S. Environmental Protection Agency software EPA-PMF 5.0, applying the multilinear engine (ME) technique via the inbuilt ME-2 program.[57] This has been widely used in ambient PM source apportionment studies.[54,58-63] The entire PMF operation process can be seen below in Figure 3.3.

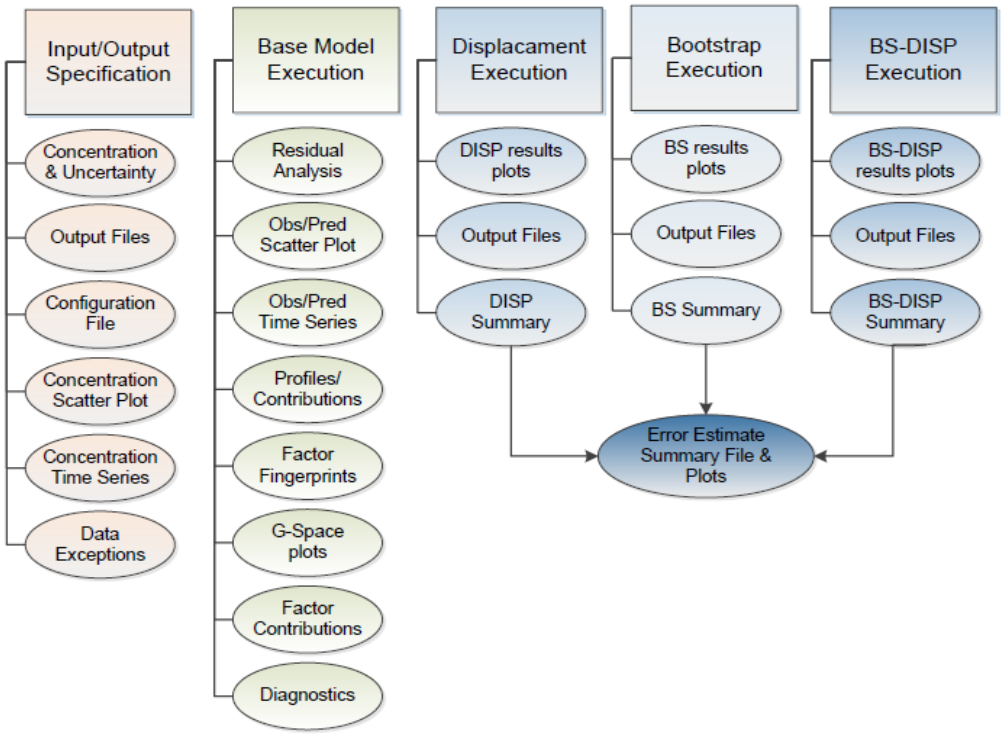


Figure 3.3: Flow chart of operations within EPA PMF – Base Model.[64]

PMF is a multivariate receptor model concept that estimates source profiles and their contributions based on a weighted least square approach.[65] The task of the PMF model in equation eleven is to obtain the unknown matrices, G and F, by the iterative treatment of a least square method.[48-49,56]

The following equations are used in the PMF analysis:

$$X = GF + E \quad \text{..... Equation (11)}$$

Where:

X = the data matrix (size m x n) consisting of n chemical components analysed in m samples

G = the source contribution to each sample (size m x p) for p factors

F = the matrix of source profile (size p x n)

E = the residual

The main mission of the iteration is to minimize the Q value, which is defined in equation twelve: [48-49,56,66]

$$Q(E) = \sum_{i=1} \sum_{j=1} \{eij/sij\}2 \quad \text{..... Equation (12)}$$

Where:

(eij) = squares of residual

(sij) = error estimates of data points

During PMF analysis, the elemental mass concentration is recalculated to their mean oxidized mass concentration, where applicable, in all the analyses.[67] Concentration data and corresponding uncertainty data are used in PMF analysis. The uncertainty is calculated using equation below: [22,67]

$$\text{Uncertainty} = (0.05 * X_{ij}) + DL_{ij} \quad \text{Equation (13)}$$

Where X_{ij} and DL_{ij} are the concentration and detection limit of the j^{th} species in the i^{th} sample, respectively. The concentration and uncertainty are needed in order to check the signal-to-noise ratio $(S/N)_j$, as an initial check of the data prior to any runs within the PMF model.

The signal-to-noise ratio represents the relationship of a certain portion of the concentration that exceeds the uncertainty. The signal-to-noise ratio is a vital marker which states which species ought to be included in the model run.[23] There are three signal-to-noise indicators in the PMF model. Species with a signal-to-noise ratio below 0.4 are marked as bad elements and excluded in the model run. The species with a signal-to-noise ratio between 0.4 ~1 are marked as weak elements and species with a signal-to-noise ratio above 1 are marked as strong elements.[67] However, in a similar study, the weak species ranged from 0.2 ~1.[48] The $PM_{2.5}$ species is set to be a “Total Variable” and assigned a higher uncertainty, and this was done for all four years. By default, the species is labelled as ‘weak’, and this will prevent it from largely influencing the outcome of the model.[67] Different factor models were run using five, six, and seven factors, which were determined from previous studies.[48-49,68]

Each base model was then run through bootstrap execution which is an error estimate, each bootstrap run was set at 100 runs for each model. The results were examined to determine if the species in each base model had values outside the interquartile ranges around the profiles.[64] The number of factors were considered appropriate if the bootstrap mapping was over 80%, which indicated that the uncertainties could be interpreted.[64]

3.3.2. UNSUPERVISED MACHINE LEARNING CLUSTER ANALYSIS

Two unsupervised ML clustering methods were used namely k-means and spectral clustering. Principal Component Analysis (PCA) was also conducted as an additional comparison as other studies have applied it in source apportionment.[35,58,69-70] K-means, spectral clustering, and PCA were all run in R statistical programme. The inbuilt R command 'prcomp' and package 'ggplot2' were used in the PCA analysis. Similar to the other clustering methods the data were scaled prior to PCA analysis. The packages used for the k-means and spectral clustering can be seen in section 3.2.6.5.

The determined number of optimal clusters were used in the clustering models, however, due to the possible number of sources found in the study area, the number of optimal clusters were increased in this portion of the study. Spectral clustering has been used for source apportionment in a study by Kumar et al.[43] Similar to the k-means clustering, the descriptive and statistics for each cluster were provided to possibly determine the source of each cluster. Then the proportions of each species in a cluster were used to determine the main sources of $PM_{2.5}$ during April 2017 to February 2021.

Non-parametric tests such as Wilcoxon rank-sum test and Kruskal Wallis tests were performed to determine any significant differences between or among the formed clusters in each model.

3.4. ETHICS APPROVAL

Ethical approval was only granted after a successful oral defence in front of the Academic Advisory committee (AAC), which occurred on 10 February 2021 (Appendix 4). After reviewer input, the title of the project changed and was returned for approval from the AAC (Appendix 5).

First, ethical approval (Reference No: 433/2021) was obtained from the Research Ethics Committee of Faculty of Health Sciences at the University of Pretoria, 26 August 2021 (see Appendix 6). After a change of study population for the first objective a second ethics approval was granted (Appendix 7). Toward the end of the project, only unsupervised ML methods were used and so the title of the project was amended and final ethics approval was granted (Appendix 8).

All identifying information of participants is omitted in the reporting of this study.

3.5. REFERENCES

1. Mehta N, Harish V, Bilimoria K, Morgado F, Ginsburg S, Law M, et al. Knowledge and attitudes on artificial intelligence in healthcare: A provincial survey study of medical students. medRxiv. 2021; 10(1) doi:10.15694/mep.2021.000075.1.

2. Ding L, Zhu D, Peng D, Zhao Y. Air pollution and asthma attacks in children: A case-crossover analysis in the city of Chongqing, China. Environmental Pollution. 2017; 220(Part A):348-53. doi:10.1016/j.envpol.2016.09.070.

3. Wichmann J. Heat effects of ambient apparent temperature on all-cause mortality in Cape Town, Durban and Johannesburg, South Africa: 2006-2010. Science of the Total Environment. 2017; 587-588:266-72. doi:10.1016/j.scitotenv.2017.02.135.

4. Maclure M. The case-crossover design: A method for studying transient effects on the risk of acute events. American journal of epidemiology. 1991; 185(11):1174-83. doi:10.1093/aje/kwx105.

5. Lombardi DA. The case-crossover study: A novel design in evaluating transient fatigue as a risk factor for road traffic accidents. Sleep. 2010; 33(3):283-4. doi:10.1093/sleep/33.3.283.

6. Carracedo-Martínez E, Taracido M, Tobias A, Saez M, Figueiras A. Case-crossover analysis of air pollution health effects: A systematic review of methodology and application. Environmental Health Perspectives. 2010; 118(8):1173-82. doi:10.1289/ehp.0901485.

7. Liu Y, Pan J, Zhang H, Shi C, Li G, Peng Z, et al. Short-term exposure to ambient air pollution and asthma mortality. American Journal of Respiratory and Critical Care Medicine. 2019; 200(1):24-32. doi:10.1164/rccm.201810-1823OC.

8. Jaakkola JJ. Case-crossover design in air pollution epidemiology. European Respiratory Journal. 2003; 40:81s-5s. doi:10.1183/09031936.03.00402703.

9. Maclure M. The case-crossover design: A method for studying transient effects on the risk of acute events. American journal of epidemiology. 1991; 133(2):144-53.

10. Liu H, Tian Y, Cao Y, Song J, Huang C, Xiang X, et al. Fine particulate air pollution and hospital admissions and readmissions for acute myocardial infarction in 26 Chinese cities. Chemosphere. 2018; 192:282-8.

11. Cox B, Gasparrini A, Catry B, Fierens F, Vangronsveld J, Nawrot TS. Ambient air pollution-related mortality in dairy cattle: Does it corroborate human findings? Epidemiology (Cambridge, Mass.). 2016; 27(6):779.

12. Olutola B. Influence of air pollution on the respiratory health of residents in the Vaal and Highveld Air Pollution Priority Areas of South Africa Pretoria: University of Pretoria; 2019.
13. World Health Organization. Classifications of diseases ICD-11. 2020. Available from: https://www.who.int/classifications/icd/en/.

14. South African Air Quality Information System. 2022. Available from: http://saaqis.environment.gov.za/.

15. Adebayo-Ojo TC, Wichmann J, Arowosegbe OO, Probst-Hensch N, Schindler C, Künzli N. Short-term joint effects of PM10, NO2 and SO2 on cardio-respiratory disease hospital admissions in Cape Town, South Africa. International Journal of Environmental Research and Public Health. 2022; 19(1):495.

16. Olutola B, Wichmann J. Does apparent temperature modify the effects of air pollution on respiratory disease hospital admissions in an industrial area of South Africa? Clean Air Journal. 2021; 31(2):1-11. doi:10.17159/caj/2021/31/2.11366.

17. Shirinde J, Wichmann J. Temperature modifies the association between air pollution and respiratory disease mortality in Cape Town, South Africa. International Journal of Environmental Health Research. 2022: 17;1-10. doi: 10.1080/09603123.2022.2076813.

18. South African Weather Service. Vaal triangle priority network monthly report – October 2019. 2019 AQI-VTPA-MER-2019-OCTOBER-001.

19. Lu P, Xia G, Zhao Q, Xu R, Li S, Guo Y. Temporal trends of the association between ambient temperature and hospitalisations for cardiovascular diseases in Queensland, Australia from 1995 to 2016: A time-stratified case-crossover study. PLoS medicine. 2020; 17(7):e1003176. doi:10.1371/journal.pmed.1003176.

20. Mwase N, Olutola B, Wichmann J. Temperature modifies the association between air pollution and respiratory disease hospital admissions in an industrial area of South Africa: The Vaal Triangle Air Pollution Priority Area. Clean Air Journal. 2022; 32(2) doi:10.17159.caj.2022.32.2.14588.

21. Hadeed SJ, O'Rourke MK, Burgess JL, Harris RB, Canales RA. Imputation methods for addressing missing data in short-term monitoring of air pollutants. Science of the Total Environment. 2020; 730 doi:10.1016/j.scitotenv.2020.139140.

22. Junger WL, De Leon AP. Imputation of missing data in time series for air pollutants. Atmospheric Environment. 2015; 102:96-104. doi:10.1016/j.atmosenv.2014.11.049.

23. Gass K, Klein M, Chang HH, Flanders WD, Strickland MJ. Classification and regression trees for epidemiologic research: An air pollution example. Environmental Health. 2014; 13(1):1-10. doi:10.1186/1476-069X-13-17.

24. Alfeus Anna. Air pollution source apportionment and joint effects on mortality in Cape Town, South Africa: University of Pretoria; 2021.

25. Gass K, Klein M, Sarnat SE, Winquist A, Darrow LA, Flanders WD, et al. Associations between ambient air pollutant mixtures and pediatric asthma emergency department visits in three cities: A classification and regression tree approach. Environmental Health. 2015; 14:58. doi:10.1186/s12940-015-0044-5.

26. Xu M YW, Tong S, Jia L, Liang F, Pan X. Non-linear association between exposure to ambient temperature and children's hand-foot-and-mouth disease in Beijing, China. Plos One. 2015; 10(5):e0126171.

27. Song X, Wang S, Hu Y, Yue M, Zhang T, Liu Y, et al. Impact of ambient temperature on morbidity and mortality: An overview of reviews. Science of the Total Environment. 2017; 586:241-54.

28. Wichmann J. Heat effects of ambient apparent temperature on all-cause mortality in Cape Town, Durban and Johannesburg, South Africa: 2006–2010. Science of the Total Environment. 2017; 587:266-72.

29. Lokotola CL, Wright CY, Wichmann J. Temperature as a modifier of the effects of air pollution on cardiovascular disease hospital admissions in Cape Town, South Africa. Environmental Science and Pollution Research. 2020; 27:16677-85 doi:10.1007/s11356-020-07938-7.

30. Wichmann J, Voyi K. Ambient air pollution exposure and respiratory, cardiovascular and cerebrovascular mortality in Cape Town, South Africa: 2001-2006. International Journal of Environmental Research and Public Health. 2012; 9(11):3978-4016. doi:10.3390/ijerph9113978.

31. Nhung NTT, Amini H, Schindler C, Joss MK, Dien TM, Probst-Hensch N, et al. Short-term association between ambient air pollution and pneumonia in children: A systematic review and meta-analysis of time-series and case-crossover studies. Environmental pollution. 2017; 230:1000-8.

32. Zhang S, Li G, Tian L, Guo Q, Pan X. Short-term exposure to air pollution and morbidity of COPD and asthma in east Asian area: A systematic review and meta-analysis. Environmental research. 2016; 148:15-23.

33. Zhang Z, Wang J, Lu W. Exposure to nitrogen dioxide and chronic obstructive pulmonary disease (COPD) in adults: A systematic review and meta-analysis. Environmental Science and Pollution Research. 2018; 25(15):15133-45.

34. Zhang Y, Wang SG, Xia Y, Shang KZ, Cheng YF, Xu L, et al. Association between ambient air pollution and hospital emergency admissions for respiratory and cardiovascular diseases in Beijing: A time series study. Biomedical and Environmental Sciences. 2015; 28(5):352-63.

35. Malik A, Tuckfield B Applied unsupervised learning with R : Uncover hidden relationships and patterns with k-means clustering, hierarchical clustering, and PCA. Birmingham: Packt Publishing Ltd. 2019.

36. Chan AK, Wozny TA, Bisson EF, Pennicooke BH, Bydon M, Glassman SD, et al. Classifying patients operated for spondylolisthesis: A k-means clustering analysis of clinical presentation phenotypes. Neurosurgery. 2021; 89(6):1033-41. doi:10.1093/neuros/nyab355.

37. Timmerman ME, Ceulemans E, De Roover K, Van Leeuwen K. Subspace k-means clustering. Behavior Research Methods. 2013; 45(4):1011-23. doi:10.3758/s13428-013-0329-y.

38. Govender P, Sivakumar V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980-2019). Atmospheric Pollution Research. 2020; 11(1):40-56. doi:10.1016/j.apr.2019.09.009.

39. Jain AK, Murty MN, Flynn PJ. Data clustering a review. ACM Computing Surveys. 1999; 31(3):264-323. doi:10.1145/331499.331504.

40. Pollard D. Strong consistency of k-means clustering. The Annals of Statistics. 1981; 9(1):135-40.

41. Long Z-Z, Xu G, Du J, Zhu H, Yan T, Yu Y-F. Flexible subspace clustering: A joint feature selection and k-means clustering framework. Big Data Research. 2021; 23 doi:10.1016/j.bdr.2020.100170.

42. Riches NO, Gouripeddi R, Payan-Medina A, Facelli JC. K-means cluster analysis of cooperative effects of CO, NO2, O3, PM2.5, PM10, and SO2 on incidence of type 2 diabetes mellitus in the US. Environmental Research. 2022; 212(Part B) doi:10.1016/j.envres.2022.113259.

43. Kumar V, Sahu M, Biswas P. Source apportionment of particulate matter by application of machine learning clustering algorithms. Aerosol and Air Quality Research. 2022; 22(3):210240. doi:10.4209/aaqr.210240.

44. Morissette L, Chartier Sylvain. The k-means clustering technique: General considerations and implementation in mathematica. Tutorials in Quantitative Methods for Psychology. 9(1):15-24.

45. Goubko M, Veremyev A. Bilinear matrix equation characterizes Laplacian and distance matrices of weighted trees. Discrete Applied Mathematics. 2021; 305:1-9. doi:10.1016/j.dam.2021.08.025.

46. Chakraborty S, Nagwani NK, Dey L. Performance comparison of incremental k-means and incremental dbscan algorithms. arXiv preprint arXiv:1406.4751. 2014.

47. Khan K, Rehman SU, Aziz K, Fong S, Sarasvady S, editors. Dbscan: Past, present and future. The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014); 2014: IEEE.

48. Adeyemi A, Molnar P, Boman J, Wichmann J. Source apportionment of fine atmospheric particles using positive matrix factorization in Pretoria, South Africa. Environmental Monitoring and Assessment. 2021; 193(11):716. doi:10.1007/s10661-021-09483-3.

49. Howlett-Downing C, Boman J, Molnár P, Shirinde J, Wichmann J. PM2.5 chemical composition and geographical origin of air masses in Pretoria, South Africa. Water, Air, & Soil Pollution. 2022; 233(7) doi:10.1007/s11270-022-05746-y.

50. Mwase N. Human health risks of inhalable exposure to PM2.5 in Pretoria, South Africa University of Pretoria; 2019.

51. Johannesson S, Gustafson P, Molnar P, Barregard L, Sallsten G. Exposure to fine particles (PM2.5 and PM1) and black smoke in the general population: Personal, indoor, and outdoor levels. . J Expo Sci Environ Epidemiol. 2007; 177(7):613-24.

52. Pekkanen J, Timonen K, Tiittanen P, Vallius M, Lanki T, et.al. Ultra study manual and data book, ultra. Kuopio: Kuopio University Printing Office, 2000.

53. Götschi T, Oglesby L, Mathys P, Monn C, et.al. Comparison of black smoke and PM2.5 levels in indoor and outdoor environments of four European cities. Environmental Science & Technology. 2002; 36:1191-97.

54. Molnár P, Sallsten G. Contribution to PM(2.5) from domestic wood burning in a small community in Sweden. Environmental Science: Processes & Impacts. 2013; 15(4):833-8. doi:10.1039/c3em30864b.

55. Molnár P, Tang L, Sjöberg K, Wichmann J. Long-range transport clusters and positive matrix factorization source apportionment for investigating transboundary PM2.5 in Gothenburg, Sweden. Environmental Science: Processes & Impacts. 2017; 19(10):1270-7. doi:10.1039/c7em00122c.

56. Novela RJ, Gitari WM, Chikoore H, Molnar P, Mudzielwana R, Wichmann J. Chemical characterization of fine particulate matter, source apportionment and long-range transport clusters in Thohoyandou, South Africa. Clean Air Journal. 30(2):1-2. doi:10.17159/caj/2020/30/2.8735.

57. Paatero P. The multilinear engine - a table-driven, least squares program for solving multilinear problems, including the n-way parallel factor analysis model. Journal of Computational and Graphical Statistics. 1999; 8:854–88.

58. Belis CA, Karagulian F, Larsen BR, Hopke PK. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. Atmospheric Environment. 2013; 69.

59. Kim S, Kim TY, Yi SM, Heo J. Source apportionment of PM2.5 using positive matrix factorization (PMF) at a rural site in Korea. Journal of Environmental Management. 2018; 214:325-34. doi:10.1016/j.jenvman.2018.03.027.

60. Ogundele LT, Owoade OK, Olise FS, Hopke PK. Source identification and apportionment of PM2.5 and PM2.5−10 in iron and steel scrap smelting factory environment using PMF, PCFA and UNMIX receptor models. Environmental Monitoring and Assessment 2016; 188(10):1-21. doi:10.1007/s10661-016-5585-8.

61. Park J, Kim H, Kim Y, Heo J, Kim S-W, Jeon K, et al. Source apportionment of PM2.5 in seoul, South Korea and Beijing, China using dispersion normalized PMF. Science of the Total Environment. 2022; 833 doi:10.1016/j.scitotenv.2022.155056.

62. Tshehla C, Djolov G. Source profiling, source apportionment and cluster transport analysis to identify the sources of PM and the origin of air masses to an industrialised rural area in Limpopo. Clean Air Journal. 2018; 28 doi:10.17159/2410-972X/2018/v28n2a18.

63. Vestenius M, Hopke PK, Lehtipalo K, Petäjä T, Hakola H, Hellén H. Assessing volatile organic compound sources in a Boreal Forest using positive matrix factorization (PMF). Atmospheric Environment. 2021; 259 doi:10.1016/j.atmosenv.2021.118503.

64. US Environmental Protection Agency. Epa positive matrix factorization (PMF) 5.0 fundementals and user guide. 2014. Available from: https://www.epa.gov/air-research/positive-matrix-factorization-model-environmental-data-analyses.

65. Hopke PK. Review of receptor modeling methods for source apportionment. Journal of the Air & Waste Management Association. 2016; 66(3):237-59. doi:10.1080/10962247.2016.1140693.

66. Williams J, Petrik L, Wichmann J. PM2.5 chemical composition and geographical origin of air masses in Cape Town, South Africa. Air Quality, Atmosphere & Health. 2020; 14(3):431-42. doi:10.1007/s11869-020-00947-y.

67. Howlett-Downing C. The association between sources of air pollution and respiratory health in Pretoria, South Africa: University of Pretoria; 2022.

68. Chen M, Wang P, Chen Q, Wu J, Chen X. A clustering algorithm for sample data based on environmental pollution characteristics. Atmospheric Environment. 2015; 107:194-203. doi:10.1016/j.atmosenv.2015.02.042.

69. Demir S, Saral A, Ertürk F, Kuzu L. Combined use of principal component analysis (PCA) and chemical mass balance (CMB) for source identification and source apportionment in air pollution modeling studies. Water, Air, & Soil Pollution : An International Journal of Environmental Pollution. 2010; 212(1-4):429-39. doi:10.1007/s11270-010-0358-4

CHAPTER 4: KNOWLEDGE, ATTITUDES AND PERCEPTIONS OF POSTGRADUATE STUDENTS TOWARDS USE OF ARTIFICIAL INTELLIGENCE IN PUBLIC HEALTH SURVEY

This chapter presents the results of the survey conducted from 13 June to 19 June 2022, among Postgraduate Diploma in Public Health students at the School of Health Systems and Public Health (SHSPH) at the University of Pretoria (UP). The survey was to assess the knowledge, attitudes and perceptions of postgraduate diploma students towards use of artificial intelligence in public health. A manuscript was submitted on 19 May 2023, to Journal of Public Health, which is currently under review.

4.1. RESULTS

4.1.1. SECTION A: DEMOGRAPHICS

There were 758 responses that were received, 618 questionnaires were completed in full with no missing values (81.5% response rate). Most respondents (82%) were female, 17.9% were male, and less than 1% chose not to indicate their gender, (Table 4.1). Majority of respondents were younger than 40 years old. Most respondents (83.5%) did not have any computer science backgrounds. Less than 15% of respondents had attended or viewed AI related talks or lectures.

Table 4.1: Summary study demographics (N=618).

	n(%)		n (%)		n(%)
Gender		**Residence**		**Highest Qualification**	
Male	108(17.5)	Eastern Cape	54(8.7)	Bachelors	501(81.1)
Female	507(82)	Free State	16(2.6)	Masters	29(4.7)
		Gauteng	252(40.8)	Doctorate	5(0.8)
Age		Kwa-Zulu Natal	83(13.4)	Other	72(11.7)
20-24	12(1.9)	Limpopo	38(6.2)	Rather not say	11(1.8)
25-30	210(34)	Mpumalanga	48(7.8)	**Background in comp science**	
31-34	125(20.2)	Northern Cape	12(1.9)	Yes	81(13.1)
35-40	134(21.7)	North-West	35(5.7)	No	516(83.5)
41-45	67(10.8)	Western Cape	28(4.5)	Rather not say	21(3.4)
46-50	42(6.8)	Outside SA	45(7.3)	**Attended AI talks/lectures**	
51-55	17(2.8)	Rather not say	7(1.1)	Yes	92(14.9)
56-60	3(0.5)			No	508(82.2)
Rather not say	8(1.3)			Rather not say	18(2.9)
				Received training computer programming/coding	
				Yes	39(6.3)
				No	571(92.4)
				Rather not say	8(1.3)

n-number of participants per group

4.1.2. SECTION B: KNOWLEDGE OF ARTIFICIAL INTELLIGENCE TERMINOLOGY

Respondents' understanding of general terminology varied among the different terms. The majority (77.9%) of the respondents at least agreed that they understand what Artificial Intelligence is and 51.4% similarly agreed that they understand Machine Learning. Other terms such as Neural Networks, Deep Learning and Algorithms that are associated with/related to deeper knowledge of Artificial Intelligence, was unknown to most of the respondents (Table 4.2).

Table 4.2: General knowledge about AI terminology.

	Strongly disagree	Disagree	Agree	Strongly agree	Unsure
AI	5.2% (32/168)	7.1% (44/618)	56.5% (349/618)	21.8% (135/618)	9.4% (58/618)
ML	7.4% (46/618)	23.8% (147/618)	40.1% (248/618)	11.3% (70/216)	17.3% (107/618)
NN	18.8% (116/618)	38.8% (240/618)	19.3% (119/618)	4.2% (26/618)	18.9% (117/618)
DL	13.3% (82/618)	29.1% (180/618)	34% (210/618)	5.8% (36/618)	17.8% (110/618)
Algorithm	19.4% (120/618)	28.4% (174/618)	27.4% (169/618)	8.7% (54/618)	16.3% (101/618)

4.1.3. SECTIONC: PERCEPTIONS OF ARTIFICIAL INTELLIGENCE

4.1.3.1. PERCEPTIONS OF AI AND TASK PERFORMANCE

Table 4.3 shows how likely respondents thought AI could perform tasks at an individual patient care level. Respondents thought it at least likely that AI could perform administrative, diagnostic and prognostic task e.g. the vast majority thought AI would be likely to read and interpret diagnostic imaging (89.1%). However, respondents did not think it likely the AI would be able to perform tasks that involved direct care to patients. Only 28% of respondents thought AI could provide empathetic care and only 26% felt that it could perform psychiatric/personal counselling.

Table 4.3: The perceived ability of AI to eventually perform a specific task at individual health level.

	Extremely unlikely	Unlikely	Likely	Extremely likely	Unsure
Provide patients with preventative health recommendations (e.g. exercise, diet, wellness).	3.6% (22/618)	11.5% (71/618)	54.1% (334/618)	29.8% (184/618)	1.1% (7/618)
Analyse patient information to reach a diagnosis.	5.2% (32/618)	11.3% (70/618)	49.8% (308/618)	30.7% (190/618)	2.9% (18/618)
Analyse patient information to establish possible prognosis.	3.6% (22/618)	12.1% (75/618)	56.8% (351/618)	23.8% (147/618)	3.7% (23/618)
Read and interpret diagnostic imaging (such as X rays).	1.5% (9/618)	6.2% (38/618)	44.8% (277/618)	44.3% (274/618)	3.2% (20/618)
Evaluate when to refer patients to other health professionals.	5.2% (32/618)	17.5% (108/618)	52.8% (326/618)	20.1% (124/618)	4.5% (28/618)
Formulate personalised treatment plans for patients	5.2% (32/618)	18.1% (112/618)	51% (315/618)	21.5% (133/618)	4.2% (26/618)
Formulate personalised medication prescriptions for patients.	4.5% (28/618)	19.3% (119/618)	49.2% (304/618)	22.2% (137/618)	% (/618)
Provide empathetic care to patients.	33.5% (207/618)	33.3% (206/618)	19.6% (121/618)	8.4% (52/618)	5.2% (32/618)
Monitor patient compliance to prescribed medications, exercise and dietary recommendations.	10.5% (65/618)	19.26% (119/618)	46.4% (287/618)	19.6% (121/618)	4.2% (26/618)
Provide psychiatric/personal counselling.	31.9% (197/618)	35.8% (221/618)	18.5% (114/618)	8.7% (54/618)	5.2% (32/618)
Perform surgery (e.g. robotic surgery).	13.8% (85/618)	17.2% (106/618)	38.8% (240/618)	25.9% (160/618)	4.4% (27/618)

Table 4.4 shows respondents who thought it likely for AI to provide preventative health recommendations (31.1%) and read and interpret diagnostic imaging (29.7%), felt that this would be possible in the next 5 to 10 years. Overall, respondents thought AI performance of other tasks would be possible within 11 to 25 years.

Table 4.4: Expected time students perceived AI to eventually perform specific tasks at individual health level.

	N	0-4 yrs	5-10 yrs	11-25 yrs	26-50 yrs	≥ 50 yrs
				%		
Provide patients with preventative health recommendations (e.g. exercise, diet, wellness).	518	7.5	31.1	29.9	20.1	11.4
Analyse patient information to reach a diagnosis.	497	9.7	27.2	30.4	22.7	10.1
Analyse patient information to establish possible prognosis.	497	9.9	29.4	29.6	20.9	10.3
Read and interpret diagnostic imaging (such as X rays).	549	14.6	29.7	26.1	19.1	10.6
Evaluate when to refer patients to other health professionals.	448	11.4	28.1	32.1	17.6	10.7
Formulate personalised treatment plans for patients	446	11.2	26	33.2	20	9.6
Formulate personalised medication prescriptions for patients.	438	10.3	27.2	29.7	19.6	13.2
Provide empathetic care to patients.	173	12.1	26	30.1	24.9	6.9
Monitor patient compliance to prescribed medications, exercise and dietary recommendations.	406	12.8	25.1	25.4	23.2	13.6
Provide psychiatric/personal counselling.	166	11.5	18.1	31.3	27.7	11.5
Perform surgery (e.g. robotic surgery).	399	17.5	22.1	29.1	19.8	11.5

N- the total number of students that selected “likely” or “Extremely likely”, %- percentage of students from N

The majority of respondents thought it possible for AI to perform tasks at a health systems level (Table 4.5). Among those who thought it likely, approximately 29% responded that AI could be able to perform most of these tasks within the next 5 to 10 years (Table 4.6). Table 4.5 also shows that respondents thought it likely that AI could perform tasks at a population health level, with 61.9% thinking it likely that AI could conduct population health surveillance and outbreak prevention. However, of the

respondents who thought it likely for AI to perform tasks at a population level, approximately 30% thought that this could occur within 11 to 25 years (Table 4.6).

Table 4.5: The perceived ability of AI to eventually perform specific tasks at health systems, and population health levels.

	Extremely unlikely	Unlikely	Likely	Extremely likely	Unsure
			Health Systems		
Provide documentation (e.g., update medical records) about patients	3.1% (19/618)	7% (43/618)	46.1% (285/618)	41.1% (254/618)	2.8% (17/618)
Assist hospitals in capacity planning and human resource management	6.2% (38/618)	19.6% (121/618)	47.6% (294/618)	22.6% (140/618)	4.1% (25/618)
Provide recommendations for quality improvement in practices/hospitals	6% (37/618)	20.6% (127/618)	50% (309/618)	19.4% (120/618)	4.1% (25/618)
			Population health		
Conduct population health surveillance and outbreak prevention.	8.9% (53/618)	24.1% (149/618)	41.8% (258/618)	20.1% (124/618)	5.5% (34/618)
Select the best population health interventions.	7.1% (44/618)	27% (167/618)	47.4% (293/618)	13.1% (81/618)	5.3% (33/618)

Table 4.6: Expected time students perceived AI to eventually perform a specific task at health systems, and population health levels.

	N	0-4 yrs	5-10 yrs	11-25yrs	26-50yrs	≥50yrs
	%					
		Health Systems				
Provide documentation (e.g., update medical records) about patients	537	21.2	25.9	26.1	17.3	9.5
Assist hospitals in capacity planning and human resource management	429	18.7	29.4	26.6	17.5	7.9
Provide recommendations for quality improvement in practices/hospitals	426	18.1	29.1	26.3	18.5	8
		Population health				
Conduct population health surveillance and outbreak prevention.	379	15.6	27.7	28.8	18.7	9.2
Select the best population health interventions.	370	11.6	29.7	32.2	18.1	8.4

4.1.3.2. PERCEPTIONS OF AI AND IMPACT ON PUBLIC HEALTH CAREERS

Table 4.7 shows respondents' perceptions about AI and the impact on public health careers. Approximately 77% of respondents had either strongly agreed or agreed that AI would reduce job availability to them. Majority of respondents (69%) strongly agreed or agreed that AI would reduce the number of jobs in the public health field. While 24% of the students disagreed that AI would reduce job availability in the public health field. Finally, 40% of respondents agreed or strongly agreed that AI has or will impact their choice of public health speciality.

Table 4.7: Perceived impact of AI on public health careers.

	Strongly agree	Agree	Disagree	Strongly disagree	Unsure
Artificial Intelligence will reduce the number of jobs available to me.	39.5% (244/618)	37.5% (232/618)	14.7% (91/618)	4.4% (27/168)	3.9% (24/618)
Artificial Intelligence will reduce the number of jobs in certain public health	30.1% (186/216)	39.3% (243/618)	20.4% (126/618)	4.2% (26/618)	6% (37/618)
Artificial Intelligence will/already did impact my choice of public health specialty selection.	12% (74/618)	28.2% (174/618)	34.6% (214/618)	11.7% (72/618)	13.6% (84/618)

4.1.3.3. PERCEPTIONS OF AI AND ETHICS

Table 4.8 shows respondents' strongly agreed or agreed that AI in public health would introduce new ethical (84.9%), social (77.4%) and health equity (77.2%) challenges. When asked if respondents thought the South African healthcare system was currently well prepared to deal with challenges related to AI, 78.6% disagreed or strongly disagreed.

Table 4.8: Perceived ethical challenges from AI.

	Strongly agree	Agree	Disagree	Strongly disagree	Unsure
AI in public health will raise new ethical challenges.	42.2% (261/168)	42.7% (264/618)	6.2% (38/618)	3.6% (22/618)	5.3% (33/618)
AI in public health will raise new social challenges.	35% (216/618)	42.4% (262/618)	10% (62/618)	4.2% (26/216)	5.2% (8.4/618)
AI in public health will raise new challenges around health equity.	36.4% (225/618)	40.8% (252/618)	9.6% (59/618)	3.6% (22/618)	9.7% (60/618)
The South African healthcare system is currently well prepared to deal with challenges related to AI.	3.1% (19/618)	6.3% (39/618)	28.6% (177/618)	50% (309/618)	12% (74/618)

4.1.3.4. PERCEPTIONS OF AI AND PUBLIC HEALTH EDUCATION

Table 4.9 shows respondents' perception about AI and public health education. A small majority (52.9%) of respondents agreed that their current public health education was adequately preparing them to work alongside AI and there was very strong

support (91.4%) for the inclusion of AI in public health training (Table 7). This inclusion could start at an undergraduate level (76 %).

Table 4.9: Integration of AI into public health education.

	Strongly agree	**Agree**	**Disagree**	**Strongly disagree**	**Unsure**
My public health education is adequately preparing me for working alongside AI tools	13.6% (84/618)	39.3% (243/618)	22.8% (141/618)	6.5% (40/618)	17.8% (110/618)
My public health training should include training on AI competencies (e.g. what is AI, how will it impact us, what are the challenges it raises).	41.9% (259/618)	49.5% (306/618)	3.2% (20/618)	1.6% (10/216)	3.7% (23/618)
Every public health student should be required to receive training in AI competencies.	39.2% (242/618)	47.4% (293/618)	6.5% (40/618)	2.1% (13/618)	4.9% (30/618)
	Under-graduate	**Post-graduate**	**Not necessary**	**unsure**	
Training in AI competencies should begin as a:	76.5% (473/618)	15.1% (93/618)	1.8% (11/618)	6.6% (41/618)	

An analysis of the open-ended responses was done. These responses were non-compulsory open-ended questions on whether they had any comments or concerns about AI in public health.

There were 327 free text responses to the question “Do you have any comments on the topic of AI in public health?” Examples of some of the responses are highlighted in Table 4.10.

Table 4.10: Themes of comments or concerns about AI in public health from respondents.

Theme	N (%)	Example
Positive comments	120 (36.7)	"AI in public health can assist in data management which plays a huge role in public health." "In terms of waiting times in healthcare institutions it will help to reduce and intervene on the current staff shortages."
Positive comments with some reservations	54 (16.5)	"AI is a very controversial topic but I believe that it could be beneficial to some aspect of public health such as surveillance. " "AI may be beneficial to reduce workload in the healthcare system for example reduce queues for medication, the bad side of it is that it will reduce job opportunities and people will end up unemployed which will cripple the economy of the country"
Negative comments	128 (39.1)	"Already there is a problem of huge unemployment is South Africa and the economy is getting worse daily and artificial intelligence will require money to be well introduced and to deal with the challenges it will come with. Our government healthcare facilities are poorly functional to a point where they cannot even feed their own patient so educating their staff members is gonna be a real problem as it is now." "My concern is the security gaps that comes with the use of AI, There need to be competent IT specialists and security expects who work in public health space to provide competency for healthcare workers. The AI will help the public health facilities to move from the paper system to the digital systems that is easily accessible."
Neutral	25 (7.7)	"AI is somehow around and in utilisation in several sectors and already have impact. I can only say its influence is increasing in our daily lives and professions." "I am yet to read on AI as it is a new concept to me"

Respondents were also asked to give a general reflection on what AI within their department in the next five years. There were 418 responses. Some examples of the responses are highlighted in Table 4.11.

Table 4.11: Themes of reflection on what AI will look like in 5 years within their department.

Theme	N (%)	Example
Beneficial impact	192, (45.9)	"In 5yrs from now it means clients will be able to just get their parcels without coming to the clinic, getting their prescriptions and repeated medications at pickup points." "In my field we already have automated systems, with their further development the pathologist will be able to run tests, resulting in low numbers of medical technologists employed"
Detrimental impact	58, (13.9)	"It will take away a lot of jobs" "AI will be a disaster in public health sector as we have observed during Covid-19 pandemic in South Africa. Our system is way far behind in preparing for AI."
No change	76 (18.2)	"It will be nowhere in the periphery where I work" "I'm still not sure if they'll be able to develop anything close enough to bedside patient care. So it will be far in development"
Change not possible	56 (13.4)	"I do not think it will have been introduced in the healthcare industry in this country" "I do not see it happening due to money constraints, equipment's to run AI needs money that our government don't have now and its becoming worse every year"
Neutral	36 (8.6)	"I am uncertain of what anything will look like in 5 years including AI. Technology is being created to make life easier for us as human beings. But everything comes at a price at a cost. Nothing is ever neutral, each gift brings pros and cons. So will see what AI will bring." "It is too complicated to comprehend"

4.1.4. DEMOGRAPHICS' CORRELATION WITH PERCETIONS

The female respondents' showed better understanding of terms such as ML ($p=0.0003$), NN ($p=0.01$) and an algorithm ($p=0.01$) than the male respondents. Respondents' who had a background in computer sciences showed better understanding of ML ($p=0.0003$), NN ($p=0.01$), DL ($p=0.001$) and an algorithm ($p=0.01$). Respondents who had not attended AI talks or lectures showed less understanding of all AI terminology ($p<0.05$). Generally demographic factors such as age, gender and highest qualification, did not significantly influence respondents perception on whether AI could eventually perform a specific task at an individual level ($p>0.05$). However in some instances respondents who had a background in computer science and attended AI talks/lectures thought AI was more likely to perform some individual tasks. Demographic factors did not significantly influence the respondents'

perceptions on whether AI could eventually perform a specific task at health systems or population health levels ($p>0.05$).

Similarly, the demographic factors did not significantly influence the way respondents' perceived AI impact on public health careers or ethical challenges from AI ($p>0.05$). There was one exception from the latter where all demographic factors showed to influence respondents' perceptions on whether 'The South African healthcare system is currently well prepared to deal with challenges related to AI ($p<0.05$). Respondents who had a background in computer science ($p=0.04$) and attended AI talk/lectures ($p=0.01$) felt that their public health education was not adequately preparing them for working alongside AI tools. However, these demographic factors did not significantly influence the perceptions of other questions asked in this section.

4.2. DISCUSSION

The aim of this study was to assess the attitudes and perceptions regarding AI among public health postgraduate diploma at University of Pretoria, South Africa. The recent interest in AI and its different applications has gained traction and research into the topic is increasing. However, there are still on-going misconceptions and attitudes towards the concept of AI that still need to be addressed. [1-2] Understanding how AI in public health is perceived and engaging with university students to explore how they understand AI can inform the integration of AI in higher education within South Africa.

The survey respondents were mainly female students. In some health studies, there is often a slightly larger female population involved [3-4] unless the study is male specific [5] otherwise the gender balance is often dependent on the country and career of interest in which the study is taking place.[6] The higher proportion of females in this study could be considered a confounding factor when gender was shown to significantly influence certain perceptions. The majority of the respondents were in the 25 to 40 year old age group, which is a common age group of online postgraduate students.[4-5,7] Similar to the Mehta et al. (2021)[8] study, the majority of the respondents in this study did not have any background in computers or had been exposed to AI lectures or talks. The respondents' lack of exposure to AI lectures or talks is evident in their responses regarding familiarity with AI terminology. Their lack of familiarity with terms such as deep learning and neural networks is similar to what was reported in a Canadian cohort.[8]

Respondents were confident that AI would be able to carry out many tasks at different levels of public health. These identified tasks were tasks that could not provide direct care to patients, such as empathetic care or counselling. Respondents felt that administrative, diagnostic and prognosis tasks were better suited for AI advancements in public health. Mehta et al.,[8] also found that their participants were more confident the AI would be able to perform 'objective' tasks such as diagnosis, prognosis, interpreting imaging and formulating prescriptions as compared to tasks requiring more person-centred skills, personal counselling and providing empathetic care. The findings in this study found that gender, age, level of qualification, a background in computer science and the attendance of AI talks or lectures, did not show to have a significant impact on the perceptions made by respondents. However, Stai et al., [3] found that age made a significant difference in participants' perception of AI's ability to perform certain tasks such as surgery. A meta-analysis done by Hauk, Hüffmeier and Krumm,[9] found it was a misconception that younger generations would be more likely to engage in advancing technology, but rather technology is better taken up by how easy it is to use rather than the age of the individual. The finding of this study seems to echo the latter sentiment as age did not impact the respondents' perceptions of AIs' ability to perform certain tasks in public health.

Some issues were raised about AI affecting job availability at an individual and general level in public health. The general perception was that AI would increase unemployment. In similar studies conducted in Canada, Spain and Turkey, the perception that the use of AI would negatively affect the number of available jobs was also recorded as a concern.[6,8,10] Unemployment in South Africa is a general concern as the current unemployment rate is close to 30%.[11] This high unemployment rate could be the reason why respondents focused on this aspect. This high unemployment rate could be the reason why respondents focused on this aspect. The concern about job availability after the introduction of AI in public health could also be a result of fear of possible redundancy of certain jobs in public health. Because AI can run certain repetitive tasks and leave the more challenging tasks for healthcare professionals, there is a concern that this would then change or reduce their roles in the workplace [12-13] In spite of the positive benefits AI in healthcare and public health may promote, the concern of job replacement and job loss is a prominent issue. Another reason why perceptions of AI may be taken negatively, could revert back to the misconceptions

around the topic. AI can be thought to be a type of technology that allows computers to think, but that is not the case. [1,14] While AI is a part of computer science, the findings showed that having a background in computer science or attending AI talks or lectures did not influence the perceptions of the effects of AI in public health careers.

A rise in ethical challenges was also highlighted as an issue that could result from introducing AI into public health. Respondents perceived that the introduction of AI would raise ethical, social and health equity issues. Concerns about hacking of private information was also raised. Patient and hospital information is a very sensitive matter in health and there is already a difficulty in acquiring such data.[15] There is therefore a legitimate concern about data being shared and liability should patient information be wrongfully accessed. [15-16] Although this was not highlighted in the study findings it is an interesting factor to consider when assessing possible ethical issues around AI in public health.

Health equity was another concern highlighted by the respondents, specifically concerning how AI implementation could increase the gap between lower- and upper-income groups. This concern is valid due to the current, severe socio-economic disparities in the country.[17-20] In contrast to this concern, some research has found that the use of AI technologies is optimistic and could be a means for developing nations' ability to address their health disparities.[21-24] Respondents overwhelmingly agreed that the South African healthcare system was currently not prepared to handle AI-related challenges. The health department has been reported to have mismanaged funds and experiences a shortage of healthcare facilities and staff to meet the demand of the population.[25] These shortcomings were highlighted during the recent Covid-19 pandemic that further strained the healthcare system.[26-27] The manner in which the pandemic was handled could be a reason why a large proportion of respondents do not feel confident that AI related challenges would be dealt with by the current healthcare system.

Gender, age, having a background in computer science, attending AI talk or lecture, showed a significant association with the ethical challenges of AI and the country's readiness to address these challenges that may arise. Similar studies did not look at the associations of the demographics and AI and ethics. [3,6,8] However some did find that in general, female respondents' had more positive perceptions towards AI in

dentistry[6] and the publics' opinion of AI in surgery.[3] Although similar to this project's findings, the distribution of females to males was not proportional.

The majority of respondents agreed that being introduced to AI competencies should be done early in undergraduate training. Due to the 'newness' of the topic of AI, people may be very unsure about it but may be willing to learn more.[6,8] Although with current misconceptions, people may have no interest at all to learn more on the topic.[10] A lot of effort needs to be done to restructure the misconceptions on AI, reducing the popularity and marketing definition of AI and inform its useful scientific capabilities.[1,28] Whilst AI may be perceived as a complex concept, curriculum content for AI in public health and medicine can be structured around the basic understanding of AI concepts, limitations, and relevant ethical and legal implications.[8,29-31] This educational strategy could better prepare public health students to accept and understand AI.

The main strength to the study is that it is the first study conducted at a South Africa university to address the subject of AI in public health. The study aligns with one of the nine strategic interventions of the National Digital Health Strategy for South Africa (2019-2024) namely "to develop enhanced digital health technical capacity and skilled workforce for digital technology support and implementation."[32] The findings of this study provide baseline information as a foundation for similar studies to be done at other South Africa tertiary institutions. A study limitation is that this study explored the topic among one group of students at one university and so the findings are not generalizable.

The study provides an argument for an introductory AI course in undergraduate and/or postgraduate health professions and public health. This introduction can encourage further AI research by public health professionals in the country as there are no current courses offered within public health curricula that offer AI and ML. However, the introduction of AI in public health education is limited by the lack of expertise on the subject. Although AI in computer science is more available, constructing a more 'user friendly' curriculum for public health students will still need more time to implement and need a balanced contribution from both computer science and public health to make it fit for purpose. It is therefore recommended that this is an area for future joint curriculum development and research collaboration that needs exploration.

4.3. CONCLUSION

In conclusion, this survey brought out a variety of views shared among future public health professionals. Although there was a general assumption about AI entering public health and performing particular tasks at different health levels, there was a general consensus that AI had the potential to increase unemployment and ethical challenges in the field. This study does create a baseline for more extensive in-depth studies to be done within an African context. It is recommended that further studies be done that include participants from more - and different - programmes within health faculties in different settings. These studies could better inform the possible introduction of AI into undergraduate or postgraduate health professional programmes.

4.4. REFERENCES

1. Emmert-Streib F, Yli-Harja O, Dehmer M. Artificial intelligence: A clarification of misconceptions, myths and desired status. Frontiers in Artificial Intelligence. 2020; 3:524339. doi:10.3389/frai.2020.52433.
9
2. Helm JM, Swiergosz AM, Haeberle HS, Karnuta JM, Schaffer JL, Krebs VE, et al. Machine learning and artificial intelligence: Definitions, applications, and future directions. Current Reviews in Musculoskeletal Medicine. 2020; 13(1):69-76. doi:10.1007/s12178-020-09600-8.

3. Stai B, Heller N, McSweeney S, Rickman J, Blake P, Vasdev R, et al. Public perceptions of artificial intelligence and robotics in medicine. Journal of Endourology. 2020; 34(10):1041-8. doi:10.1089/end.2020.0137.

4. Tiyuri A, Saberi B, Miri M, Shahrestanaki E, Bayat BB, Salehiniya H. Research self-efficacy and its relationship with academic performance in postgraduate students of tehran university of medical sciences in 2016. Journal of Education and Health Promotion. 2018; 7 doi:10.4103/jehp.jehp_43_17.

5. Dada SO, Oyewole OE, Desmennu AT. Knowledge as determinant of healthy-eating among male postgraduate public health students in a Nigerian tertiary institution. International Quarterly of Community Health Education. 2020; doi:10.1177/0272684X20972895.

6. Yüzbaşıoğlu E. Attitudes and perceptions of dental students towards artificial intelligence. Journal of Dental Education. 2021; 85(1):60-8. doi:10.1002/jdd.12385.

7. Fernandez RS, Tran DT, Ramjan L, Ho C, Gill B. Comparison of four teaching methods on evidence-based practice skills of postgraduate nursing students. Nurse Education Today. 2014; 34(1):61-6. doi:10.1016/j.nedt.2012.10.005.

8. Mehta N, Harish V, Bilimoria K, Morgado F, Ginsburg S, Law M, et al. Knowledge and attitudes on artificial intelligence in healthcare: A provincial survey study of medical students. medRxiv. 2021; 10(1) doi:10.15694/mep.2021.000075.1.

9. Hauk N, Hüffmeier, J., & Krumm, S. Ready to be a silver surfer? A meta-analysison the relationship between chronological age and technology acceptance. Computers in Human Behavior. 2018; 84,:304-19.

10. Albarrán Lozano I, Molina JM, Gijón C. Perception of artificial intelligence in Spain. Telematics and Informatics. 2021; 63 doi:10.1016/j.tele.2021.101672.

11. Marire J. Relationship between fiscal deficits and unemployment in South Africa. Journal of Economic and Financial Sciences. 2022; 15(1) doi:10.4102/jef.v15i1.693.

12. Chen Y, Stavropoulou C, Narasinkan R, Baker A, Scarbrough H. Professionals' responses to the introduction of AI innovations in radiology and their implications for future adoption: A qualitative study. BMC Health Services Research. 2021; 21(1):1-9.

13. Tursunbayeva A, Renkema M. Artificial intelligence in health-care: Implications for the job design of healthcare professionals. Asia Pacific Journal of Human Resources. n/a(n/a) doi:https://doi.org/10.1111/1744-7941.12325.

14. Liu TYA, Bressler NM. Controversies in artificial intelligence. Current opinion in ophthalmology. 2020; 31(5):324-8. doi:10.1097/icu.0000000000000694.

15. Reisman M. EHRs: The challenge of making electronic data usable and interoperable. Pharmacy and Therapeutics. 2017; 42(9):572-5.

16. Petersson L, Larsson I, Nygren JM, Nilsen P, Neher M, Reed JE, et al. Challenges to implementing artificial intelligence in healthcare: A qualitative interview study with healthcare leaders in Sweden. BMC Health Services Research. 2022; 22(1):850. doi:10.1186/s12913-022-08215-8.

17. Gordon T, Booysen F, Mbonigaba J. Socio-economic inequalities in the multiple dimensions of access to healthcare: The case of South Africa. BMC Public Health. 2020; 20(1):1-13.

18. Mutyambizi C, Booysen F, Stokes A, Pavlova M, Groot W. Lifestyle and socio-economic inequalities in diabetes prevalence in South Africa: A decomposition analysis. PloS One. 2019; 14(1):e0211208.

19. Rispel L. Analysing the progress and fault lines of health sector transformation in South Africa. South African Health Review. 2016; 2016(1):17-23.

20. Wilson F. Historical roots of inequality in South Africa. Economic History of Developing Regions. 2011; 26(1):1-15.

21. Akpanudo S. Application of artificial intelligence systems to improve healthcare delivery in Africa. Primary Health Care. 2022; 12(1):1-4.

22. Marcus JL, Sewell, W.C., Balzer, L.B., & Krakower, D.S. . Artificial intelligence and machine learning for HIV prevention: Emerging approaches to ending the epidemic. Current HIV/AIDS Reports. 2020; 17 (3):171-9.

23. Owoyemi A, Owoyemi J, Osiyemi A, Boyd A. Artificial intelligence for healthcare in Africa. Frontier Digital Health. 2020; 2:6. doi:10.3389/fdgth.2020.00006.

24. Vinuesa R, Azizpour H, Leite I, Balaam M, Dignum V, Domisch S, et al. The role of artificial intelligence in achieving the sustainable development goals. Nature Communications. 2020; 11(1):233. doi:10.1038/s41467-019-14108-y.

25. Aikman N. The crisis within the South African healthcare system : A multifactorial disorder. SAJBL. 2019; 12(2):52-6. doi:10.7196/SAJBL.2019.v12i2.694.

26. Mokhele T, Manyaapelo T, Sifunda S, Dukhi N, Sewpaul R, Naidoo I, et al. Factors influencing healthcare workers' perception of South African health system capability for managing COVID-19 pandemic. The Open Public Health Journal. 2022; 15(1) doi:10.2174/18749445-v15-e2204070.

27. Taylor A, Feuvre DL, Taylor B. COVID-19: The South African experience. Interventional Neurology. 2021; 27(1_suppl):50-3. doi:10.1177/15910199211035905.

28. Cukurova M, Luckin R, Kent C. Impact of an artificial intelligence research frame on the perceived credibility of educational research evidence. International Journal of Artificial Intelligence in Education. 2020; 30(2):205-35. doi:10.1007/s40593-019-00188-w.

29. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthcare Journal. 2019; 6(2):94-8. doi:10.7861/futurehosp.6-2-94.

30. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digital Medicine. 2018; 1 doi:10.1038/s41746-018-0061-1.

31. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digital Medicine. 2020; 3(1) doi:10.1038/s41746-020-0294-7.

32. Department of Health. National digital health strategy for South Africa (2019-2024). Pretoria, South Africa. 2019.

CHAPTER 5: IMPUTATION OF AIR POLLUTION DATA

This chapter summarises the imputation methods used on meteorological data (relative humidity, temperature, and wind speed) and air quality data (SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} pollutants).

5.1. RESULTS

5.1.1. METEOROLOGICAL DATA

Relative humidity (%), temperature (°C) and wind speed (m/s) data for six stations located in the Vaal Triangle Airshed Priority Area (VTAPA) South Africa, were used. Table 5.1 show descriptive statistics and proportion on missing data of the meteorological conditions before imputation. Proportion of missing data ranged between ~10% and ~22%. Table 5.2 shows descriptive statistics after mice imputation.

.

Table 5.1: Descriptive statistics for the six monitoring stations in the VTAPA on relative humidity, temperature and wind speed, before imputation.

	RH	Temperature	Wind speed	RH	Temperature	Wind speed
	Diepkloof			Sharpeville		
Mean	45.09	17.16	2.71	48.75	17.90	2.65
Minimum	1.2	0.57	0.12	9.37	3.66	0.36
1st quartile	32.89	13.7	2.08	37.93	14.22	1.87
Median	46.01	17.46	2.66	49.05	18.725	2.52
3rd quartile	57.91	20.86	3.28	59.67	21.71	3.26
Maximum	85.82	32.48	5.94	97.1	31.05	6.55
NA	352	378	444	484	491	477
%	10.52	11.29	13.27	14.46	14.67	14.25
	Kliprivier			Three Rivers		
Mean	52.77	16.45	2.23	51.43	17.32	2.30
Minimum	4.75	-1.45	0.28	10.82	1.21	0.38
1st quartile	42.76	12.14	1.48	41.37	13.25	1.61
Median	54.03	17.55	1.90	51.79	18.33	2.14
3rd quartile	63.32	20.77	2.49	61.57	21.53	2.84
Maximum	96.33	34.49	14.00	99.03	29.5	6.61
NA	514	510	449	626	629	615
%	15.36	15.24	13.41	18.70	18.79	18.37
	Sebokeng			Zamdela		
Mean	49.06	17.10	2.61	50.54	15.86	2.26
Minimum	8.67	-1.13	0.26	10.35	-1.45	0.12
1st quartile	37.40	13.24	1.94	40.34	11.64	1.51
Median	49.72	17.46	2.50	50.84	16.43	2.13
3rd quartile	60.87	21.24	3.20	61.12	20.31	2.91
Maximum	87.77	30.39	6.65	89.83	30.04	5.88
NA	688	680	650	365	363	748
%	20.56	20.32	19.42	10.91	10.85	22.35

NA-number of missing data, %- percentage of missingness

Table 5.2: Descriptive statistics for the six monitoring stations in the VTAPA on relative humidity, temperature and wind speed, after imputation.

	RH	Temperature	Wind speed	RH	Temperature	Wind speed
	Diepkloof			Sharpeville		
Mean	45.06	17.16	2.71	48.78	17.89	2.65
Minimum	1.20	0.57	0.12	9.37	3.66	0.36
1st quartile	34.61	14.22	2.17	39.84	14.98	1.98
Median	45.56	17.37	2.67	48.95	18.34	2.57
3rd quartile	56.47	20.52	3.20	58.47	21.26	3.17
Maximum	85.82	32.48	5.94	97.10	31.05	6.55
	Kliprivier			Three Rivers		
Mean	52.78	16.50	2.23	51.27	17.33	2.30
Minimum	4.75	-1.45	0.28	10.82	1.21	0.38
1st quartile	44.87	12.96	1.54	43.75	14.42	1.72
Median	53.74	17.14	1.96	51.28	17.95	2.19
3rd quartile	61.78	20.37	2.50	59.18	21.03	2.70
Maximum	96.33	34.49	14.00	99.03	29.50	6.61
	Sebokeng			Zamdela		
Mean	49.12	17.08	2.61	50.53	15.86	2.26
Minimum	8.67	-1.13	0.26	10.35	-1.45	0.12
1st quartile	40.96	14.43	2.08	41.96	12.28	1.68
Median	49.51	17.16	2.56	50.66	16.22	2.21
3rd quartile	58.10	20.38	3.01	59.84	19.77	2.73
Maximum	87.77	30.39	6.65	89.83	30.04	5.88

NA-number of missing data, %- percentage of missingness

5.1.2. AIR POLLUTION DATA

Figures 5.1 to 5.6 illustrate the time-series of the 24-hour averages of the five air pollutants of interest at the six different air monitoring stations from 1 January 2011 to 29 February 2020. The time-series show SO_2, NO_2, $PM_{2.5}$, and PM_{10} set against the 24-hour WHO guidelines (blue line) and National Ambient Air Quality Standards (NAAQS) of South Africa (red line). O_3 currently does not have 24-hour guidelines or standards. In the VTAPA it is evident that the 24-hour averages exceed the stricter WHO guidelines by more than the four exceedances permitted annually (Chapter 2).

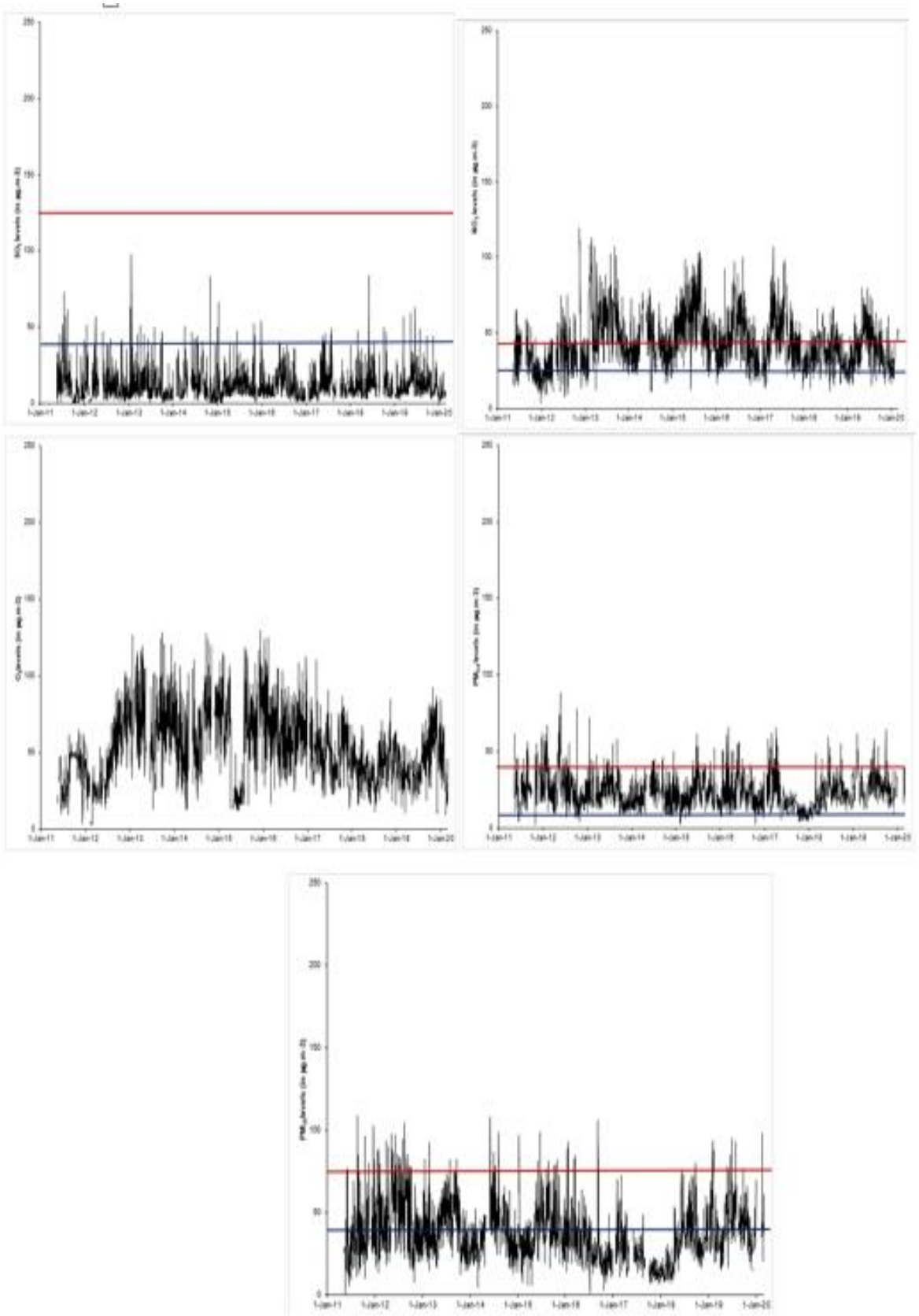


Figure 5.1: Time-series of daily SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentration (µg/m^3) in Diepkloof during 1 January 2011 to 29 February 2020.

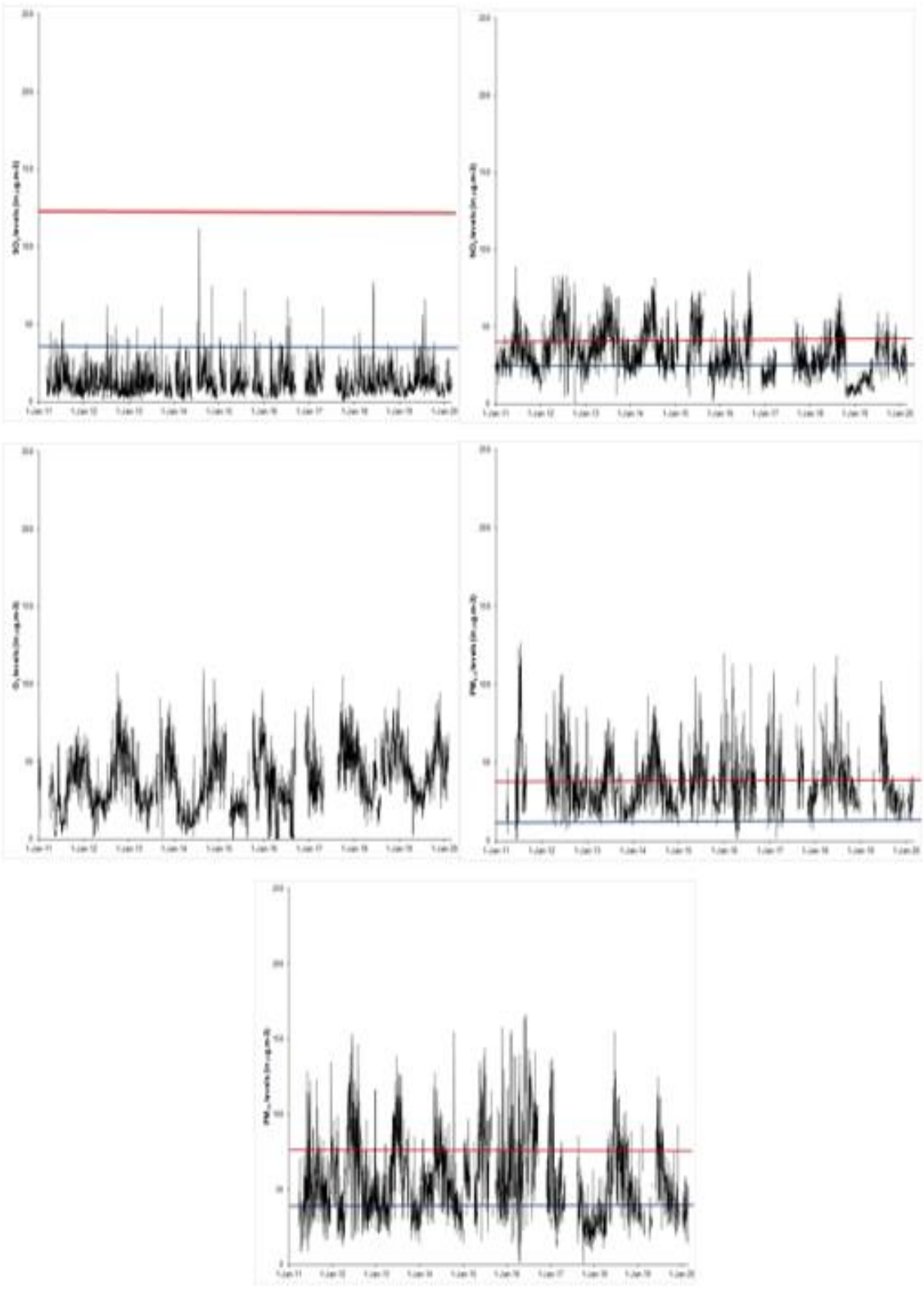


Figure 5.2: Time-series of daily SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, concentrations ($\mu g/m^3$) in Klipriver during 1 January 2011 to 29 February 2020.

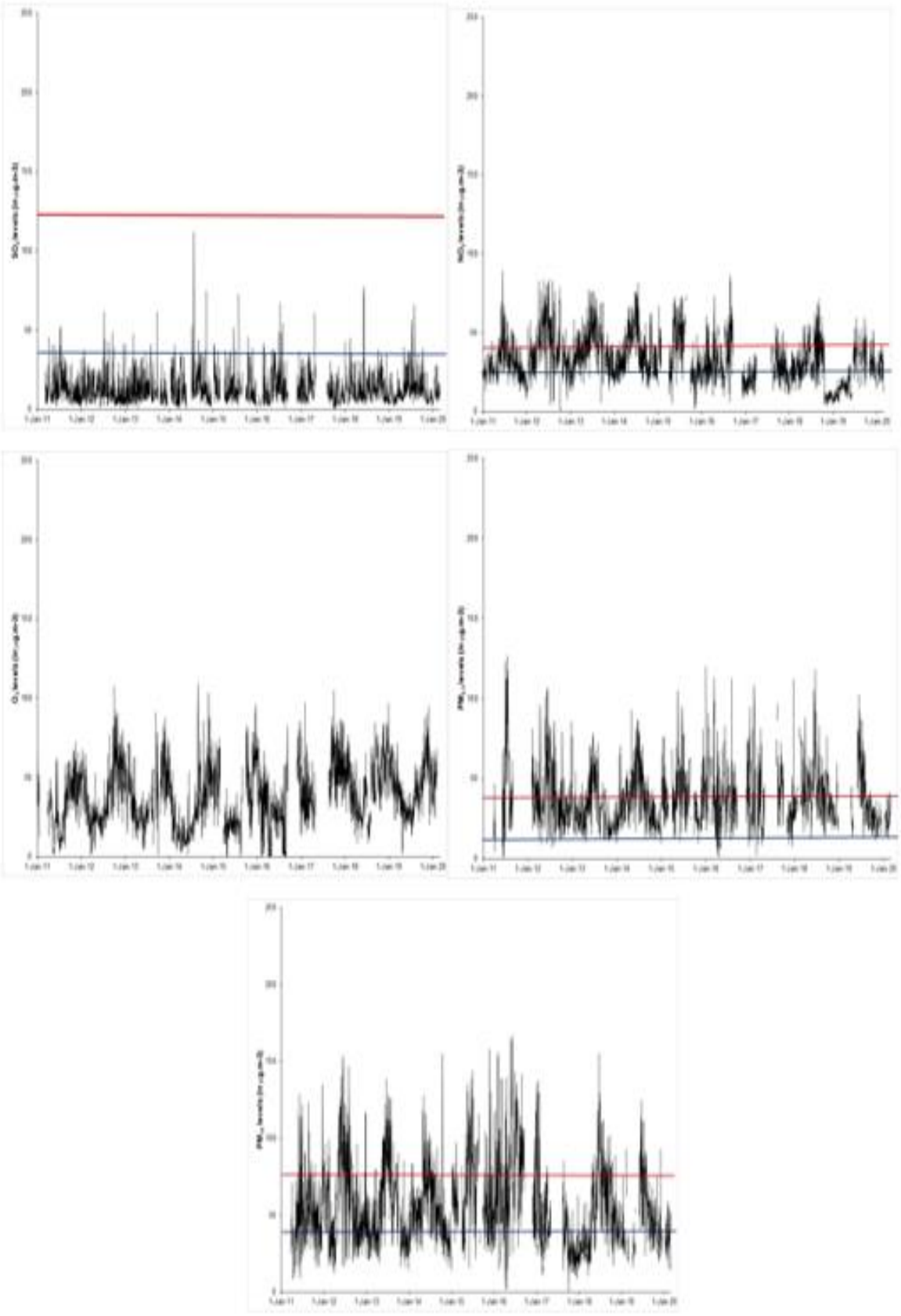


Figure 5.3: Time-series of daily SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10}, concentrations ($\mu g/m^3$) in Sebokeng during 1 January 2011 to 29 February 2020.

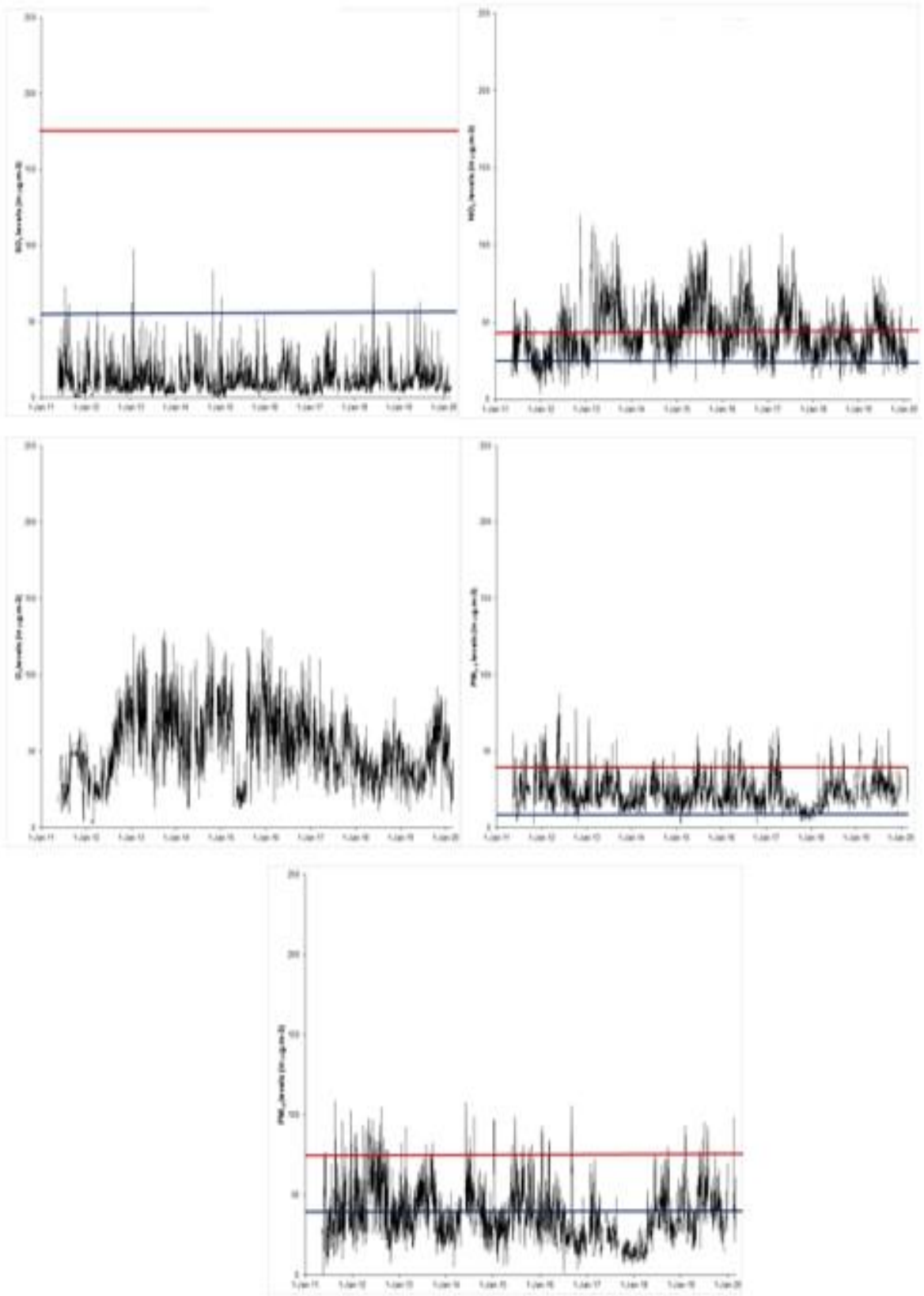


Figure 5.4: Time-series of daily SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, concentrations ($\mu g/m^3$) in Sharpeville during 1 January 2011 to 29 February 2020.

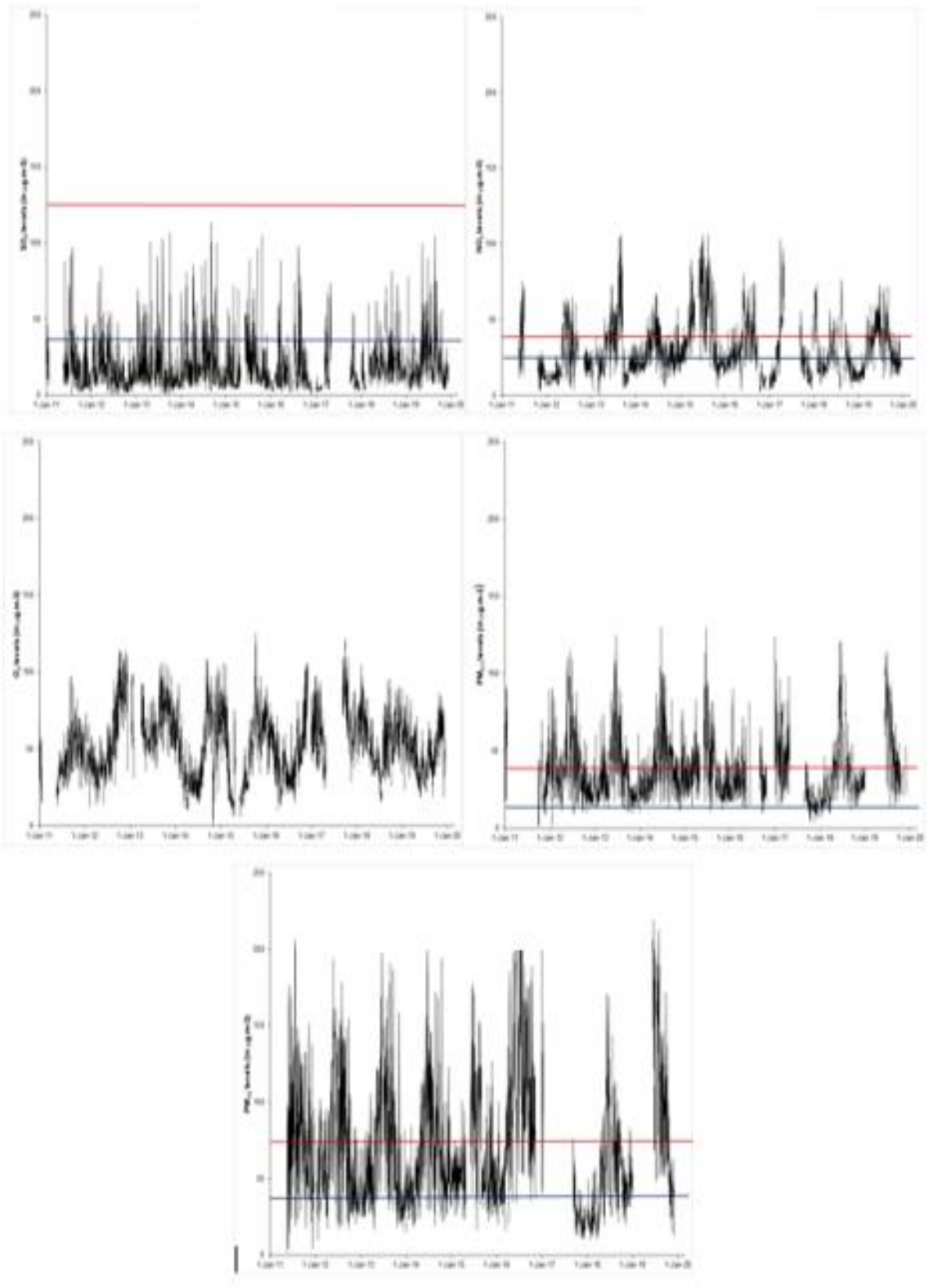


Figure 5.5: Time-series of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, concentrations in Three Rivers during 1 January 2011 to 29 February 2020.

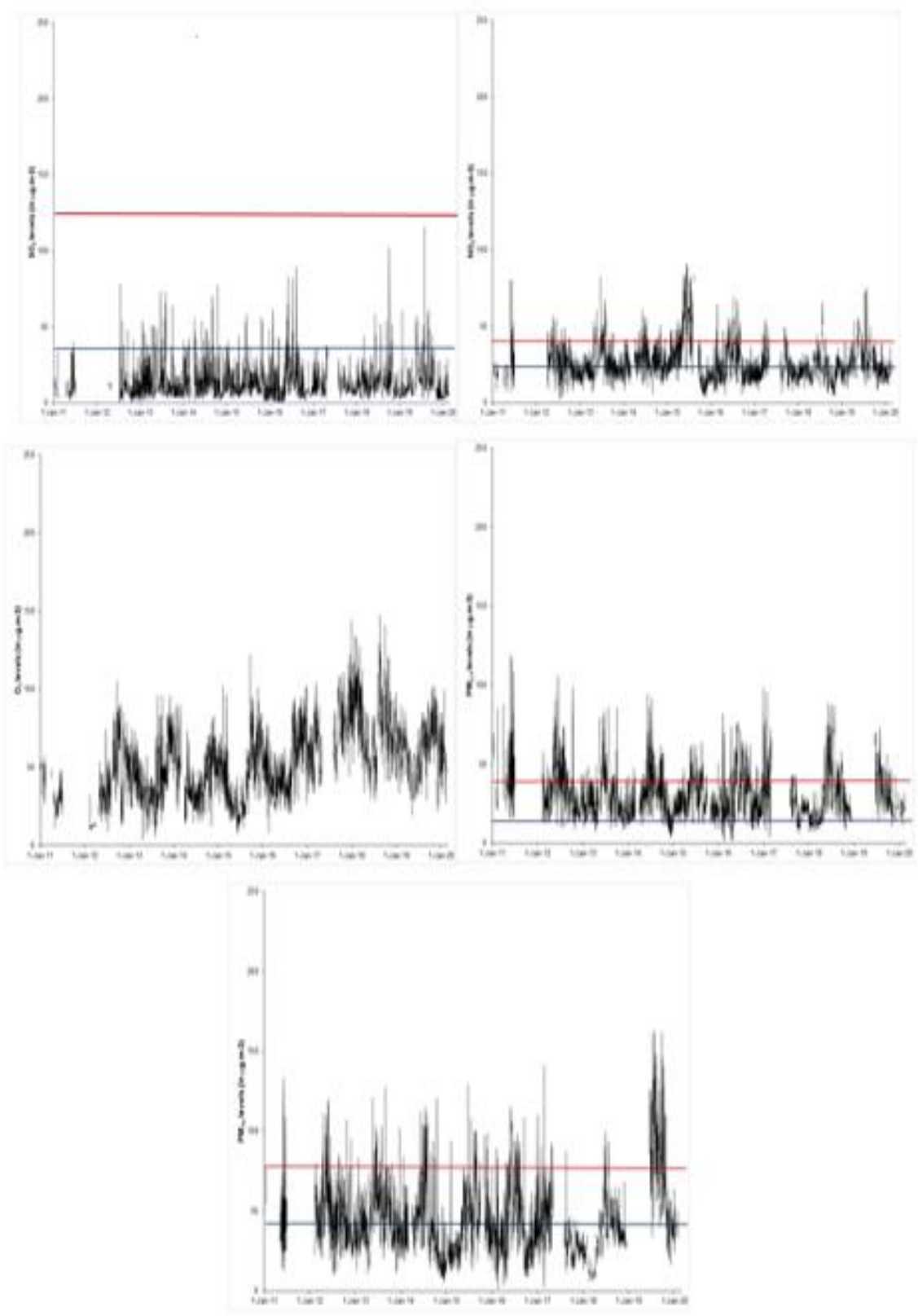


Figure 5.6: Time-series of daily SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, concentrations ($\mu g/m^3$) in Zamdela during 1 January 2011 to 29 February 2020.

Seasonal change occurred throughout the specified timeframe, i.e. 1 January 2011 to 29 February 2020. Concentration levels for SO_2, NO_2, $PM_{2.5}$, and PM_{10} showed increases in cold winter months (1 June to 31 August) and a decrease in warmer summer months (1 December to 28/29 February). O_3 showed to have an inverse trend, showing increased concentration levels in warmer summer months and decreases in colder winter months.

Table 5.3 shows a summary of statistics for the 30 datasets before imputation was implemented. The highest proportion of missing data was seen in Three Rivers for $PM_{2.5}$ (~34%), while the lowest proportion of missing data was seen in Zamdela for SO_2 (~11%).

Table 5.3: Descriptive statistics for the six monitoring stations in the VTAPA on SO_2, NO_2, O_3, $PM_{2.5}$, PM_{10}, and BC, before imputation.

	SO_2	NO_2	O_3	$PM_{2.5}$	PM_{10}	SO_2	NO_2	O_3	$PM_{2.5}$	PM_{10}
	Diepkloof					**Sharpeville**				
Mean	12.19	43.10	53.02	23.87	37.29	18.58	29.92	52.98	37.61	70.24
Minimum	0.08	3.60	3.00	2.10	0.19	0.29	0.43	0.20	1.62	1.81
1st quartile	5.57	30.91	36.37	16.60	25.31	7.88	17.07	37.19	23.09	40.35
Median	9.49	40.70	50.68	22.30	34.75	13.86	25.02	51.84	32.39	59.58
3rd quartile	15.61	52.89	68.14	29.30	46.67	23.68	39.05	67.17	47.05	92.28
Maximum	98.04	119.24	129.65	88.21	108.83	113.12	106.83	125.35	129.76	219.22
NA	492	447	441	605	580	629	802	599	1036	874
%	14.70	13.36	13.18	18.08	17.33	18.79	23.96	17.90	30.95	26.11
	Kliprivier					**Three Rivers**				
Mean	12.41	32.76	37.75	37.69	55.43	14.69	23.29	50.13	25.53	52.48
Minimum	0.11	0.84	0.04	0.06	0.045	0.21	1.51	1.12	0.49	0.94
1st quartile	6.14	22.48	23.63	24.21	34.81	7.02	15.22	33.33	16.87	32.60
Median	9.90	30.51	35.56	33.83	50.153	11.21	20.85	47.69	23.29	46.85
3rd quartile	15.91	41.25	51.14	46.60	71.05	18.06	29.65	64.60	31.64	68.18
Maximum	112.01	88.93	111.91	126.74	166.54	117.5	85.21	125.35	96.52	144.16
NA	579	537	502	1014	652	711	808	807	1164	961
%	17.30	16.04	15.00	30.30	19.48	21.24	24.14	24.11	34.78	28.71
	Sebokeng					**Zamdela**				
Mean	13.40	26.17	52.78	30.49	44.40	21.60	26.21	47.01	30.39	59.17
Minimum	0.02	1.89	3.54	0.88	1.59	0.01	0.15	0.63	0.23	0.41
1st quartile	6.01	18.06	35.09	19.34	28.22	9.22	16.76	35.69	18.93	32.56
Median	10.12	23.53	50.63	27.04	39.40	17.76	23.98	46.17	27.22	50.27
3rd quartile	16.89	31.52	68.25	37.76	54.98	29.23	33.55	57.81	38.58	76.26
Maximum	115.67	97.82	147.45	119.78	162.74	105.09	91.50	108.67	108.80	213.11
NA	903	834	731	993	988	380	1015	463	671	932
%	26.98	24.92	21.84	29.67	29.52	11.35	30.33	13.83	20.05	27.85

NA-number of missing data, %- percentage of missingness

..

5.1.2.1. PATTERNS OF MISSING DATA

There were two techniques used to illustrate missing data. Firstly, figures 5.7 to 5.12 (which used the ImputeTS package) show where missing data occurs represented along the time-series graphs. The red blocks represent the missing data, while the blue points represent the available data.

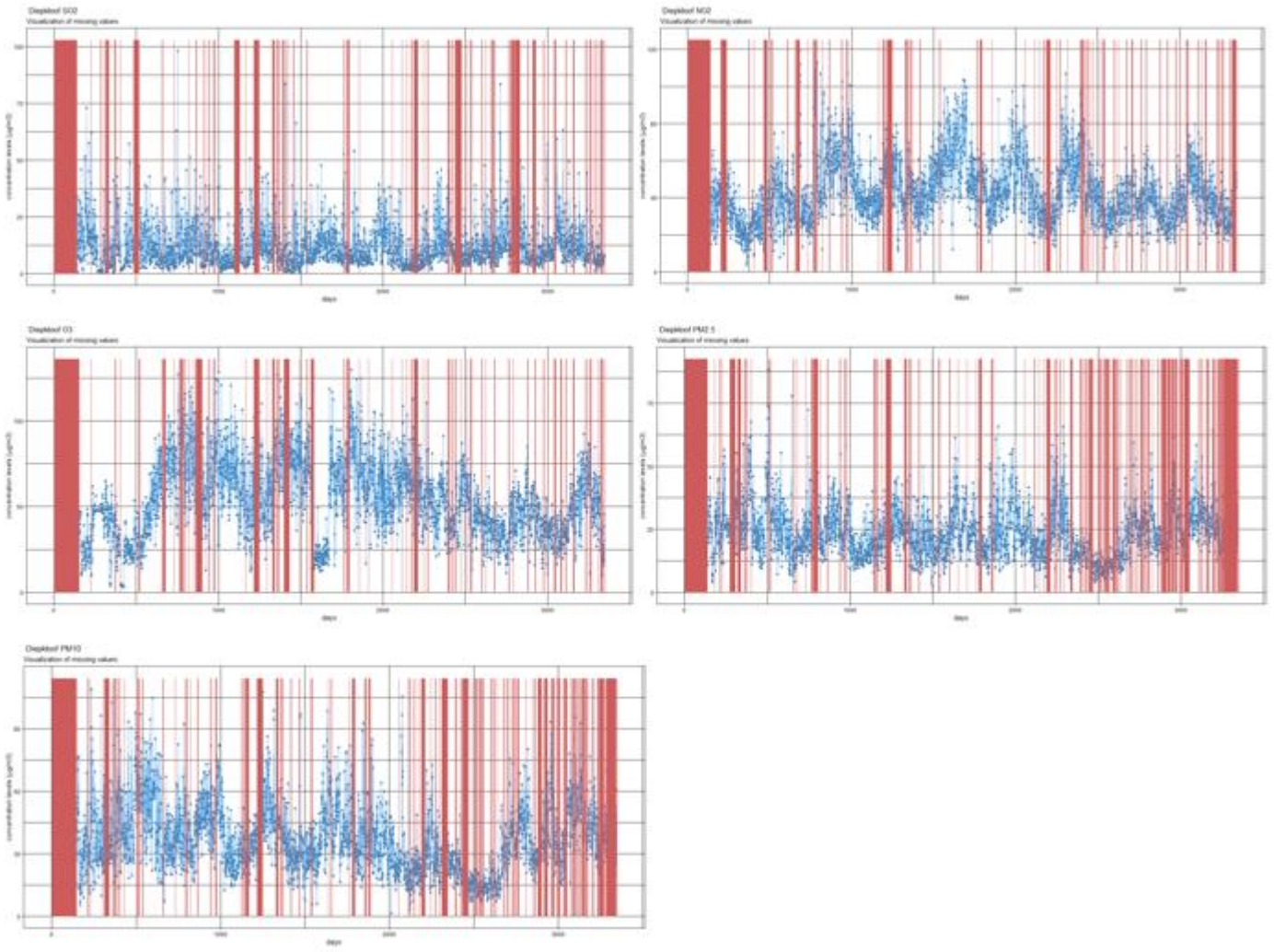


Figure 5.7: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations ($\mu g/m^3$) Diepkloof.

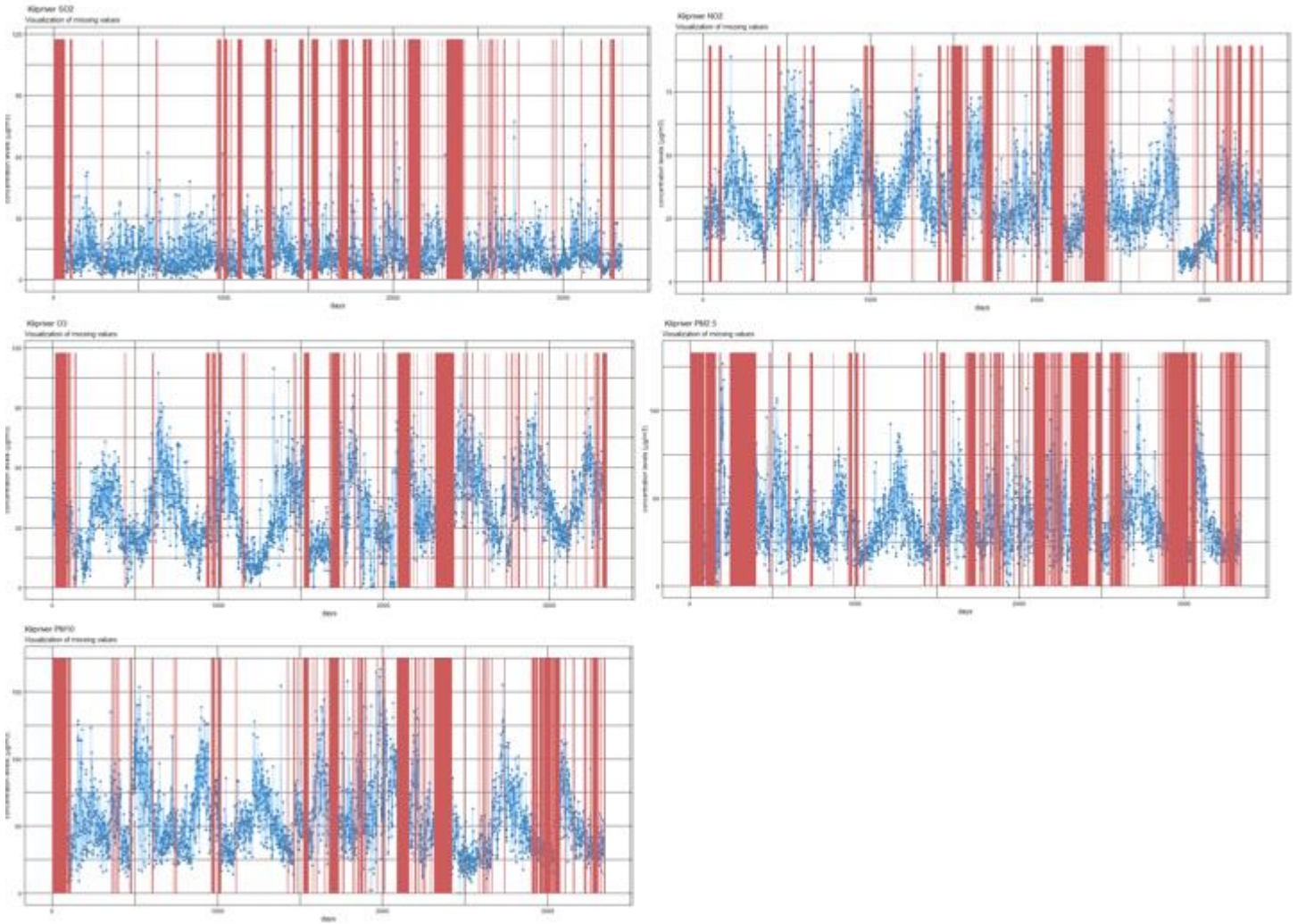


Figure 5.8: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations ($\mu g/m^3$) Kliprivier.

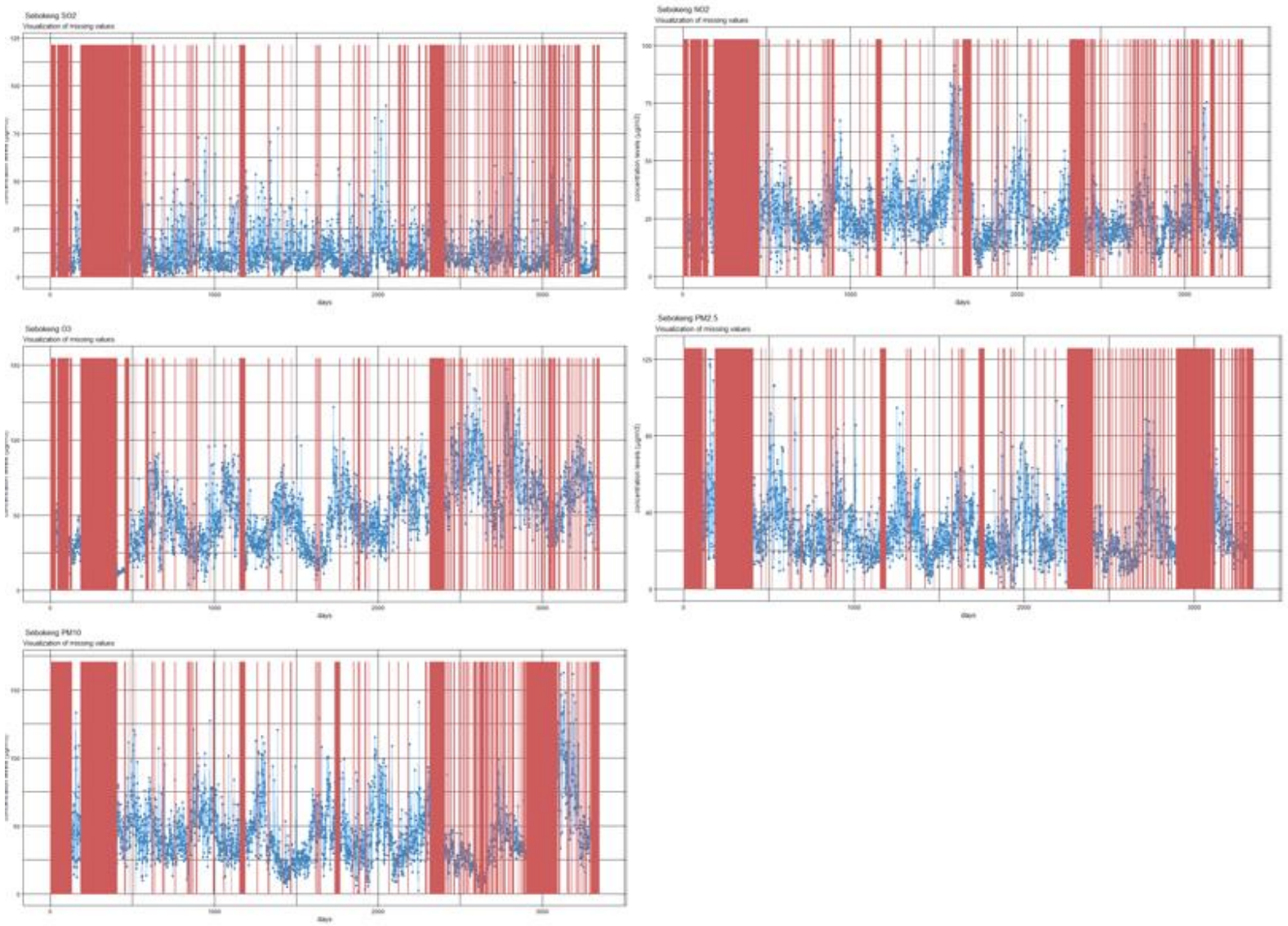


Figure 5.9: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations ($\mu g/m^3$) Sebokeng.

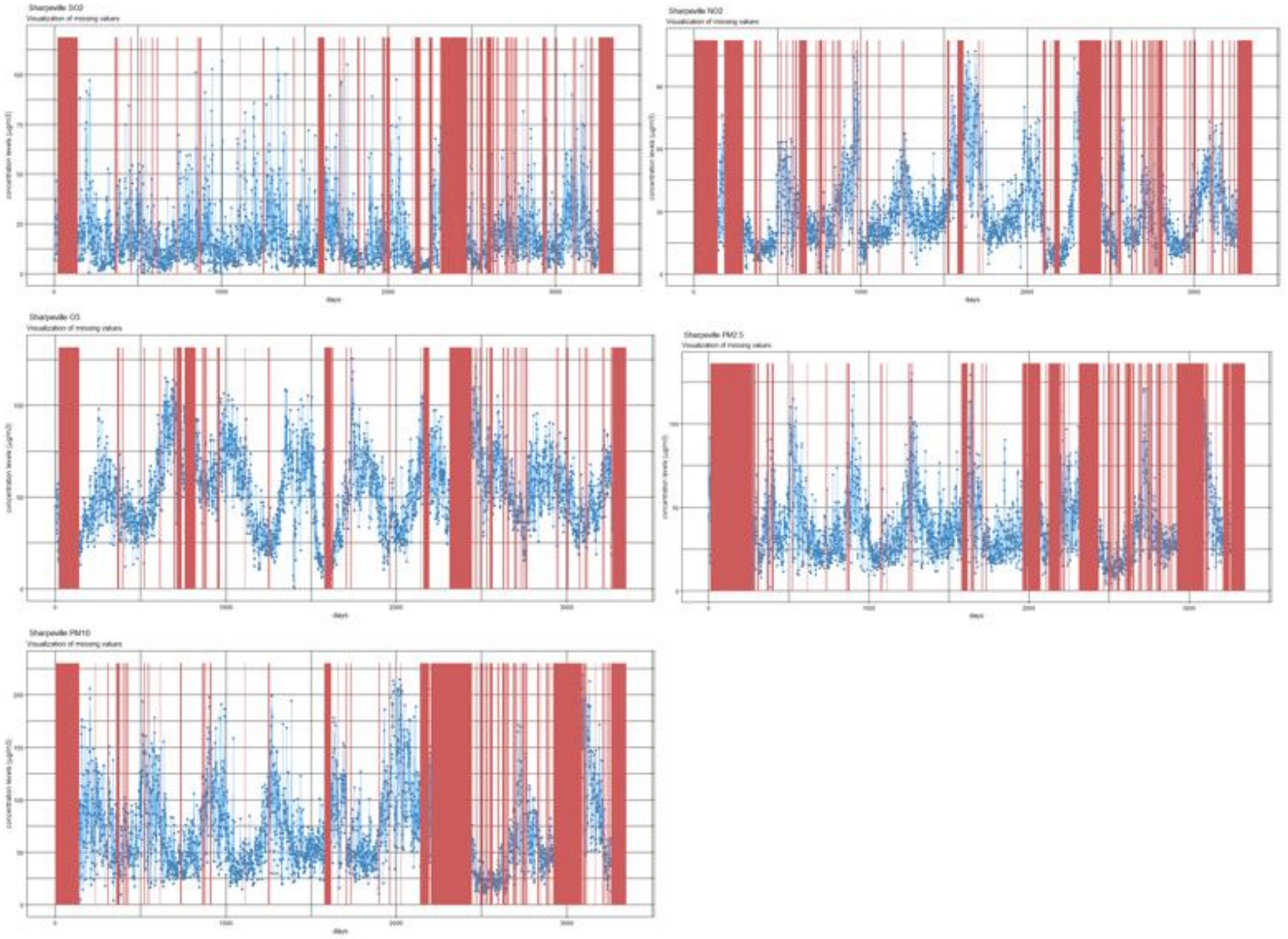


Figure 5:10: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations ($\mu g/m^3$) Sharpeville.

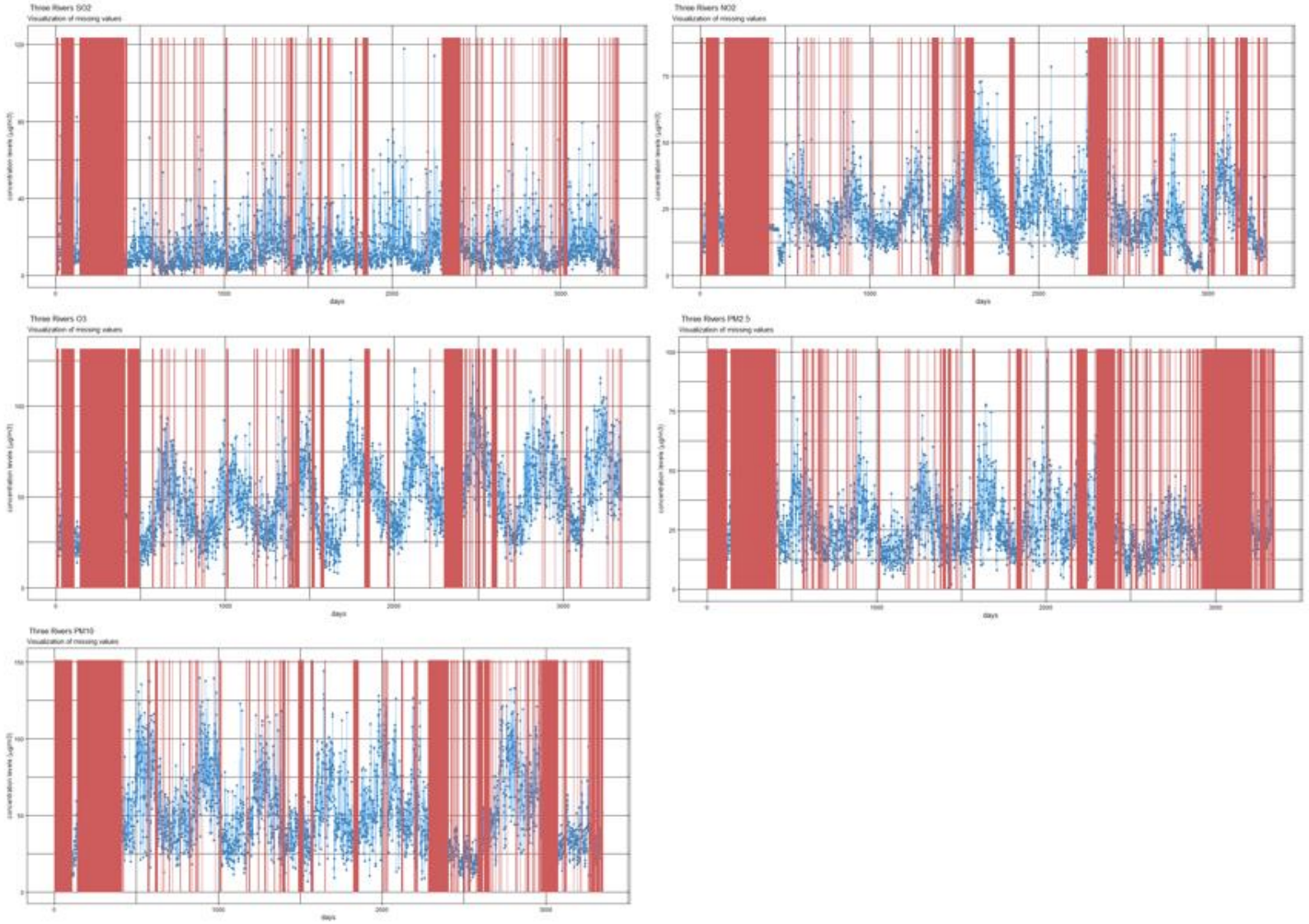


Figure 5.11: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations ($\mu g/m^3$) Three Rivers.

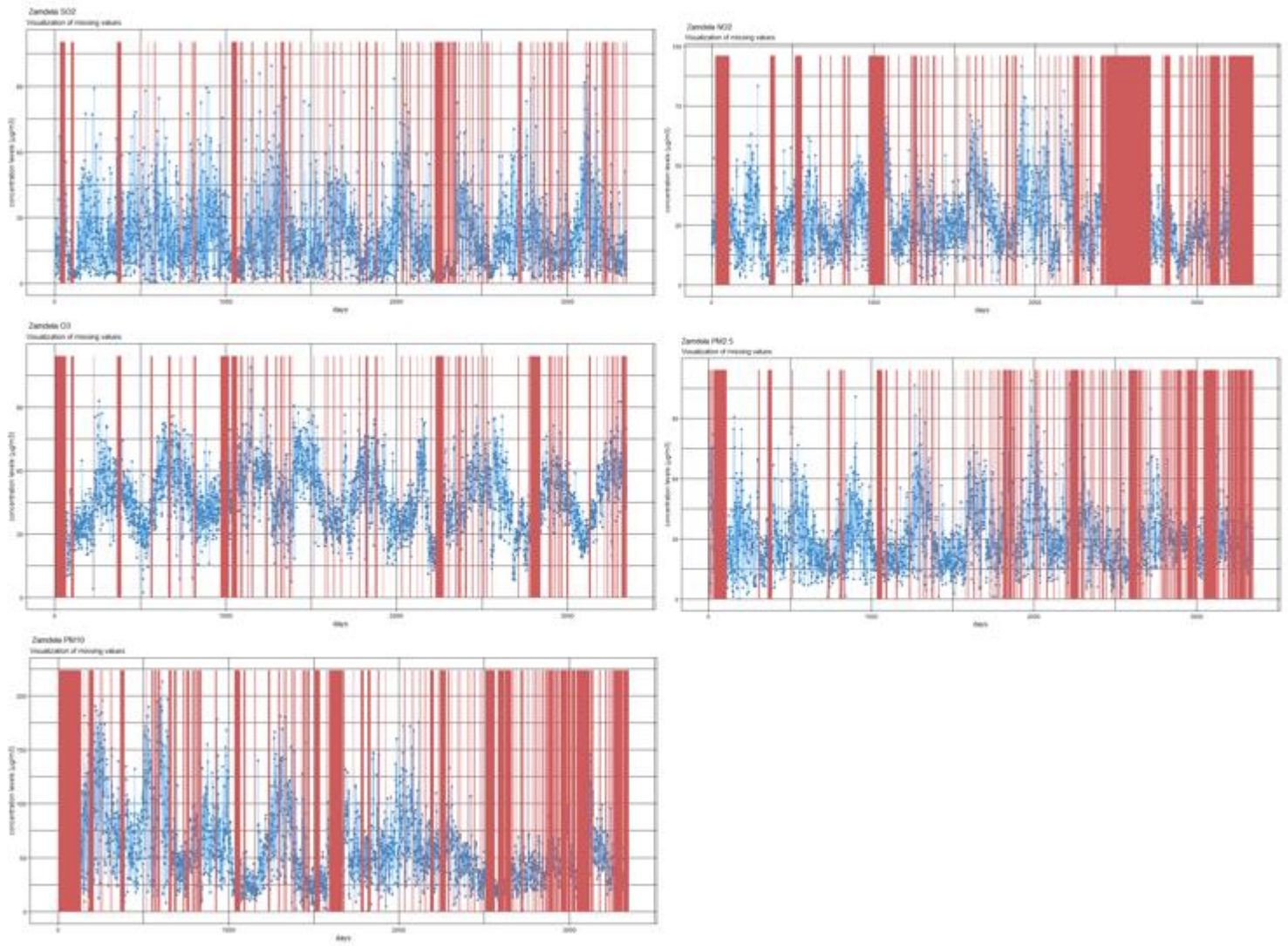


Figure 5.12: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations (µg/m^3) Zamdela.

Figures 5.13 to 5.18 (formed using the VIM package) show the patterns of missing data within the dataset, where yellow represents the missing data and blue the available/observed data. The amount of missing data, where all five pollutants were missing simultaneously in Diepkloof, Kliprivier, Sebokeng, Sharpeville, Three Rivers, and Zamdela were, 8.84%, 10.64%, 19.54%, 14.04%, 18.40%, and 8.16%, respectively. Data that was simultaneously available for all six pollutants in Diepkloof, Kliprivier, Sebokeng, Sharpeville, Three Rivers, and Zamdela were indicated as 67.25%, 57.04%, 56.83%, 53.54%, 53.06%, and 51.54%, respectively. For five out of the six areas, $PM_{2.5}$ had the most missing data, but in the sixth area, Zamdela, SO_2 was the pollutant with the least missing data.

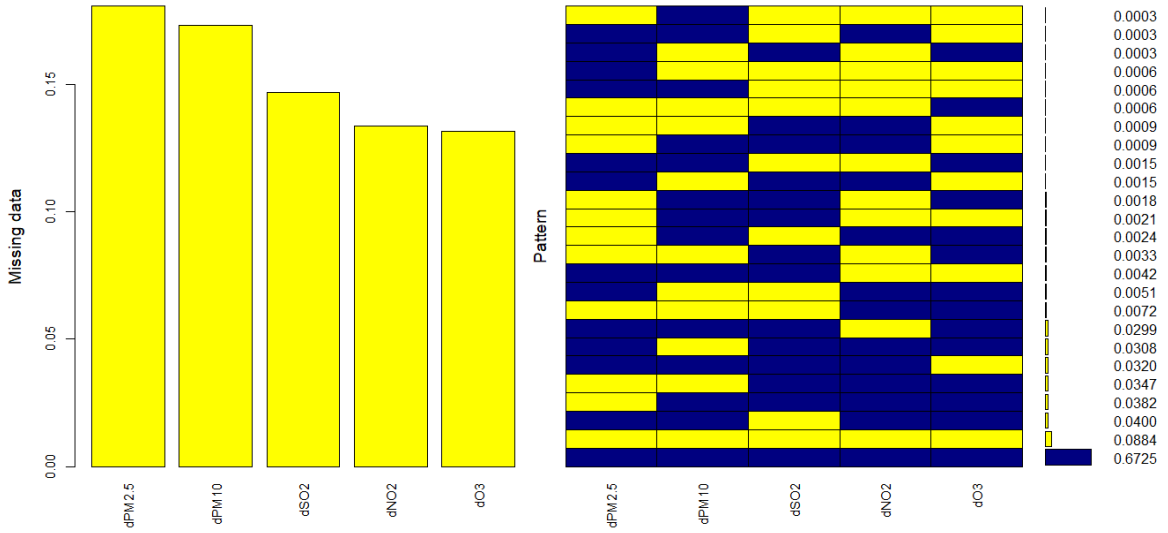


Figure 5.13: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, Diepkloof.

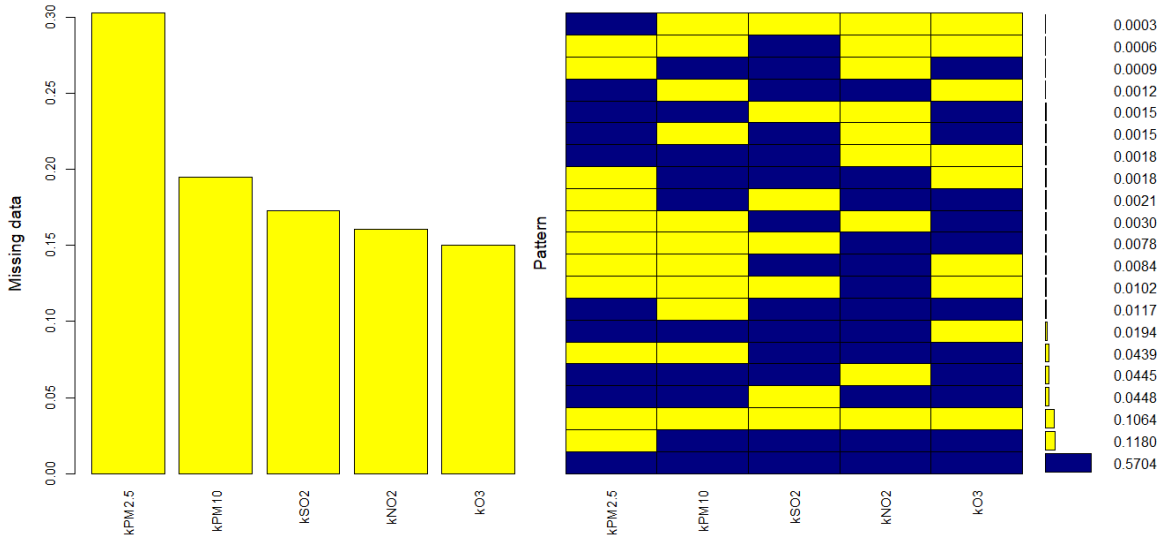


Figure 5.14: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, Kliprivier.

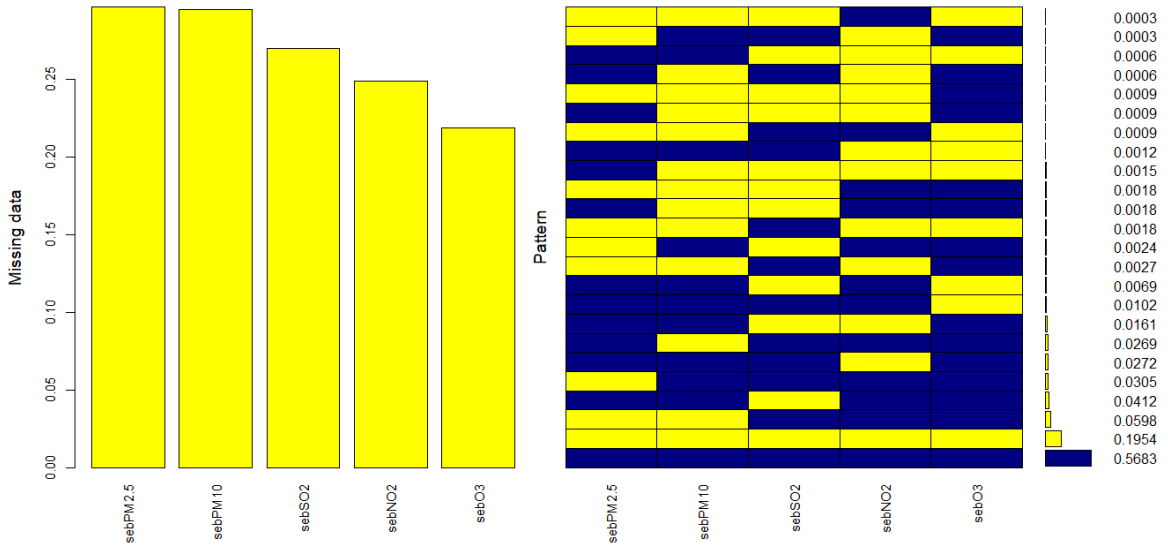


Figure 5.15: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, Sebokeng.

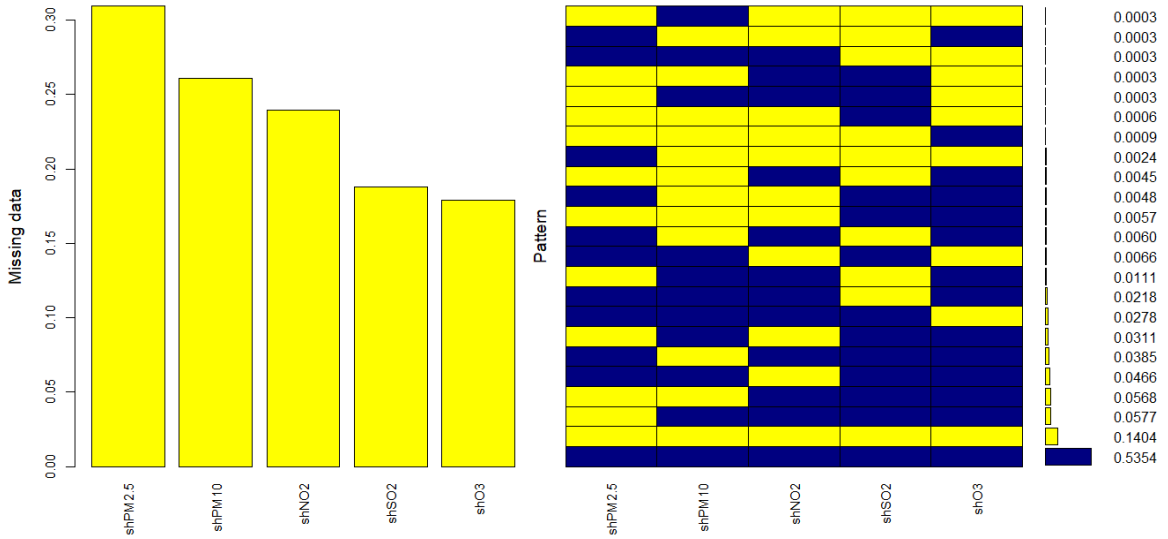


Figure 5.16: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, Sharpeville.

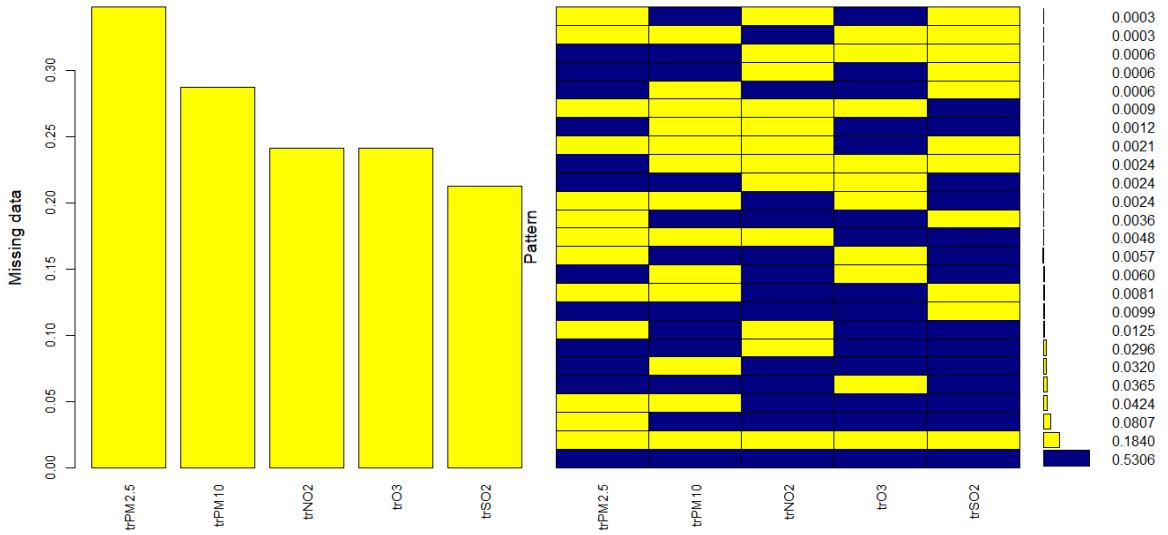


Figure 5.17: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, Three Rivers.

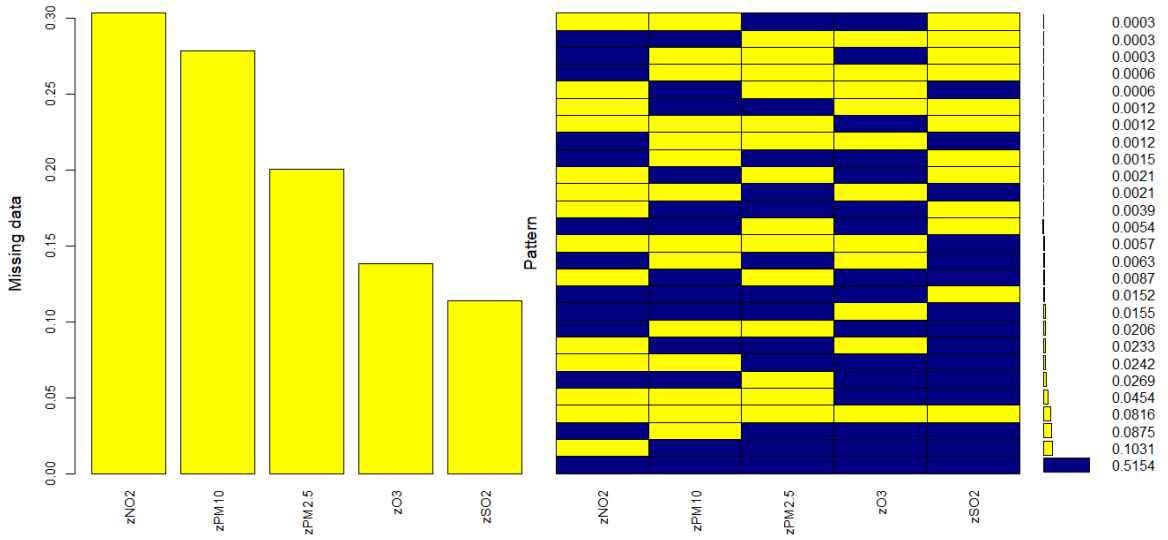


Figure 5.18: Visualisation of missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, Zamdela.

5.1.3. DATA IMPUTATIONS

Examples of simple univariate imputations such as mean, median, random, and LOCF imputation can be seen in Figure 5.19. This illustrates the ill-fit imputations for the missing data. All the imputations were run as a single pollutant and then compiled into a single figure.

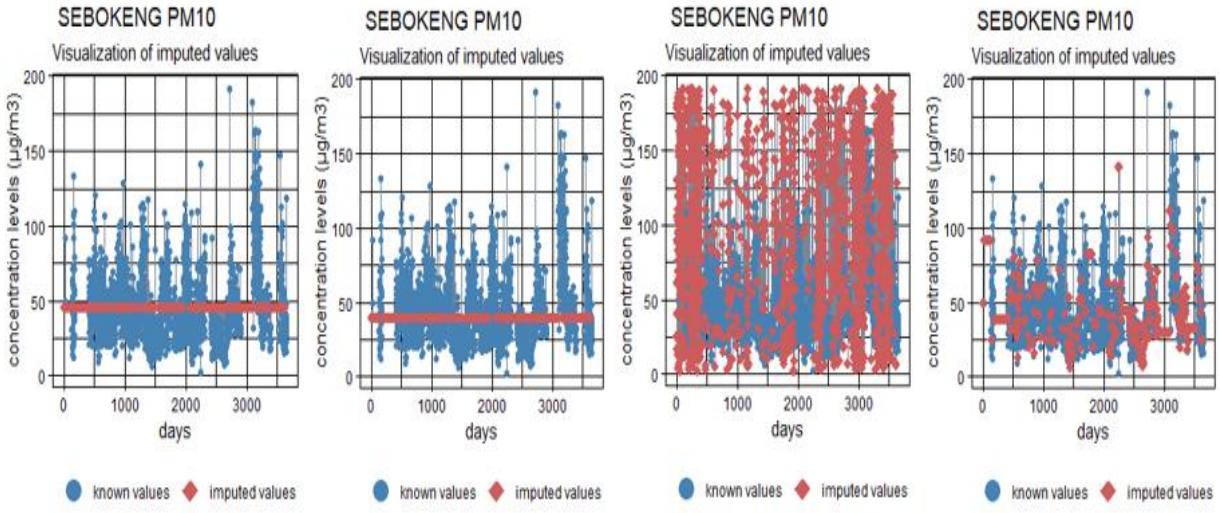


Figure 5.19: Examples of univariate imputation methods mean, median, random and last observation carried forward, respectively.

5.1.3.1. KALMAN IMPUTATION

Table 5.4 shows the descriptive statistics of the six stations for each pollutant of interest, i.e. SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}. The concentration levels remain within the range of values from the concentration levels prior to imputation. As a result, PM_{10} showed to have the highest concentration levels in Kliprivier, Sebokeng, Sharpeville, Three Rivers, and Zamdela at 166.5 µg/m^3, 162.7 µg/m^3, 219.2 µg/m^3, 144.2 µg/m^3, and 215.4 µg/m^3, respectively. SO_2 showed to have the lowest concentration levels in Diepkloof, Sebokeng, Sharpeville, Three Rivers, and Zamdela at 0.08 µg/m^3; 0.02 µg/m^3; 0.29 µg/m^3; 0.21 µg/m^3; and 0.01 µg/m^3; respectively. SO_2 mean concentration levels ranged from 12.37 µg/m^3 - 21.1 µg/m^3; NO_2 mean concentration levels ranged from 22.97 µg/m^3 - 42.58 µg/m^3; O_3 mean concentration levels ranged from 37.9 µg/m^3 - 52.84 µg/m^3; $PM_{2.5}$ mean concentration levels ranged from 24.2 µg/m^3 - 38.38 µg/m^3; and PM_{10} mean concentration levels ranged from 36.92 µg/m^3 - 68.81 µg/m^3.

Figures 5.20 to 5.25 show complete datasets after using the Kalman imputation, where the red indicates the imputed values and the blue represents original values.

Table 5.4: Descriptive statistics for the six monitoring stations in the VTAPA on SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10}, after Kalman imputation.

	SO_2	NO_2	O_3	$PM_{2.5}$	PM_{10}	SO_2	NO_2	O_3	$PM_{2.5}$	PM_{10}
	Diepkloof					**Sharpeville**				
Mean	12.38	42.58	52.84	24.20	36.92	19.19	30.10	51.06	38.38	68.81
Minimum	0.08	3.60	3.00	2.10	0.19	0.29	0.43	0.20	1.62	1.81
1st quartile	6.02	29.94	36.12	17.29	25.46	8.70	17.53	34.99	23.87	37.47
Median	10.07	40.02	50.22	22.85	34.84	14.82	26.27	49.75	33.77	59.50
3rd quartile	15.53	52.52	68.19	29.48	46.06	24.28	39.43	65.19	49.08	93.06
Maximum	98.04	119.24	129.65	88.21	108.83	113.12	106.83	125.35	129.76	219.22
SD	9.38	17.10	22.12	10.13	16.08	15.47	17.06	21.28	19.62	40.42
	Kliprivier					**Three Rivers**				
Mean	12.37	32.46	37.90	37.88	55.28	15.14	22.97	48.02	26.63	50.30
Minimum	0.11	0.84	0.04	0.06	0.05	0.21	1.51	1.12	0.49	0.94
1st quartile	6.63	23.26	24.25	25.39	34.74	7.69	14.58	31.65	18.17	31.79
Median	10.69	29.99	36.78	34.87	49.97	11.87	20.18	45.94	24.19	44.71
3rd quartile	15.54	39.84	50.50	46.51	69.31	18.69	29.50	60.70	31.28	64.84
Maximum	112.01	88.93	111.91	126.74	166.54	117.65	85.21	125.35	96.52	144.16
SD	8.54	13.97	18.38	17.39	26.87	11.75	11.79	19.82	12.50	25.11
	Sebokeng					**Zamdela**				
Mean	13.38	27.10	50.65	31.85	45.73	21.10	25.92	46.12	30.75	57.13
Minimum	0.02	1.89	3.54	0.88	1.59	0.01	0.15	0.63	0.23	0.41
1st quartile	6.77	18.93	33.08	21.54	31.19	9.28	16.47	35.18	20.01	34.45
Median	11.66	24.82	48.15	29.07	41.99	17.63	23.94	45.77	27.63	50.19
3rd quartile	16.72	32.02	64.91	39.52	56.07	28.36	31.16	57.18	39.36	70.33
Maximum	115.67	97.82	147.45	119.78	162.74	105.09	91.50	108.67	108.80	213.11
SD	10.13	12.58	22.44	14.83	21.94	16.13	12.90	16.05	15.19	32.99

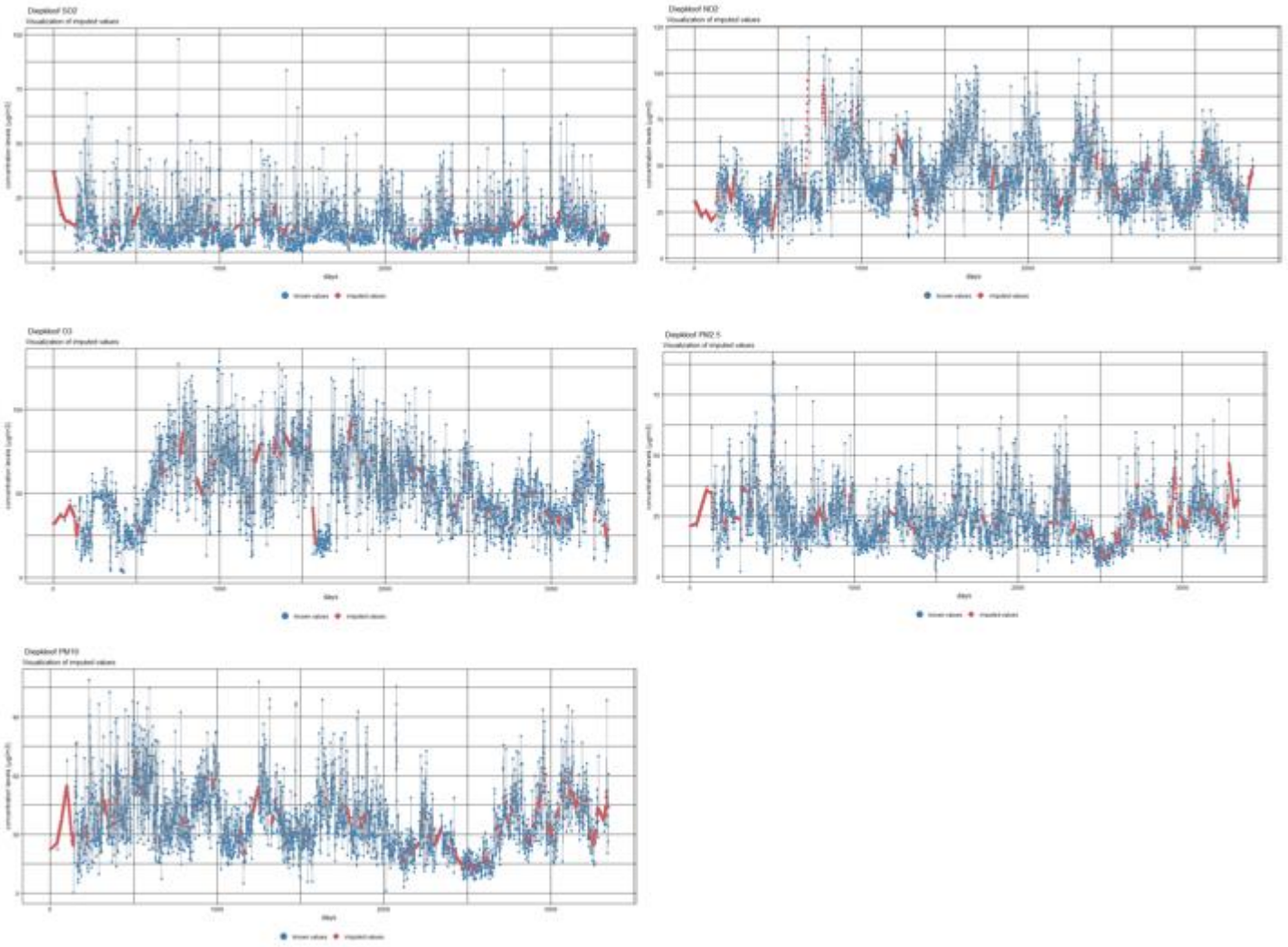


Figure 5.20: Visualisation of imputed data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations (µg/m^3), Diepkloof.

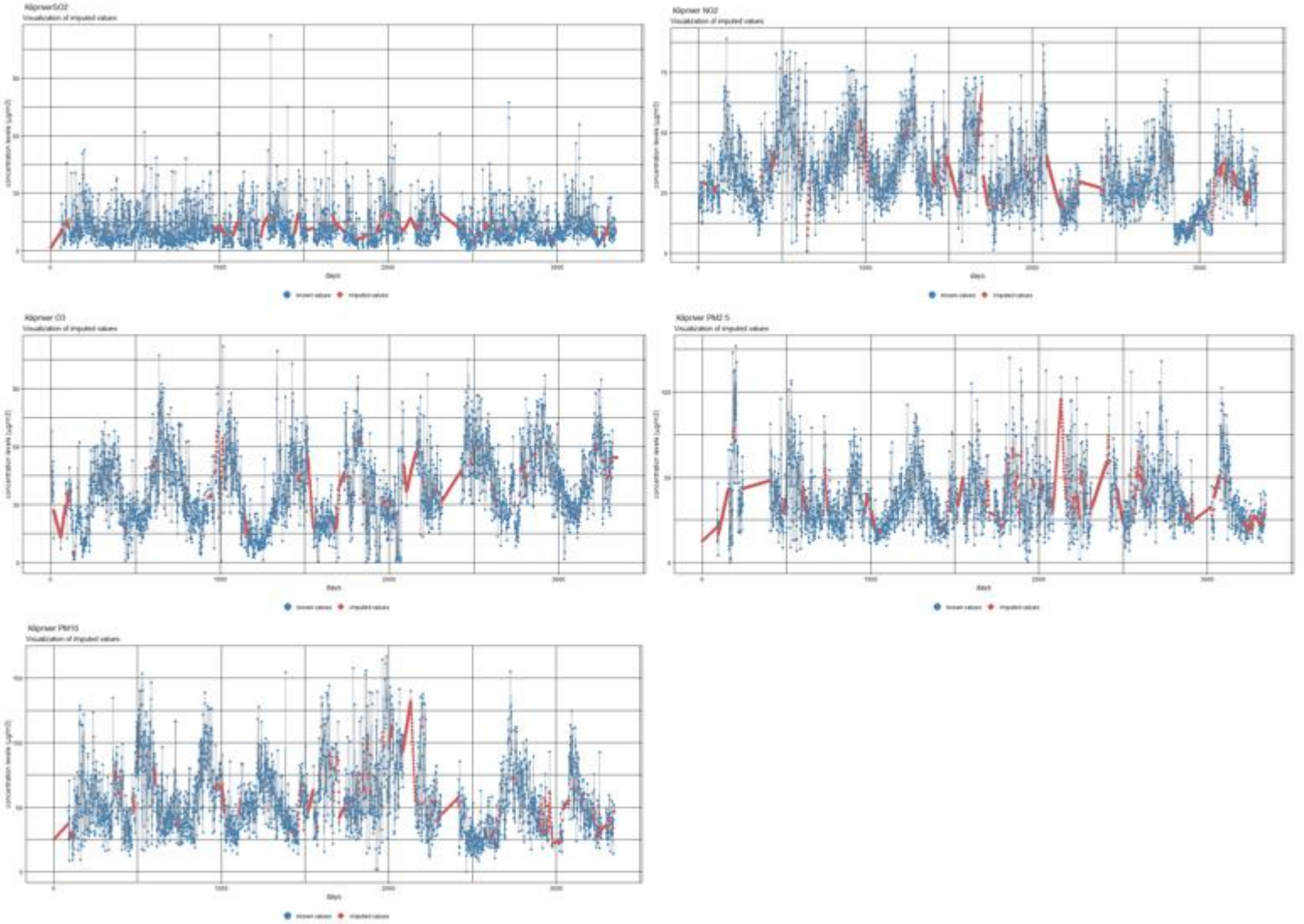


Figure 5.21: Visualisation of imputed data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, concentrations ($\mu g/m^3$) Kliprivier.

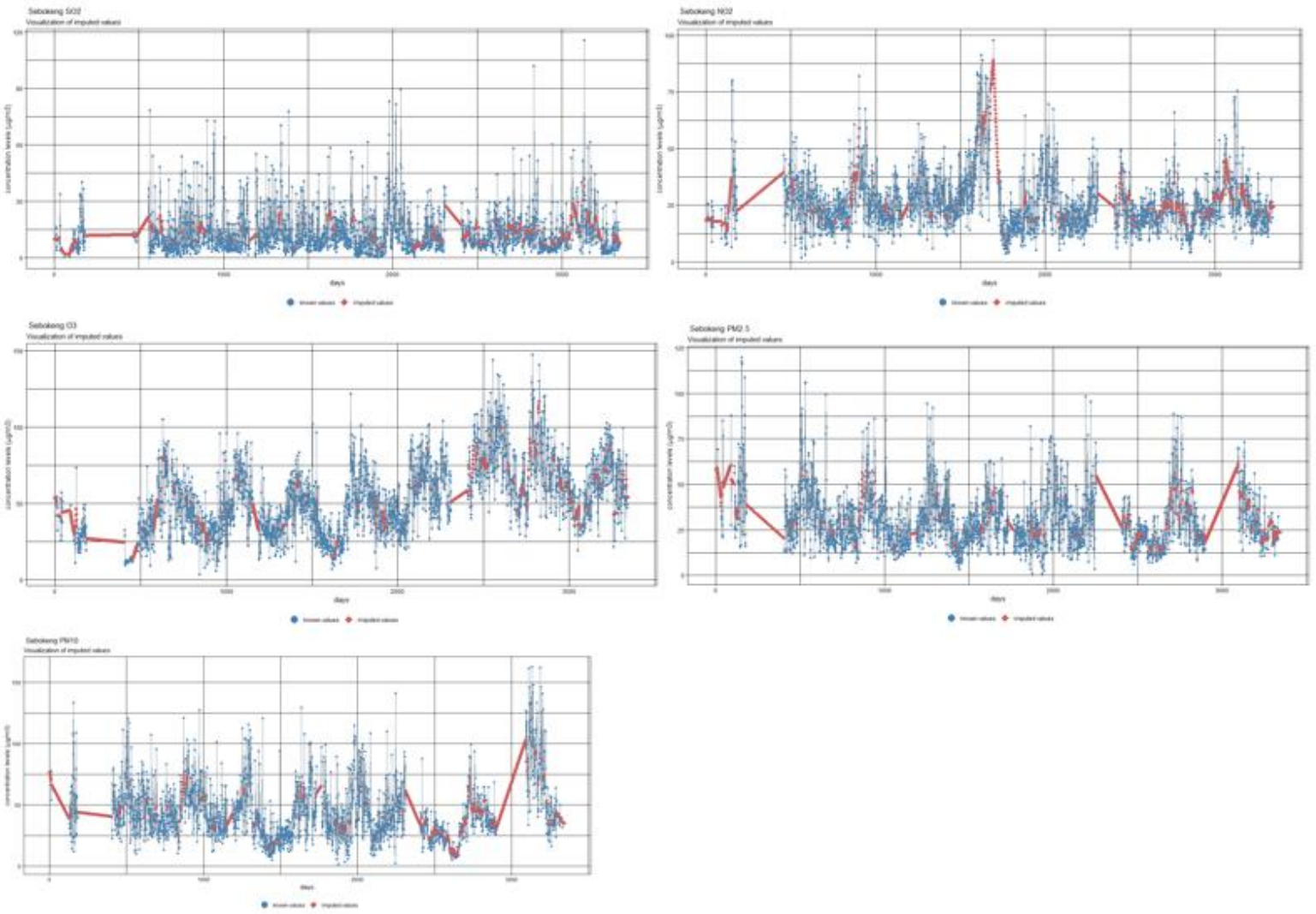


Figure 5.22: Visualisation of imputed data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations ($\mu g/m^3$), Sebokeng.

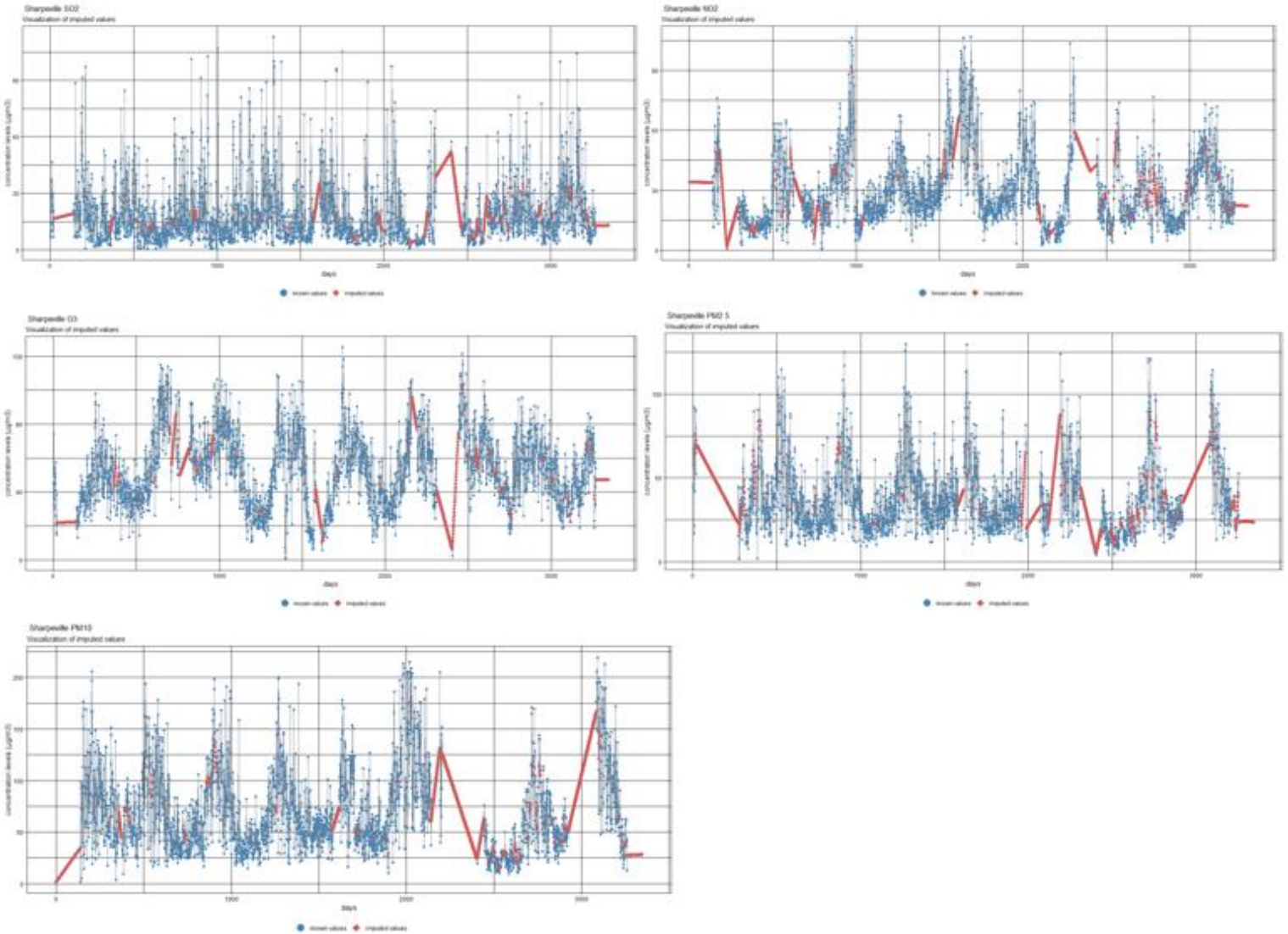


Figure 5.23: Visualisation of imputed missing data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations ($\mu g/m^3$), Sharpeville.

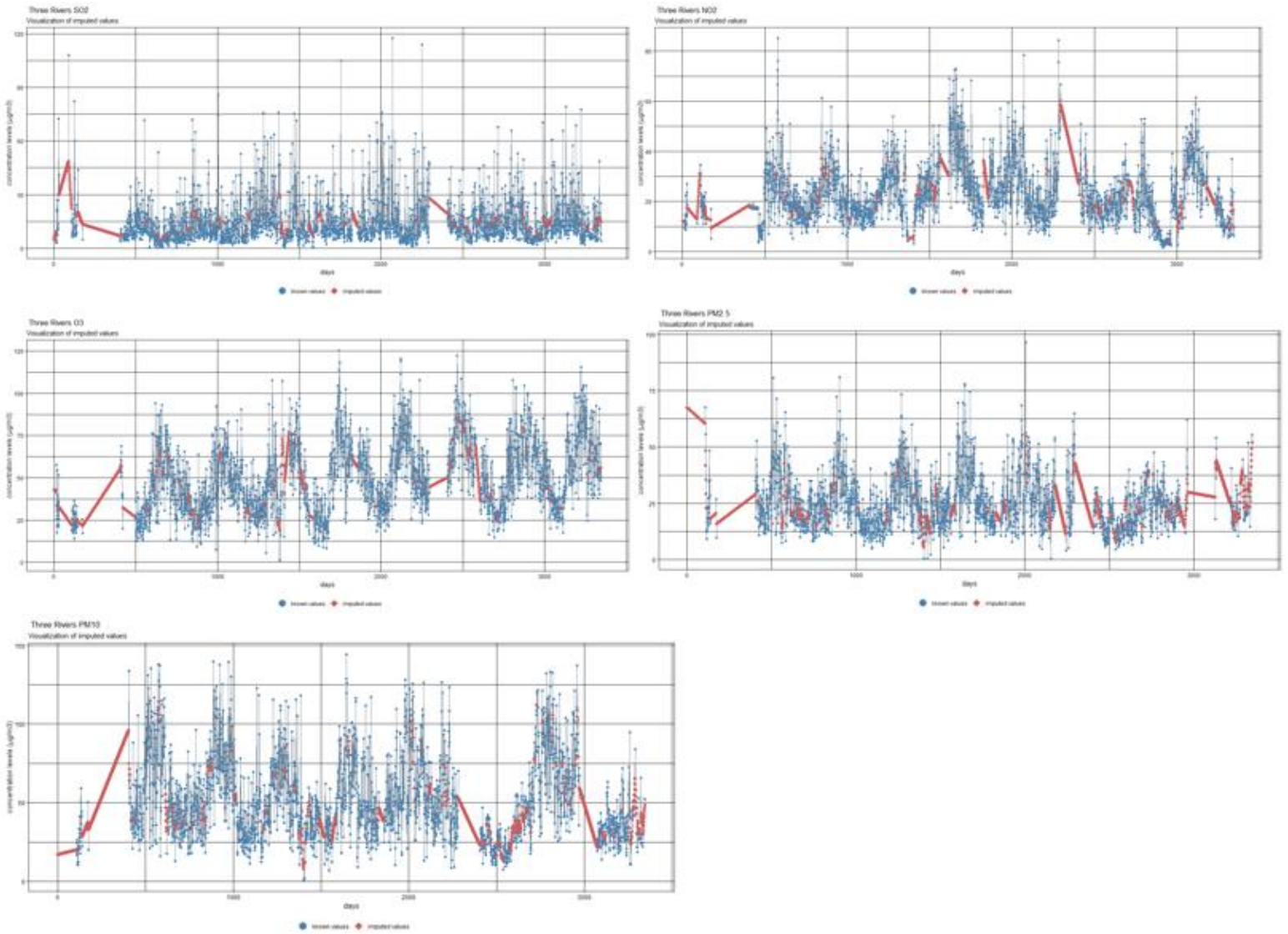


Figure 5.24: Visualisation of imputed data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations ($\mu g/m^3$), Three Rivers.

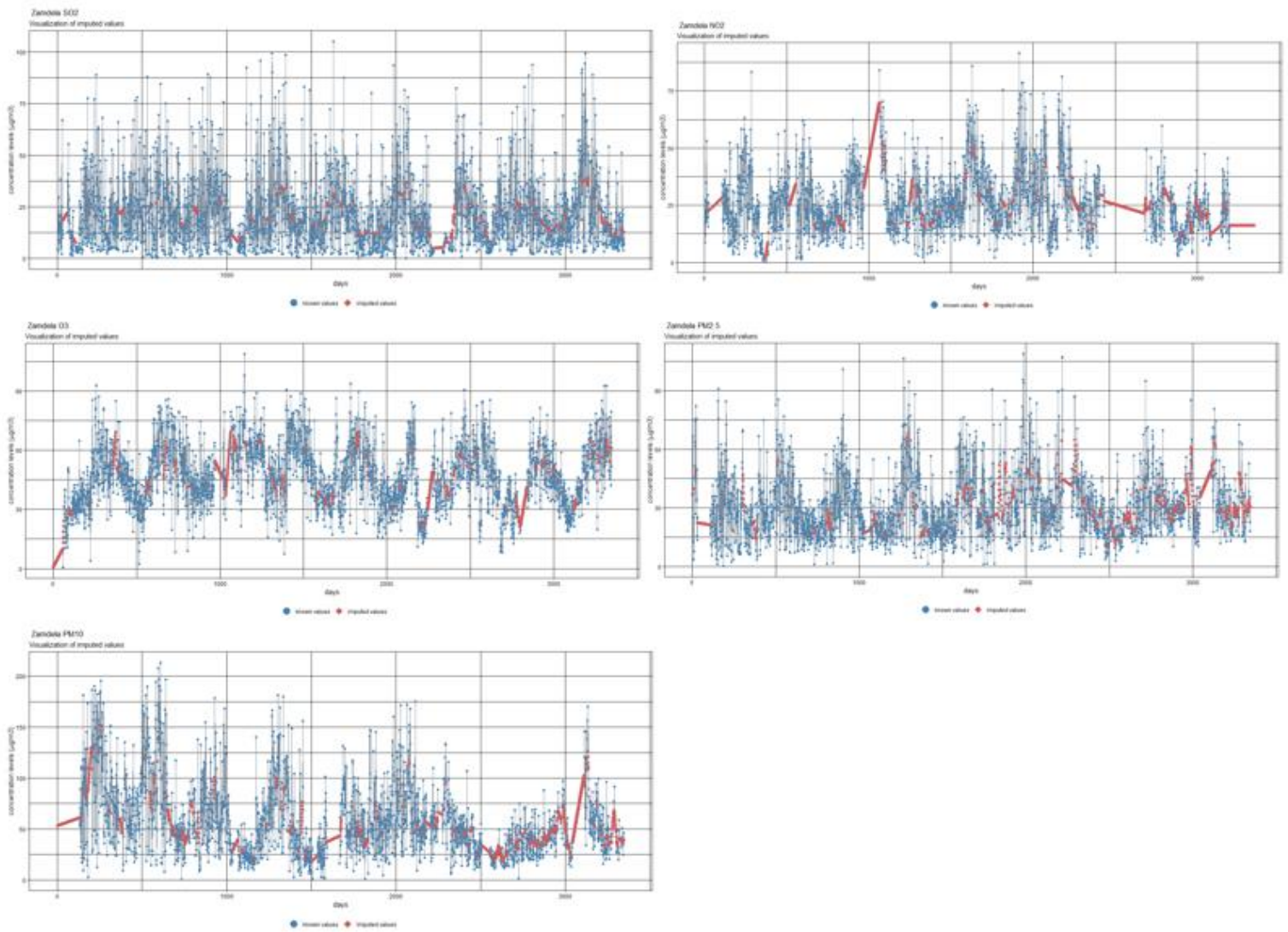


Figure 5.25: Visualisation of imputed data for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations (µg/m^3), Zamdela.

5.1.3.2. MULTIVARIATE IMPUTATION BY CHAINED EQUATIONS (MICE) IMPUTATION

The descriptive statistics for the six stations, as well as the variables at each station after mice imputation, are show in table 5.5. Similar to the Kalman imputations, the minimum and maximum values for the air monitoring stations did not differ from the values prior to imputation. The SO_2 mean concentration levels ranged from 12.08 µg/m^3 - 21.49 µg/m^3; NO_2 mean concentration levels ranged from 23.46 µg/m^3 - 43.09 µg/m^3; O_3 mean concentration levels ranged from 37.6 µg/m^3 - 52.97 µg/m^3; $PM_{2.5}$ mean concentration levels ranged from 25.09 µg/m^3 - 39.05 µg/m^3, and PM_{10} mean concentration levels ranged from 36.98 µg/m^3 - 69.6 µg/m^3.

Figure 5.26 to 5.31 shows density plots for each set of the five pollutants in each area. The imputed points (magenta graphs) match fairly well with the shape of the observed points (blue graph) at each of the 20 to 30 imputed datasets. Figures 5.32 to 5.36 show density plots of mice imputed datasets with meteorological variables, i.e. temperature, relative humidity, and wind speed were added as additional variables, (in magenta) against the original dataset (in blue). There were more imputed datasets that diverge from the original dataset.

Table 5.5: Descriptive statistics for the six monitoring stations in the VTAPA Area on SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, after mice imputation.

	SO_2	NO_2	O_3	$PM_{2.5}$	PM_{10}	SO_2	NO_2	O_3	$PM_{2.5}$	PM_{10}
	Diepkloof					**Sharpeville**				
Mean	12.08	43.09	52.98	23.69	36.98	18.42	29.80	52.97	39.05	69.60
Minimum	0.08	3.60	3.00	2.10	0.19	0.29	0.43	0.20	1.62	1.81
1st quartile	6.07	31.93	38.44	17.30	26.47	9.09	18.74	39.70	25.48	43.51
Median	9.96	41.10	51.31	22.38	34.79	14.86	26.39	52.23	35.03	62.14
3rd quartile	14.93	51.61	66.24	28.07	44.86	22.74	37.18	64.78	47.27	86.95
Maximum	98.04	119.24	129.65	88.21	108.83	113.12	106.83	125.35	129.76	219.22
SD	9.16	16.06	21.26	9.82	15.71	14.50	15.92	19.04	18.94	36.21
	Kliprivier					**Three Rivers**				
Mean	12.46	32.55	37.60	36.41	54.46	14.67	23.46	49.91	25.09	51.95
Minimum	0.11	0.84	0.04	0.06	0.05	0.21	1.51	1.12	0.49	0.94
1st quartile	6.79	23.52	24.92	25.21	36.86	7.88	16.85	36.53	18.14	35.44
Median	10.68	30.62	35.87	33.33	50.19	12.28	21.96	48.33	23.64	48.40
3rd quartile	15.35	39.70	49.12	43.29	65.64	17.53	28.52	60.65	29.51	63.08
Maximum	112.01	88.93	111.91	126.74	166.54	117.65	85.21	125.35	96.52	144.16
SD	8.47	13.70	17.85	16.24	24.84	10.95	10.45	18.37	10.16	23.16
	Sebokeng					**Zamdela**				
Mean	13.43	26.34	52.68	30.53	43.85	21.49	26.05	46.98	30.10	58.54
Minimum	0.02	1.89	3.54	0.88	1.59	0.01	0.15	0.63	0.23	0.41
1st quartile	7.18	19.73	39.05	21.67	30.69	9.90	18.22	36.90	20.00	35.16
Median	11.41	24.67	51.37	28.73	40.92	18.18	24.49	46.29	27.73	51.70
3rd quartile	16.42	30.48	64.18	36.24	51.72	28.29	31.60	56.28	36.89	73.39
Maximum	115.67	97.82	147.45	119.78	162.74	105.09	91.50	108.67	108.80	213.11
SD	9.81	10.96	20.44	13.71	20.17	15.92	11.85	14.51	14.65	32.65

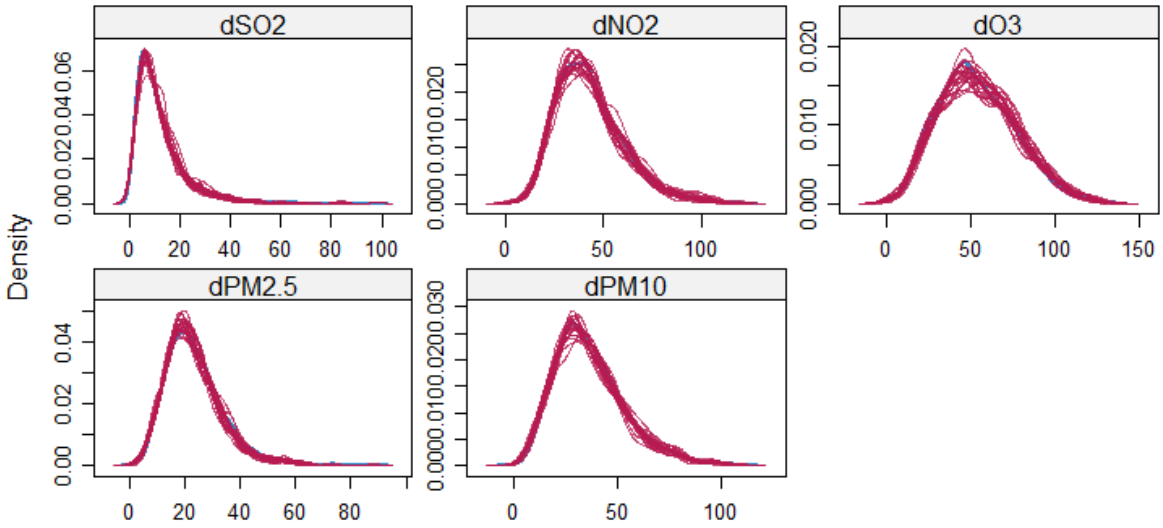


Figure 5.26: Mice imputations for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} density plots for Diepkloof.

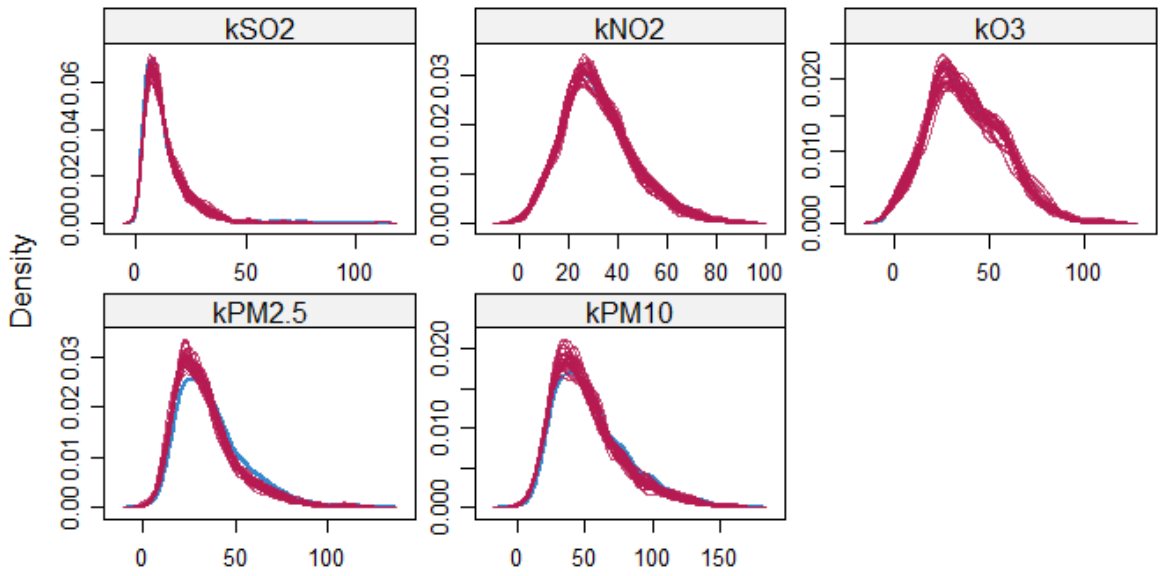


Figure 5.27: Mice imputations for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} density plots for Kliprivier.

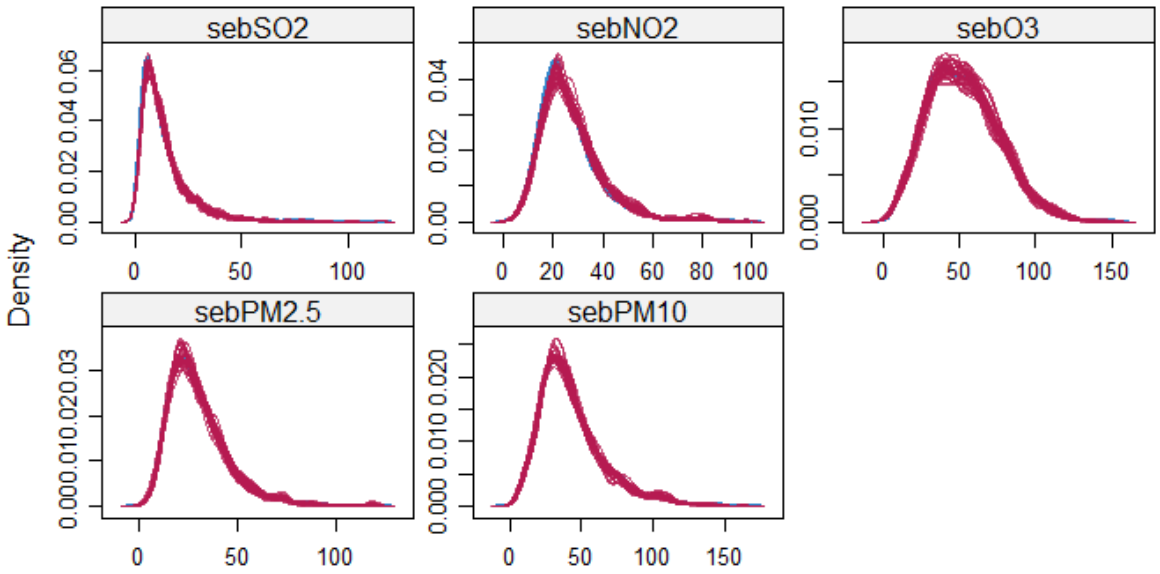


Figure 5.28: Mice imputations for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} density plots for Sebokeng.

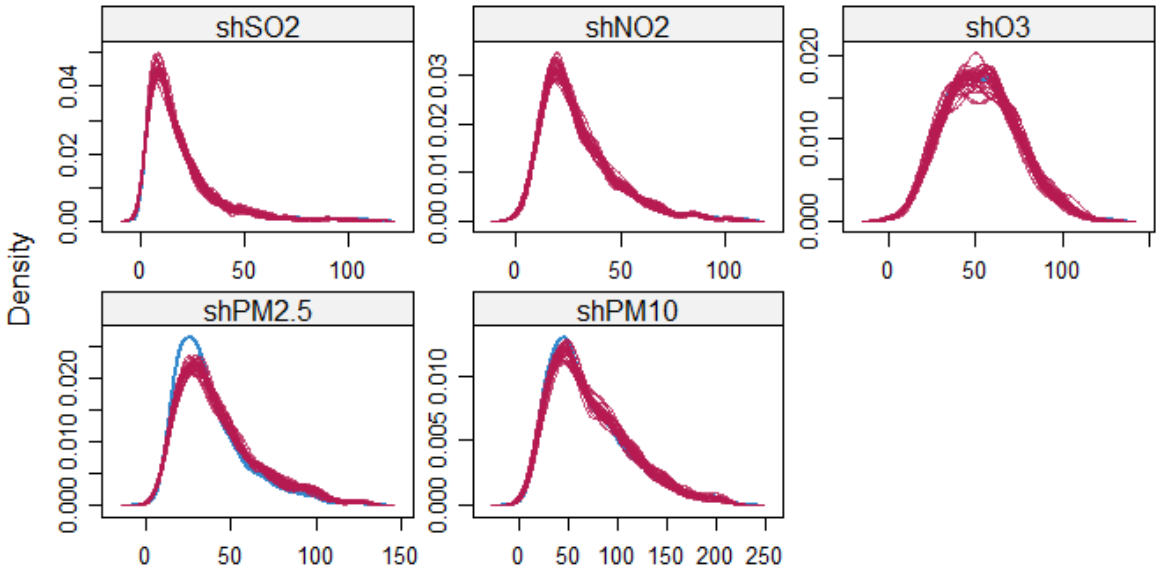


Figure 5.29: Mice imputations for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} density plots for Sharpeville.

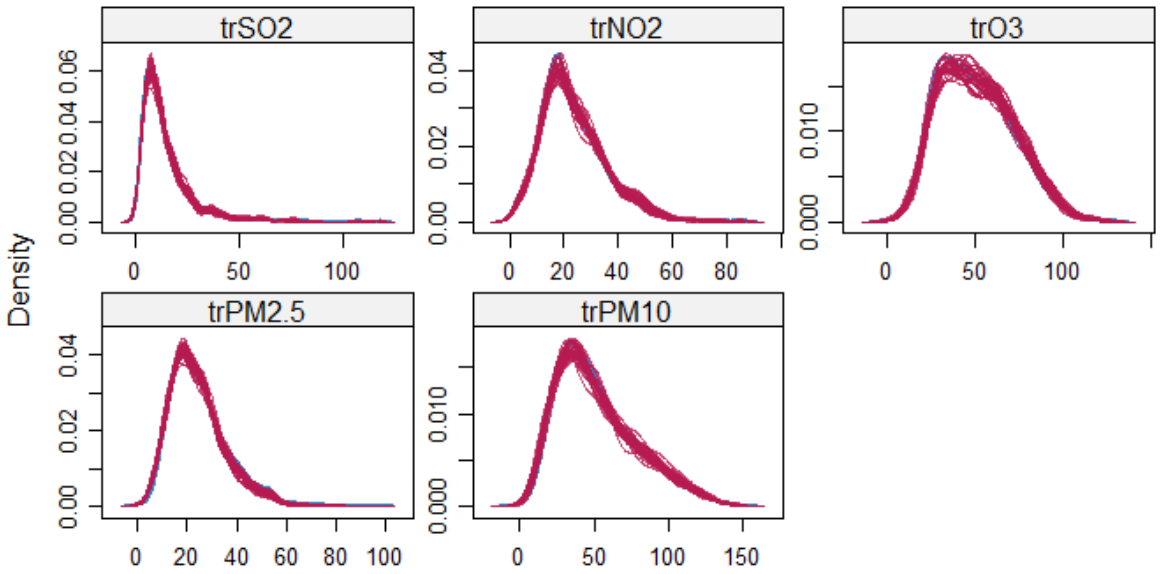


Figure 5.30: Mice imputations for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} density plots for Three Rivers.

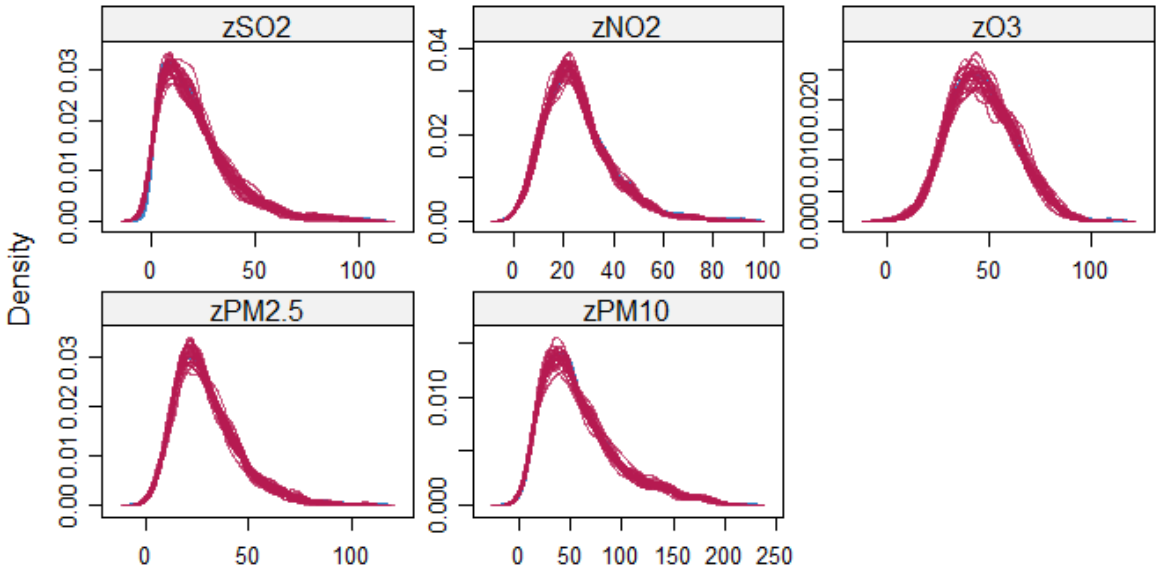


Figure 5.31: Mice imputations for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} density plots for Zamdela.

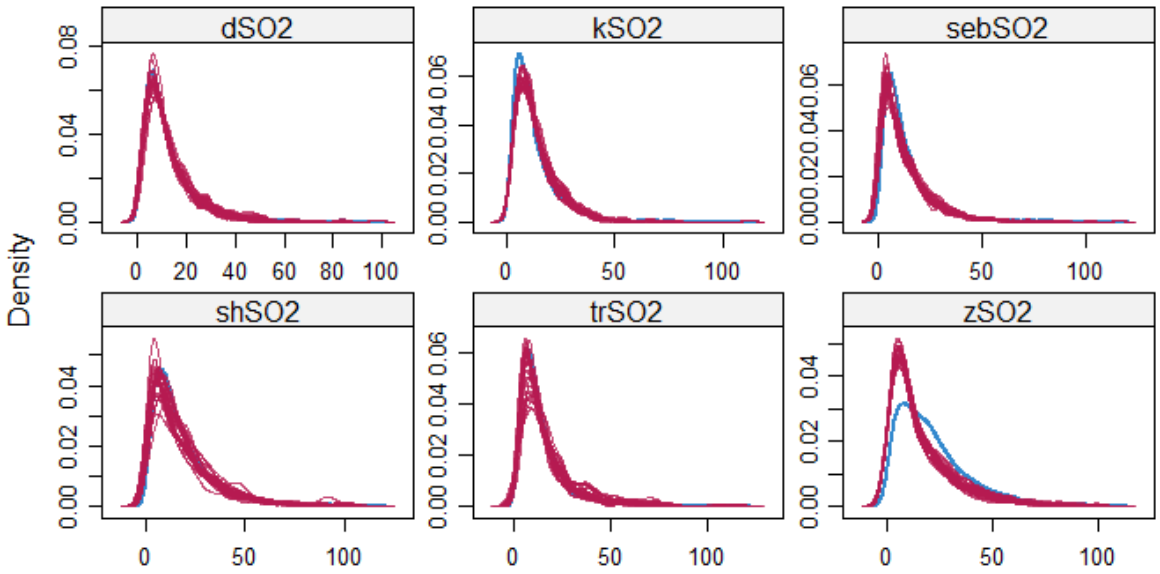


Figure 5.32: Density plots mice imputations SO_2 with meteorological data, for VTAPA stations.

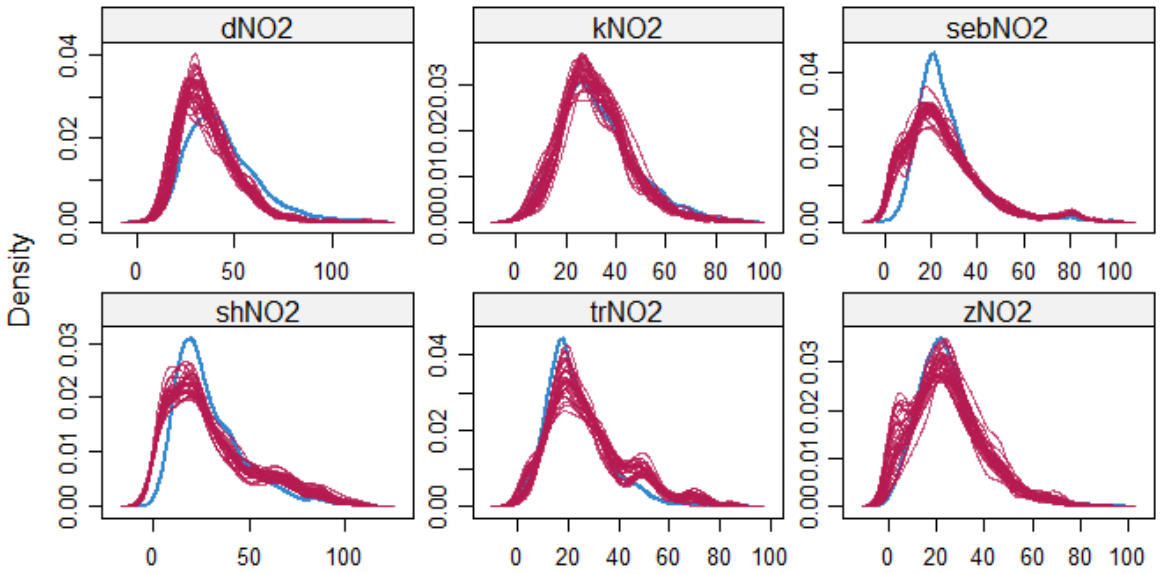


Figure 5.33: Density plots mice imputations NO_2, with meteorological data, for VTAPA stations.

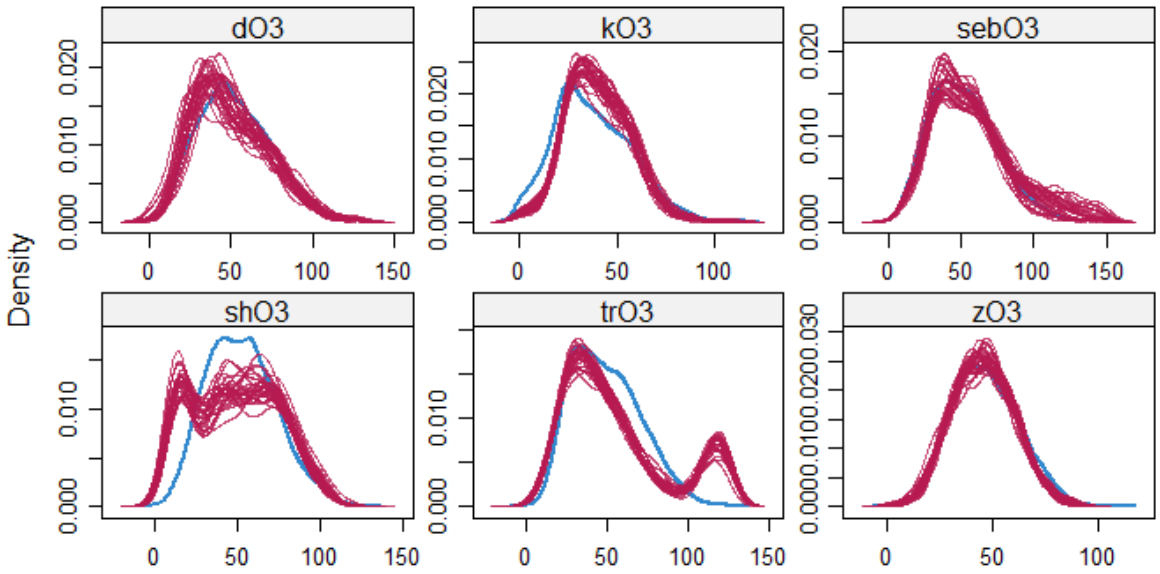


Figure 5.34: Density plots mice imputations O_3, with meteorological data, for VTAPA stations.

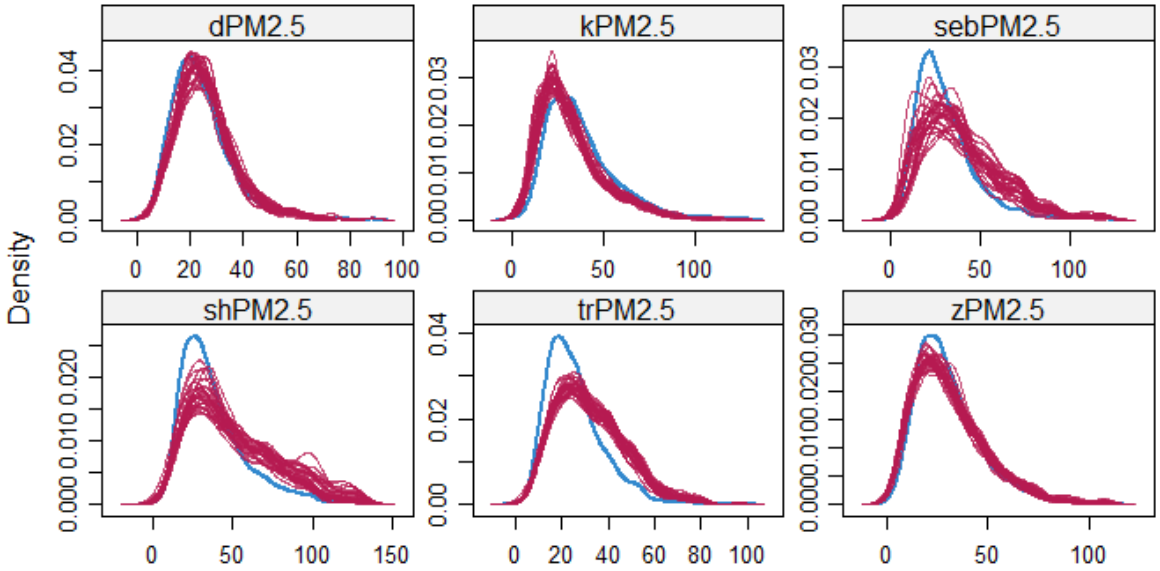


Figure 5.35: Density plots mice imputations $PM_{2.5}$, with meteorological data, for VTAPA stations.

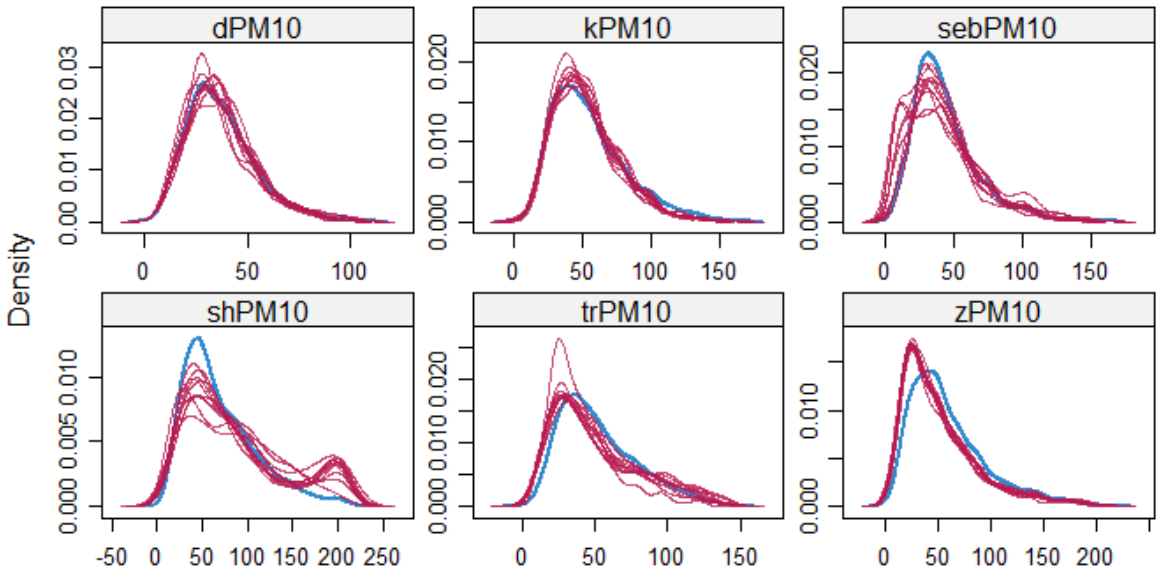


Figure 5.36: Density plots mice imputations PM10, with meteorological data, for VTAPA stations.

5.1.3.3. MULTIVARIATE TIME SERIES DATA IMPUTATION (MTSDI) IMPUTATION

Table 5.6 show the descriptive statistics after mtsdi imputation. The SO_2 mean concentration levels ranged from 12.29 µg/m^3 - 21.22 µg/m^3; NO_2 mean concentration levels ranged from 23.64 µg/m^3 - 42.27 µg/m^3; O_3 mean concentration levels ranged from 37.90 µg/m^3 - 52.82 µg/m^3; $PM_{2.5}$ mean concentration levels ranged from 23.93 µg/m^3 - 38.39 µg/m^3; and PM_{10} mean concentration levels ranged from 36.85 µg/m^3 - 56.6 µg/m^3.

Figures 5.37 to 5.41 show time-series of the imputed data sets using the mtsdi method. Figures 5.42 to 5.46 show mtsdi imputation that was run with additional meteorological data, i.e. temperature, relative humidity, and wind speed. The blue line is the level which shows the manner in which the data followed the time-series. The mtsdi imputation models were also run using the default smooth spline approach. Interestingly, the imputed data produced negative values.

Table 5.6: Descriptive statistics for the six monitoring stations in the VTAPA on SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, after mtsdi imputation (no meteorological data.

	dSO2	kSO2	sebSO2	shSO2	trSO2	zSO2
Mean	12.29	12.50	13.57	18.67	14.92	21.22
Minimum	0.08	0.11	-1.43	-3.24	-1.20	0.01
1st quartile	5.84	6.55	6.42	8.18	7.30	9.53
Median	9.94	10.35	10.63	14.48	11.76	17.29
3rd quartile	15.80	15.91	17.46	24.22	19.37	28.42
Maximum	98.04	112.01	115.67	113.12	117.65	105.09

	dNO2	kNO2	sebNO2	shNO2	trNO2	zNO2
Mean	42.27	32.78	26.66	29.88	23.64	26.22
Minimum	3.60	0.84	1.89	-1.41	1.51	0.15
1st quartile	30.33	23.42	19.02	17.41	15.77	18.82
Median	39.83	30.99	24.50	25.34	21.18	24.33
3rd quartile	51.54	40.43	31.59	39.12	29.74	31.92
Maximum	119.24	88.93	97.82	106.83	85.21	91.50

	dO3	kO3	sebO3	shO3	trO3	zO3
Mean	52.60	37.90	50.87	52.82	49.16	46.73
Minimum	3.00	0.04	3.54	0.20	1.12	0.63
1st quartile	35.36	24.43	33.66	36.80	32.91	35.58
Median	50.29	35.76	48.45	52.23	46.92	46.12
3rd quartile	67.57	50.59	64.96	66.96	63.29	57.44
Maximum	129.65	111.91	147.45	125.35	125.35	108.67

	dPM2.5	kPM2.5	sebPM2.5	shPM2.5	trPM2.5	zPM2.5
Mean	23.93	36.81	32.22	38.39	26.19	30.71
Minimum	2.10	-10.05	0.88	-4.28	0.49	0.23
1st quartile	17.05	25.18	21.08	24.10	18.08	19.64
Median	22.93	34.13	29.45	33.93	24.67	28.01
3rd quartile	28.80	44.58	40.63	48.28	31.90	38.49
Maximum	88.21	126.74	119.78	129.76	96.52	108.80

	dPM10	kPM10	sebPM10	shPM10	trPM10	zPM10
Mean	36.85	54.28	45.89	68.63	49.13	56.60
Minimum	0.19	0.05	1.59	-34.11	0.94	-5.88
1st quartile	25.91	35.42	30.44	39.51	30.46	31.50
Median	34.41	48.83	42.32	60.00	43.46	49.17
3rd quartile	45.65	67.93	56.07	90.65	62.72	73.38
Maximum	108.83	166.54	162.74	223.94	144.16	213.11

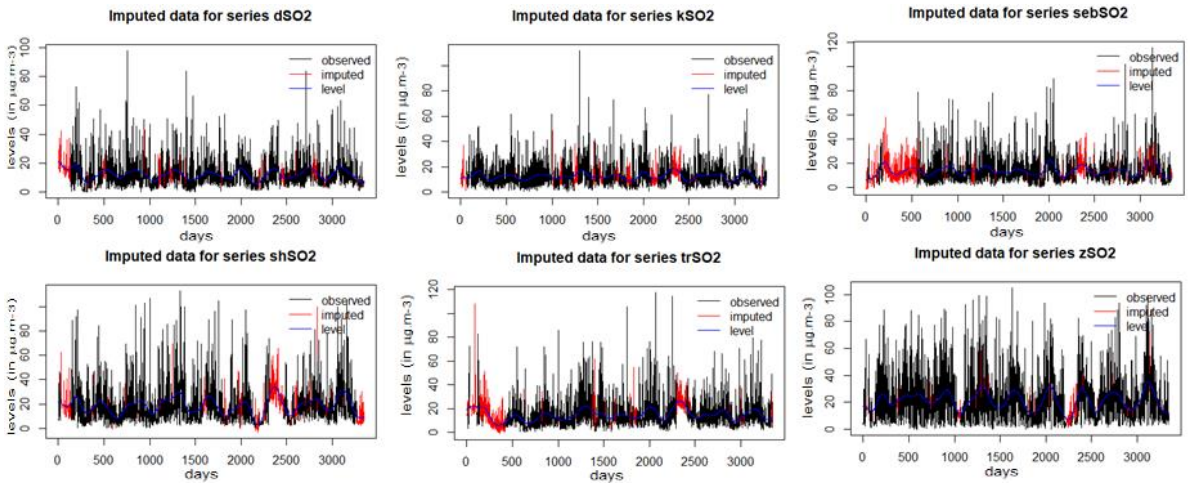


Figure 5.37: Illustration of mtsdi imputations for SO_2 concentrations ($\mu g/m^3$) at VTAPA six stations using the default smooth spline approach.

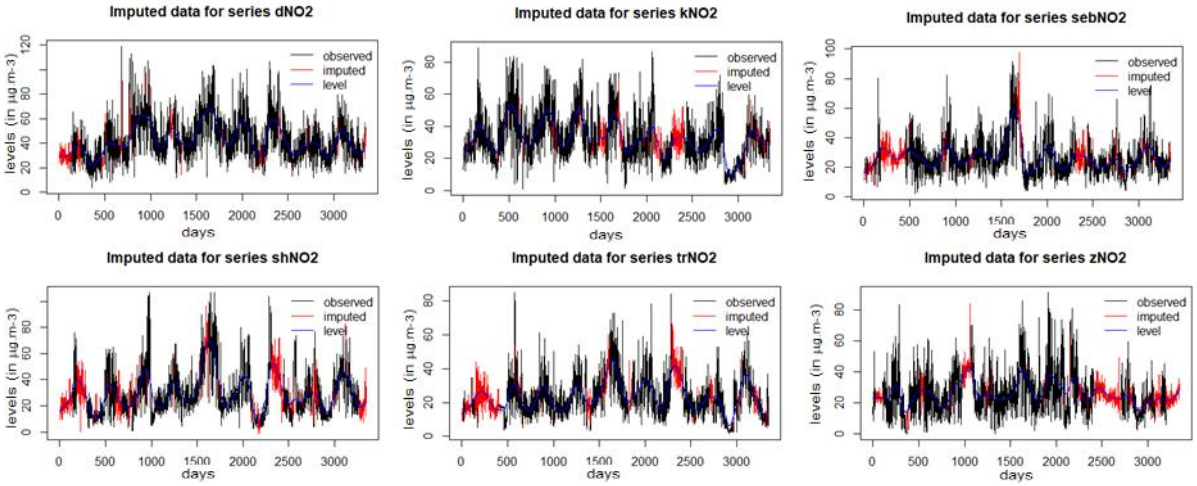


Figure 5.38: Illustration of mtsdi imputations no additional covariates for NO_2 concentrations ($\mu g/m^3$) at VTAPA six stations using the default smooth spline approach.

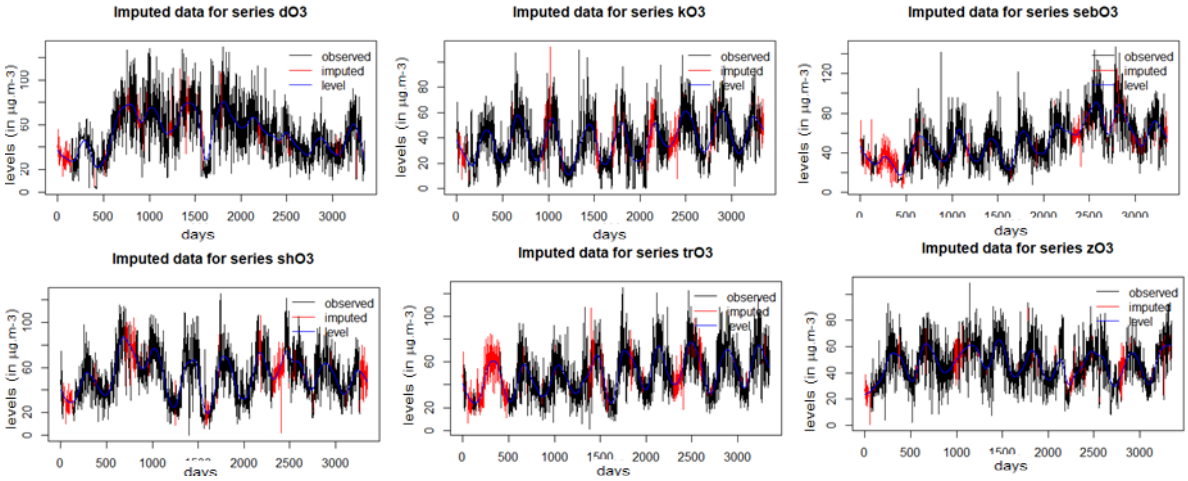


Figure 5.39: Illustration of mtsdi imputations no additional covariates for O_3 concentrations (µg/m^3) at VTAPA six stations using the default smooth spline approach.

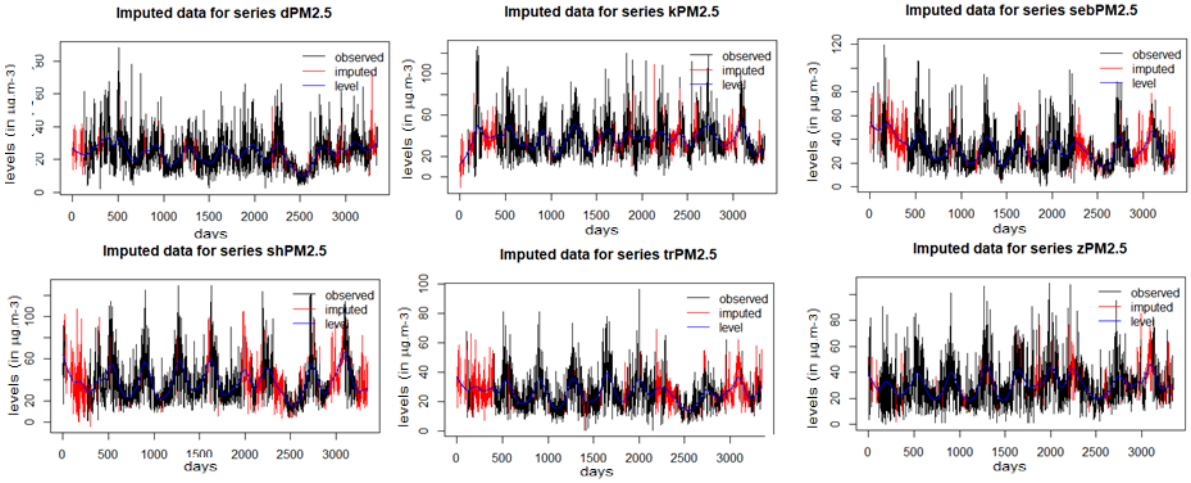


Figure 5.40: Illustration of mtsdi imputations no additional covariates for $PM_{2.5}$ concentrations (µg/m^3) at VTAPA six stations using the default smooth spline approach.

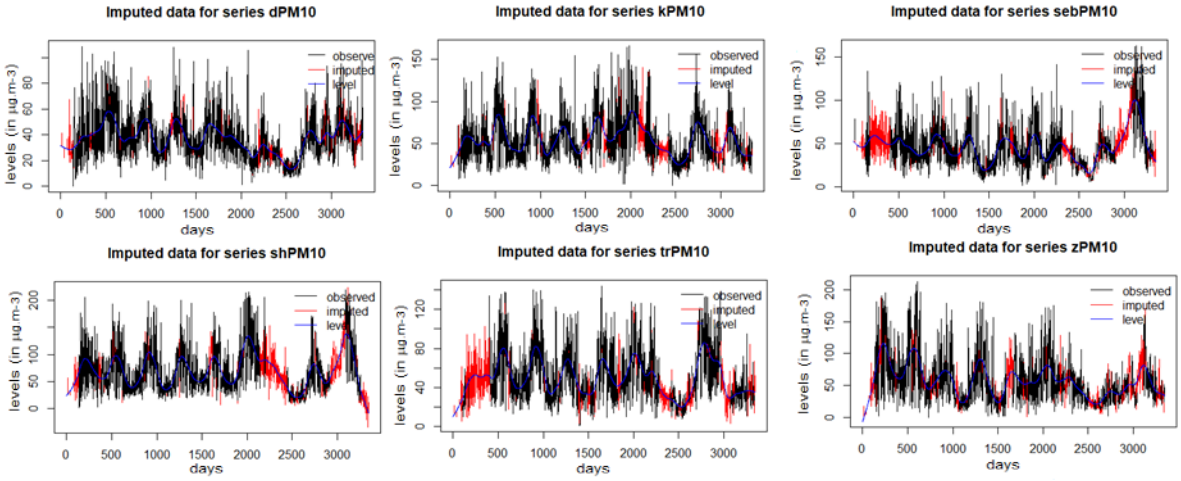


Figure 5.41: Illustration of mtsdi imputations no additional covariates for PM_{10} concentrations ($\mu g/m^3$) at VTAPA six stations using the default smooth spline approach.

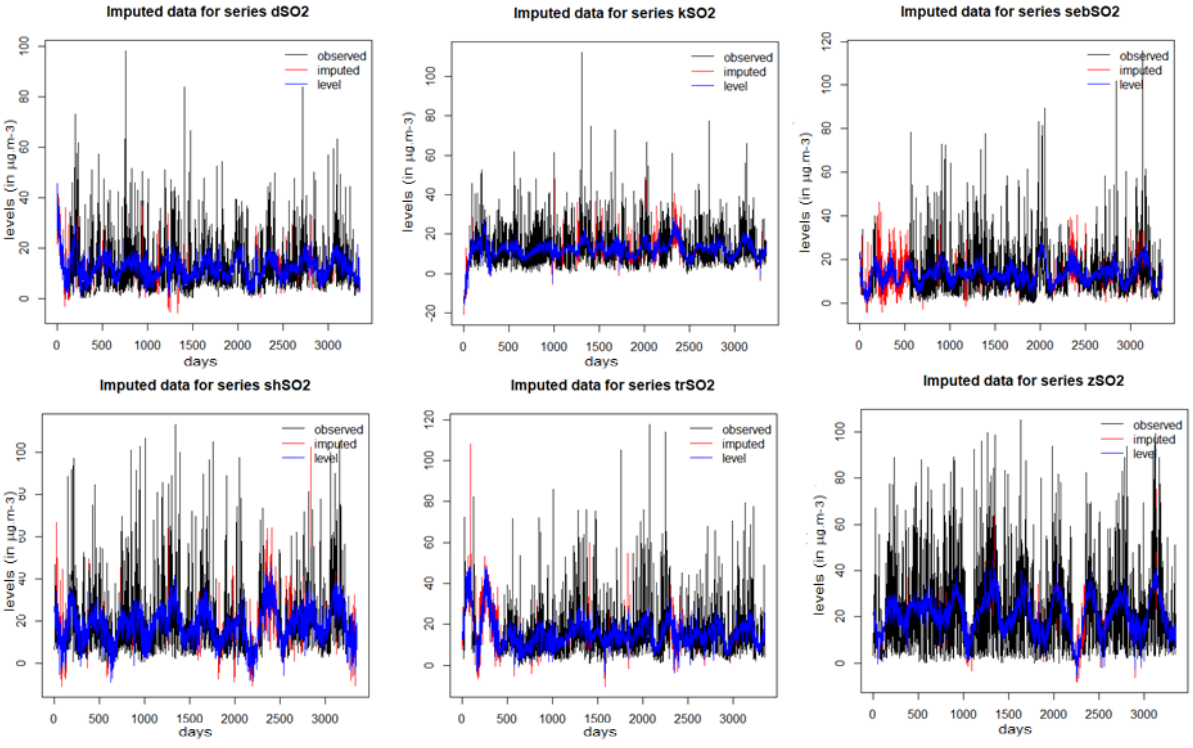


Figure 5.42: Illustration of mtsdi imputations with additional covariates for SO_2 concentrations ($\mu g/m^3$) at VTAPA six stations using the default smooth spline approach.

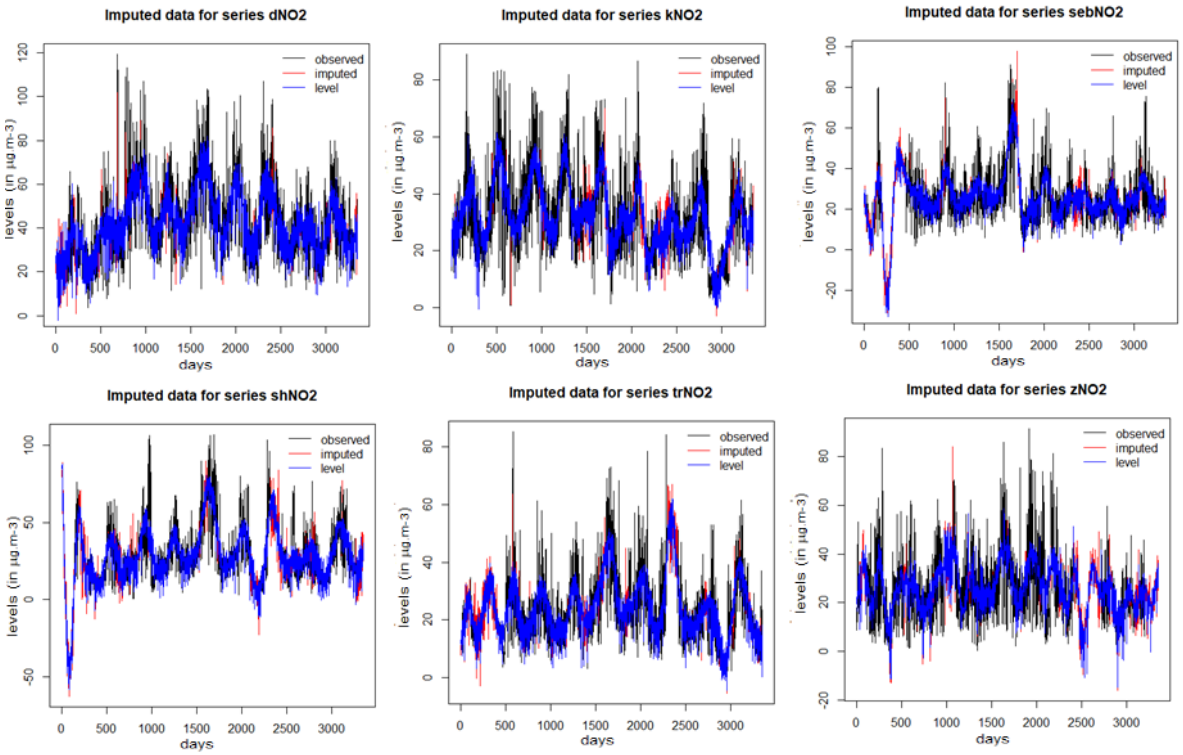


Figure 5.43: Illustration of mtsdi imputations with additional covariates for NO_2 concentrations (µg/m^3) at VTAPA six stations using the default smooth spline approach.

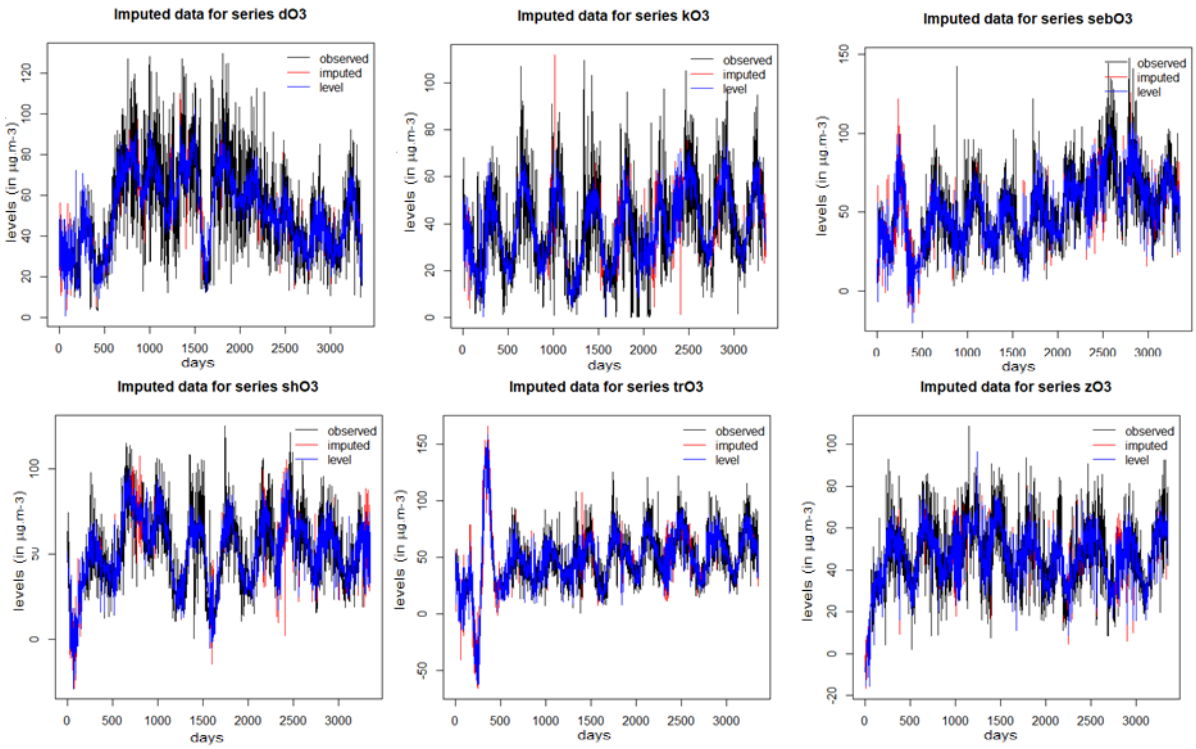


Figure 5.44: Illustration of mtsdi imputations with additional covariates for O_3 concentrations (µg/m^3) at VTAPA six stations using the default smooth spline approach.

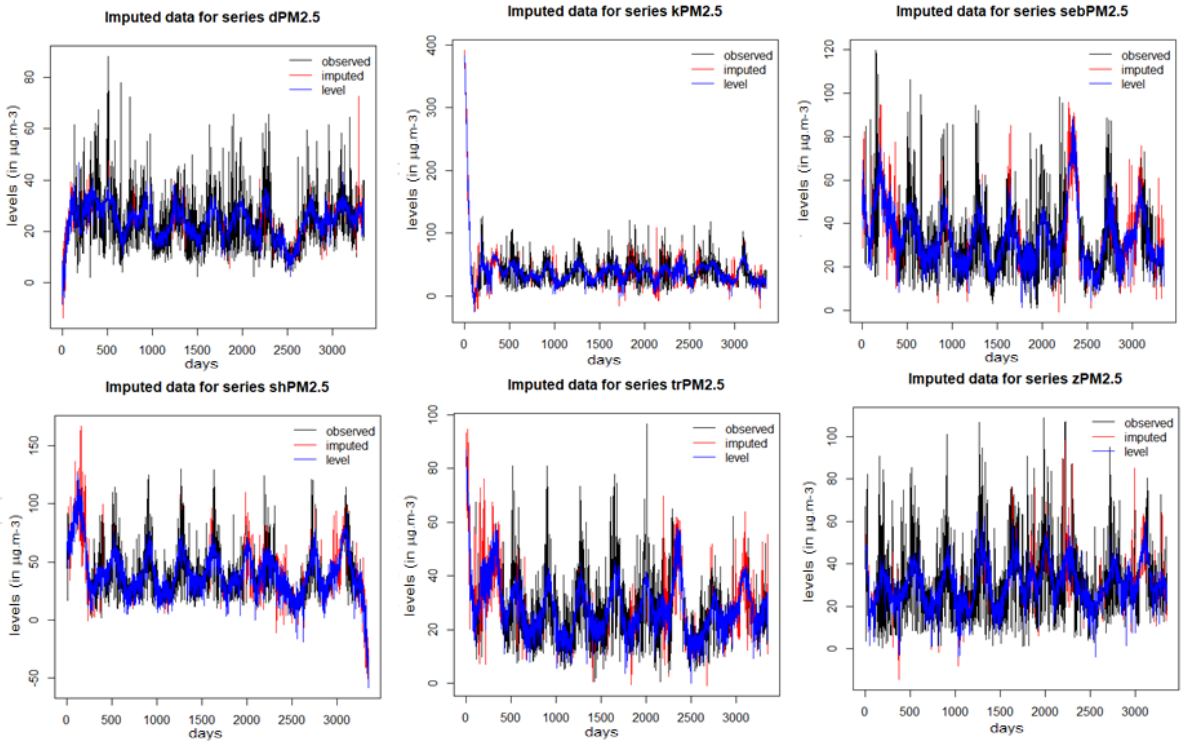


Figure 5.45: Illustration of mtsdi imputations for $PM_{2.5}$ concentrations (µg/m^3) at VTAPA six stations using the default smooth spline approach.

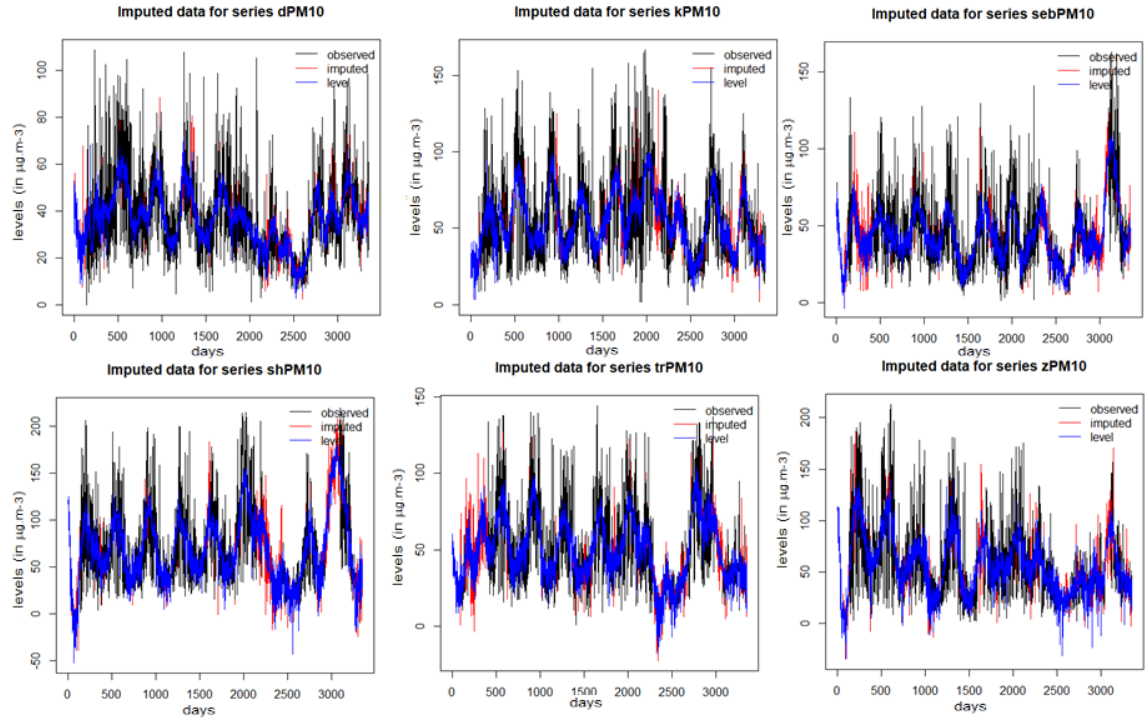


Figure 5.46: Illustration of mtsdi imputations for PM_{10} concentrations (µg/m^3) at VTAPA six stations using the default smooth spline approach.

5.1.4. OVERALL DATASETS

The imputed datasets for each pollutant were averaged across the six stations as final datasets for further analysis. Table 5.7 shows each averaged dataset from all imputation methods used. The descriptive statistics for all the datasets do not vary much from one another.

The mean SO_2 concentration levels ranged from 15.38 ± 8.21 µg/m^3 to 15.53 ± 9.06 µg/m^3, with maximums ranging from 63.57 to 69.27 µg/m^3. The NO_2 daily average concentration levels ranged from 29.82 ± 11.37 µg/m^3 to 30.24 ± 11.15 µg/m^3. The mean O_3 concentration levels ranged from 47.77 ± 15.22 µg/m^3 to 49.08 ± 12.71 µg/m^3. The $PM_{2.5}$ daily average concentration levels ranged from 30.84 ± 10.88 µg/m^3 to 32.78 ± 23.01 µg/m^{3,} with a maximum range of 80.61 to 118.33 µg/m^3. The estimated daily average concentration levels of PM_{10} ranged from 51.90 ± 22.48 to 53.13 ± 22.42 µg/m^3. The mice imputed dataset without the inclusion of meteorological conditions was the dataset used for classification and regression tree (CART) analyses.

Table 5.7: Descriptive statistics averaged SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, from the six monitoring stations in the VTAPA, after the 3 imputation methods, including mice and mtsdi methods when meteorological variables were added to imputations.

	SO_2	NO_2	O_3	$PM_{2.5}$	PM_{10}
Kalman imputation					
Mean	15.59	30.19	47.77	31.61	52.36
SD	8.21	10.52	15.22	11.05	20.60
Minimum	2.44	5.31	14.31	8.06	12.46
1st quartile	9.43	22.82	35.98	23.52	37.02
Median	13.72	28.11	46.33	30.09	49.06
3rd quartile	19.94	36.04	58.31	38.24	64.68
Maximum	63.57	80.81	103.91	80.61	131.75
Mice imputation no weather data used					
Mean	15.46	30.23	49.05	30.84	53.07
SD	7.95	9.85	13.68	10.88	19.50
Minimum	2.44	5.31	14.31	7.25	12.46
1st quartile	9.82	23.34	38.96	23.27	39.29
Median	13.67	28.62	47.99	29.13	49.78
3rd quartile	19.21	35.27	57.86	36.26	63.60
Maximum	63.57	80.81	103.91	80.61	131.75
Mice imputation with weather data used					
Mean	15.38	30.10	49.08	32.03	53.13
SD	8.91	11.32	12.71	15.89	22.42
Minimum	2.33	5.31	14.28	6.59	10.54
1st quartile	8.72	21.90	36.28	22.60	36.27
Median	13.22	27.87	48.52	29.91	49.05
3rd quartile	19.98	36.20	60.19	39.06	66.64
Maximum	63.67	80.81	103.91	84.44	131.75
Mtsdi imputation no weather data used					
Mean	15.53	30.24	48.34	31.38	51.90
SD	9.06	11.15	15.80	12.56	22.48
Minimum	2.17	5.31	9.45	6.12	5.87
1st quartile	8.72	22.16	35.92	22.01	34.94
Median	13.51	28.21	47.25	29.30	47.88
3rd quartile	20.25	36.27	59.31	38.18	65.20
Maximum	69.27	80.81	103.91	88.06	131.75
Mtsdi imputation with weather data used					
Mean	15.47	29.82	48.20	32.78	51.94
SD	8.85	11.37	16.04	23.01	14.29
Minimum	2.21	3.51	-6.66	4.04	-6.59
1st quartile	8.93	21.72	35.98	22.37	35.02
Median	13.46	27.90	47.70	30.14	48.53
3rd quartile	20.15	36.08	59.63	40.57	66.16
Maximum	68.25	80.81	103.91	118.66	131.75

SD-standard deviation

5.2. DISCUSSION

The aim of this section of the study was to prepare complete air pollution and meteorological datasets of the VTAPA for CART and unsupervised Machine Learning (ML) clustering analyses. Although mice is a commonly used imputation method, it was interesting to explore other imputation methods. Imputation can serve as a useful tool to researchers across multiple fields,[1-3] although it is not widely practiced across different diciplines.[4] It can benefit air pollution epidemiologists who often encounter missing air pollution data and experience challenges determining causal relationships with health outcomes.

Missing data are a recurring issue in environmental and occupational health studies, but imputation of missing data may assist in alleviating the limitations that can occur when using incomplete datasets.[5-7] The data in this study was assumed to be missing at random (MAR), because the missing values were not missing independent of themselves.[8] Nevertheless, the missing not at random (MNAR) assumption cannot be completely ruled out, as rigorous analysis must be conducted and, in the event of the data possibly being MNAR, some methods are still applied to treat the missing data as MAR.[4] The more difficult scenario would be having the similar or higher level of missingness under the MNAR.[6,9]

The data was also assumed to be MAR from the visualisation of the missing data. Here it can be seen that the missing values form clusters over the 3 347 observations, suggesting that during these times there could have been faults at the monitoring stations or maintenance may have occurred during those periods.[5-6,10] Once the MAR mechanism of missingness was assumed, the different methods of imputation were chosen to provide usable complete datasets that could undergo further analysis. Imputation methods are usually verified by artificially generating missing data and using performance metrics to compare the two datasets.[6,11] Therefore, a selection of imputation methods on real missing data could be subjective, dependent on the mechanism of missingness.

Due to the magnitude of all the missing data, i.e. the meteorological and air pollutant data of approximately ~10% to ~34%, imputation methods such as listwise deletion and complete case analysis were not considered. Although they are common methods

for handling data, they require the omission of records with missing values and continue to analyse the remaining data.[12] This approach is more suited towards data under the missing completely at random (MCAR) assumption with less than 5% of values missing, additionally these methods may lead to biased parameters and estimates within a dataset.[9,13] Univariate methods were equally inappropriate, although they are simple and fast to perform.[14-15] Regardless, these imputation algorithms lower the variability of the data, since the same value is imputed for all missing values. It creates a downward bias for the variance and also disturbs the covariance with the other variables in the set.[5,16]. Therefore, such imputation methods can yield biased estimates for almost all other parameters, excluding the mean or median, and if the data are not MCAR, even if the estimate of the mean could be highly biased.[17]

The South African Air Quality System (SAAQIS) under the Department of Forestry Fisheries and the Environment (DFFE) shows that there are thirteen air monitoring stations within the VTAPA. Six of the thirteen monitoring stations are serviced and maintained by the SAWS. These six monitoring stations, namely Diepkloof, Kliprivier, Sebokeng, Sharpeville, Three Rivers, and Zamdela, were included in the study because SAWS uses the New Zealand code of practice for data management.[18] The other seven stations within the area only had available data from 1 January 2019 to 29 February 2020 and did not have all the needed information for the five air pollutants of interest. Meteorological data such as relative humidity, temperature, and wind speed have been found to influence air pollution data.[19-24] Thus, the meteorological data was also imputed in order to see how it influences the multivariate imputations. Additionally, the imputed temperature, relative humidity and wind speed datasets were used as additional variables in the CART analysis.

Mice imputation has been more commonly used in air quality studies.[5-6,25-26] As a result, it was used to impute the meteorological missing data in the VTAPA. The number of imputations was dependent on the magnitude of missingness, researchers are recommended to use 20 imputations for 10-30% missing information, and 40 imputations for 50% missing information.[27] Although it has been argued that this approach is computationally taxing,[28] the proportion of missing data among all the

stations and variables did not exceed 30%, which was not computationally taxing with a model adjusted to allow for the proportion of missing data.

Although mice is a popular imputation method, two other imputation methods were applied to the air pollution datasets. Simple imputation methods – mean, median, random, and LOCF – cannot accurately impute data with a time-series nature.[14,29] This was clearly shown in figure 5.19, where the imputed data did not illustrate its variability. The Kalman algorithm, however, accounts for some of the variability of air pollution concentrations.[22,30-31] Hadeed et al,[6] found that the Kalman imputation performed well at 20-40% missing data. This imputation method uses more advanced algorithms and computation time.[14-15] It shows strong seasonality in the time-series nature of the dataset and yields good results.[32] The Kalman imputation is a favourable method to use as compared to the univariate methods. This is evident in the results showing that the Kalman imputation method did not disrupt the evident seasonal pattern of each pollutant. However, the Kalman filters used in the imputation algorithm reduces the variance within the dataset. As such, even though it does show seasonality, it does not show seasonal occurrences in a manner that is logically seen in air pollution data. The difference is apparent in comparison to the multivariate imputation methods. Nonetheless, the univariate time-series model did produce fairly reasonable results.

Multivariate imputation methods consider more complex aspects in the imputation process. It takes into consideration the correlation among the variables and accounts for the variability within the dataset.[11] The mtsdi imputation method was run under the EM algorithm and fitted with a smooth spline on each iteration. Junger and De Leon, 2015,[11] found that using the mtsdi method under the EM-ARIMA model performed best at 5% missing data, but the performance reduced at 40% missing data. This study used the default model, because using the default EM algorithm with normal multivariate data under the fitting of a spline yields precise estimates.[11] Each pollutant was imputed with the corresponding pollutant from a different area. For instance, PM_{10} was imputed from the six areas, and not from its PM_{10} relation to the other pollutants. Both area and its proximity to others were taken into consideration. The method can equally be used to consider other pollutants' influence on each other. The latter approach, however, produced models that would not converge and were omitted for the results.

The mtsdi imputation method does not rely on the variables' spatial dependency, but rather the linear dependency of the variables being imputed. The simple mtsdi method produced a fair overall dataset. The greatest strength of using the mtsdi method is that it takes into consideration the data variance and variability. However, once the meteorological conditions (i.e. wind speed, temperature, and relative humidity) were added as additional factors in the imputation runs, the imputed data did not show to be as uniform to the original dataset as compared to the data imputed without the weather variables. Unfortunately, some of the runs (e.g. O_3 imputations) did not converge well, even after as many as 1000 iterations, creating some exaggerated values. The imputed datasets produced negative concentration values that would not seem rational to use in any further analysis. While meteorological conditions are influential to air pollution concentration levels in these imputations, the factors showed to over-direct the figures. Perhaps the imputed meteorological data caused the over-direction of the air pollutant imputations.

Lastly, mice imputation was run on the different air pollutant datasets. Mice specifically operates under the MAR assumption of missing data and can produce biased results if used under MCAR or MNAR.[4] The mice predictive mean algorithm (PMM) was used, because the data met the requirement for all variables being continuous data. The PMM method draws imputations from the observed data, imputed values had the same gaps as in the original data, and were always within the range of the original dataset.[16] This method ensures that no imputations are outside the original data range and produces more reliable results.[33] PMM is also more robust to misspecification and interaction effects, therefore, it does not require an explicitly specified model from which imputations are generated.[17,34]

Mice often works by imputing via the relationship each variable has with the others and pools together to make one dataset.[4,35] In the imputation of the VTAPA data, the mice algorithm showed to be affected by extreme points that may have occurred in the observed dataset, but, overall, still produced reasonable results. Although the mice imputation method does not clearly show time variability of a dataset,[4] each run was able to compensate for the large proportions of missing data. When imputed with meteorological data, mice imputations showed that the imputed datasets did not run as well with the original dataset. However, overall datasets per air pollutant were

reasonable and produced slightly better results than the mtsdi method when meteorological variables were added. It should be mentioned that mice imputations have not always been found to perform as well as anticipated in some studies. Hadeed et al,[6] found that mice imputation did not perform well at 20-40% missing data. It is important to understand the data before applying any imputation method and consider that different extremes within a dataset can influence the performance of the method.

Overall, all the imputation methods produced reasonable imputed values. The multivariate imputation methods may be said to produce 'better' results, as they are able to capture the daily variance of each dataset. The most important goal of applying imputation methods on missing data are not just to recover missing values within a dataset, but to produce a valid analytical result in the presence of missing data.[36] Imputed values do not necessarily have to fall within plausible or possible ranges,[37] but this is dependent on the discipline and the level of missing data within the study.[36] The main problem of missing values within a dataset is that it can produce undesirable effects on the analyses of results, especially when it leads to biased parameter estimates.[13]

Completing the dataset can reduce potential bias that would have been created had the incomplete dataset been used for further epidemiological analysis.[5-6,16,38-39] The data management and cleaning done to the VTAPA dataset before imputation was a critical factor to consider. For this study, both daily and hourly data was compared to eliminate or replace outliers that were seen in the initial dataset downloaded from the SAAQIS website.[40] This was done to ensure that any wrongfully recorded extreme values found in the original dataset would not influence any of the imputation methods applied. The noted extreme values that occurred during 1 January 2011 to 29 February 2020 are suspected to have some influence of the imputed values during imputation of the multivariate imputations.

There are numerous advantages to correctly applied imputation methods. These include significantly reducing bias, increasing statistical power, and preservation of sample size.[5] It also provides robust estimations that are derived from the available data. Imputation, whether univariate or multivariate, is simple and relatively easy to implement.[5,14,29] Although there are noticeable strengths to imputation, there are also some limitations. Determining which imputation method to use on the missing data

can be challenging.[8] In some cases, data analyses must be done to determine which mechanism of missing data are to be assumed.[41-43] If an incorrect imputation method is used, it can produce underestimated or overestimated parameters and variability, which can increase bias within the dataset.[9] A disadvantage specific to using univariate imputation models is that they are time consuming to run when working with multiple datasets. Multivariate imputation, however, can demand far more computation power depending on the number of iterations and imputations requested.[44]

In conclusion, imputation is a useful tool that can be used in air pollution studies. By completing the air pollution and meteorological datasets, it provided complete and usable data for determining the joint effects of the air pollutants on hospital admissions using CART and unsupervised ML methods. Although both mice and mtsdi showed to be preferable methods, the final mice imputed datasets were used for further analyses.

5.3. REFERENCES

1. Arowosegbe OO, Röösli M, Künzli N, Saucy A, Adebayo-Ojo TC, Jeebhay MF, et al. Comparing methods to impute missing daily ground-level PM10 concentrations between 2010-2017 in South Africa. International Journal of Environmental Research and Public Health. 2021; 18(7) doi:10.3390/ijerph18073374.

2. Sweeney S, Vassall A, Guinness L, Siapka M, Chimbindi N, Mudzengi D, et al. Examining approaches to estimate the prevalence of catastrophic costs due to tuberculosis from small-scale studies in South Africa. PharmacoEconomics. 2020; 38(6):619-31. doi:10.1007/s40273-020-00898-3.

3. Vermaak C. Tracking poverty with coarse data: Evidence from South Africa. The Journal of Economic Inequality. 2012; 10(2):239-65. doi:10.1007/s10888-011-9211-2.

4. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: What is it and how does it work? International Journal of Methods in Psychiatric Research. 2011; 20(1):40-9. doi:10.1002/mpr.329.

5. Gómez-Carracedo MP, Andrade JM, López-Mahía P, Muniategui S, Prada D. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemometrics and Intelligent Laboratory Systems. 2014; 134:23-33. doi:10.1016/j.chemolab.2014.02.007.

6. Hadeed SJ, O'Rourke MK, Burgess JL, Harris RB, Canales RA. Imputation methods for addressing missing data in short-term monitoring of air pollutants. Science of the Total Environment. 2020; 730 doi:10.1016/j.scitotenv.2020.139140.

7. Junninen H, Niska H, Tuppurainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality data sets. Atmospheric Environment. 2004; 38(18):2895-907. doi:10.1016/j.atmosenv.2004.02.026.

8. Rubin DB. Inference and missing data. Biometrika. 1976; 63(3) doi:10.2307/2335739.

9. Kang H. The prevention and handling of the missing data. Korean Journal of Anesthesiology. 2013; 64(5):402-6. doi:10.4097/kjae.2013.64.5.402.

10. Quinteros ME, Lu S, Blazquez C, Cárdenas-R JP, Ossa X, Delgado-Saborit JM, et al. Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile. Atmospheric Environment. 2019; 200:40–9.

11. Junger WL, De Leon AP. Imputation of missing data in time series for air pollutants. Atmospheric Environment. 2015; 102:96-104. doi:10.1016/j.atmosenv.2014.11.049.

12. Somasundaram R, Nedunchezhian R. Evaluation of three simple imputation methods for enhancing preprocessing of data with missing values. International Journal of Computer Applications. 2011; 21(10):14-9.

13. Abidin N, Ismail A, Emran N. Performance analysis of machine learning algorithms for missing value imputation. International Journal of Advanced Computer Science and Applications. 2018; 9(6). doi:10.14569/IJACSA.2018.090660.

14. Moritz S, Bartz-Beielstein T. Imputets: Time series missing value imputation in R. RJ. 2017; 9(1):207. doi:10.32614/RJ-2017-009.

15. Moritz S. Package imputets,. 2017. Available from: http://cran.r-project.org/web/packages/imputeTS/imputeTS.pdf.

16. Karahalios A, Baglietto L, Carlin JB, English DR, Simpson JA. A review of the reporting and handling of missing data in cohort studies with repeated assessment of exposure measures. BMC Medical Research Methodology. 2012; 12(1):1-10. doi:10.1186/1471-2288-12-96.

17. Van Buuren S. Flexible imputation of missing data. Boca Raton, FL: CRC Press; 2012.

18. South African Weather Service. Vaal triangle priority network monthly report – October 2019. 2019. AQI-VTPA-MER-2019-OCTOBER-001.

19. Chen K, Glonek G, Hansen A, Williams S, Tuke J, Salter A, et al. The effects of air pollution on asthma hospital admissions in Adelaide, South Australia, 2003-2013: Time-series and case-crossover analyses. Clinical & Experimental Allergy. 2016; 46(11):1416-30. doi:10.1111/cea.12795.

20. Chen M, Wang P, Chen Q, Wu J, Chen X. A clustering algorithm for sample data based on environmental pollution characteristics. Atmospheric Environment. 2015; 107:194-203. doi:10.1016/j.atmosenv.2015.02.042.

21. Fayiga AO, Ipinmoroti MO, Chirenje T. Environmental pollution in Africa. Environment, Development and Sustainability. 2018; 20(1):41-73. doi:10.1007/s10668-016-9894-4.

22. Poole JA, Barnes CS, Demain JG, Bernstein JA, Padukudru MA, Sheehan WJ, et al. Impact of weather and climate change with indoor and outdoor air quality in asthma: A work group report of the AAAAI environmental exposure and respiratory health committee. Journal of Allergy and Clinical Immunology. 2019; 143(5):1702-10. doi:10.1016/j.jaci.2019.02.018.

23. Taghvaee S, Sowlat MH, Mousavi A, Hassanvand MS, Yunesian M, Naddafi K, et al. Source apportionment of ambient PM2.5 in two locations in Central Tehran using the positive matrix factorization (PMF) model. Science of the Total Environment. 2018; 628-629:672-86. doi:10.1016/j.scitotenv.2018.02.096.

24. Tshehla C, Djolov G. Source profiling, source apportionment and cluster transport analysis to identify the sources of PM and the origin of air masses to an industrialised rural area in Limpopo. Clean Air Journal. 2018; 28(2):54-66. doi:10.17159/2410-972X/2018/v28n2a18.

25. Hajmohammadi H, Heydecker B. Multivariate time series modelling for urban air quality. Urban Climate. 2021; 37 doi:10.1016/j.uclim.2021.100834.

26. Van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011; 45(1):1-67.

27. Allison P. Why you probably need more imputations than you think. 2012. Available from: https://statisticalhorizons.com/more-imputations.

28. Graham JW, Olchowski AE, Gilreath TD. How many imputations are really needed? Some practical clarifications of multiple imputation theory. Prevention Science. 2007; 8(3):206-13. doi:10.1007/s11121-007-0070-9.

29. Moritz S, Sardá A, Bartz-Beielstein T, Zaefferer M, Stork Jr. Comparison of different methods for univariate time series imputation in R. 2015.

30. Hou X, Fei D, Kang H, Zhang Y, Gao J. Seasonal statistical analysis of the impact of meteorological factors on fine particle pollution in China in 2013–2017. Natural Hazards. 2018; 93:677–98. doi:10.1007/s11069-018-3315-y.

31. Sanguineti PB, Lanzaco BL, López ML, Achad M, Palancar GG, Olcese LE, et al. PM2.5 monitoring during a 10-year period: Relation between elemental concentration and meteorological conditions. Environmental Monitoring and Assessment. 2020; 192(5):313. doi:10.1007/s10661-020-08288-0.

32. Cichowicz R, Wielgosiński G, Fetter W. Dispersion of atmospheric air pollution in summer and winter season. Environmental Monitoring and Assessment : An International Journal Devoted to Progress in the Use of Monitoring Data in Assessing Environmental Risks to Man and the Environment. 2017; 189(12):1-10. doi:10.1007/s10661-017-6319-2.

33. Chinomona A, Mwambi H. Multiple imputation for non-response when estimating HIV prevalence using survey data. BMC Public Health. 2015; 15(1):1059.

34. Randahl D. Raoul: An R-package for handling missing data. Uppsala universitet, Statistiska institutionen; 2016.

35. Liu Y, Brown SD. Comparison of five iterative imputation methods for multivariate classification. Chemometrics and Intelligent Laboratory Systems. 2013; 120:106-15. doi:10.1016/j.chemolab.2012.11.010.

36. Nguyen CD, Carlin JB, Lee KJ. Model checking in multiple imputation: An overview and case study. Emerging Themes in Epidemiology. 2017; 14(1):8. doi:10.1186/s12982-017-0062-6.

37. Rodwell L, Lee KJ, Romaniuk H, Carlin JB. Comparison of methods for imputing limited-range variables: A simulation study. BMC Medical Research Methodology. 2014; 14:57. doi:10.1186/1471-2288-14-57.

38. Chen J, Hunter S, Kisfalvi K, Lirio RA. A hybrid approach of handling missing data under different missing data mechanisms: Visible 1 and varsity trials for ulcerative colitis. Contemporary Clinical Trials. 2021; 100 doi:10.1016/j.cct.2020.106226.

39. Lupattelli AP, Wood MEP, Nordeng HP. Analyzing missing data in perinatal pharmacoepidemiology research: Methodological considerations to limit the risk of bias. Clinical Therapeutics. 2019; 41(12):2477-87. doi:10.1016/j.clinthera.2019.11.003.

40. South African Air Quality Information System. 2022. Available from: http://saaqis.environment.gov.za/.

41. Lee KJ, Tilling KM, Cornish RP, Little RJA, Bell ML, Goetghebeur E, et al. Framework for the treatment and reporting of missing data in observational studies: The treatment and reporting of missing data in observational studies framework. Journal of Clinical Epidemiology. 2021; 134:79-88. doi:10.1016/j.jclinepi.2021.01.008.

42. Li C. Little's test of missing completely at random. The Stata Journal. 2013; 13(4):795-809. doi:10.1177/1536867X1301300407.

43. Little RJA, Rubin DB. Statistical analysis with missing data. Newy York, United States: John Wiley & Sons, Incorporated; 2002.

44. van Smeden M, Groenwold RHH, Moons KGM. A cautionary note on the use of the missing indicator method for handling missing data in prediction research. Journal of Clinical Epidemiology. 2020; 125:188-90. doi:10.1016/j.jclinepi.2020.06.007.

CHAPTER 6: THE ASSOCIATION OF JOINT EFFECTS OF AIR POLLUTION ON RESPIRATORY AND CARDIOVASCULAR DISEASE HOSPITAL ADMISSIONS

This chapter presents the results of the CART analysis applied to determine the joint effect of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} on respiratory and cardiovascular hospital admissions in Vereeniging and Vanderbijlpark, in the Vaal Triangle Airshed Priority Area (VTAPA), from January 2011 to February 2020.

6.1. AIR POLLUTION AND WEATHER CONDITIONS

Table 6.1 summarises the daily air pollutants and weather conditions observed in the Vaal Triangle Airshed Priority Area (VTAPA) from 2 January 2011 to 29 February 2020, after mice imputation was applied, as discussed in Chapter 5. The air and meteorological data are described after being set at a two-day cumulative lag.

Table 6.1: Summary statistics of daily air pollutants and meteorological conditions in VTPA, South Africa, 2 January 2011 – 29 February 2020 (3346 days).

Variable	Mean	Min	P25	Median	P75	Max
SO_2 ($\mu g.m^{-3}$)	15.45	2.44	9.83	13.67	19.17	63.57
NO_2 ($\mu g.m^{-3}$)	30.24	5.31	23.34	28.64	35.35	80.81
O_3 ($\mu g.m^{-3}$)	49.04	14.31	38.93	48.02	57.87	103.91
$PM_{2.5}$ ($\mu g.m^{-3}$)	30.84	7.25	23.31	29.13	36.28	80.61
PM_{10} ($\mu g.m^{-3}$)	59.07	12.46	39.29	49.81	63.47	131.75
Relative Humidity (%)	49.57	13.10	40.79	50.06	58.54	85.06
Temperature (°C)	16.97	0.35	14.02	17.55	20.29	29.75
TAPP (°C)	15.46	-2.19	12.48	16.09	18.86	28.10

Abbreviations: SO_2: sulphur dioxide; NO_2: nitrogen dioxide; O_3: Ozone; $PM_{2.5}$: particulate matter with an aerodynamic diameter of less than 2.5 μm; PM_{10}: particulate matter with an aerodynamic diameter of less than 10 μm; Tapp: apparent temperature; P25: 25th percentile; P75: 75th percentile

The daily average concentration level on PM_{10} was 59.07 μg/m^3, the highest among the five pollutants. Approximately 14% (475/3 346) of the data exceeded the South African (SA) standard and 61.7% (2065/3 346) of the WHO guideline. The daily maximum concentration level was 131.72 μg/m^3. The $PM_{2.5}$ daily average concentration was 30.84 μg/m^3 with a minimum and maximum of 7.25 μg/m^3 and 80.61 μg/m^3, respectively. Approximately 17.5% (587/3 346) of days had concentrations

above the 24-hour average NAAQS and 97% (3262/3 346) exceeded the twenty-four-hour average concentrations in the WHO guidelines. The NO_2 average daily concentration was 30.24 µg/m^3, ~14% (488/3 346) of the 24-hour average NAAQS, and 67.5% (2260/3 346) above the WHO 24-hour average guideline. The mean SO_2 concentration was 15.45 µg/m^{3,} with 1.46% (46/3 346) of the days above the WHO 24-hour average guideline, and no daily average concentrations above the NAAQS. O_3 had a mean concentration level of 49.04 µg/m^3.

The daily average apparent temperature (Tapp) measured was 15.46°C and temperature it was 16.97ºC. The minimum and maximum Tapp was -2.19°C and 28.10ºC, respectively. Temperature had a minimum and maximum were 0.35°C and maximum 29.75°C. The relative humidity ranged between 13.10% and 85.06%, with an average of 49.57%.

Figures 6.1 to 6.5 illustrate the time-series of the 24-hour concentration averages of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, during the study period. Kruskal Wallis tests showed significant differences in seasons for all five air pollutants (p-value < 0.001). For SO_2, NO_2, $PM_{2.5}$, and PM_{10}, higher concentrations were generally seen in winter (June to August) and lower concentrations in summer (December to February). However, for O_3, the concentration levels were generally higher in summer and lower in winter.

Figures 6.6 to 6.7 illustrate the time-series of the 24-hour averages of relative humidity, Tapp, and temperature during the study period. Seasonal trends can be seen for relative humidity, Tapp, and temperature. The relative humidity dropped in the colder months, May to August, and peaked during the warmer months, September to April. Tapp and temperature (°C) peaked in the summer months, September to April, and declined in the cold months, June to August, which is commonly experienced in South Africa.

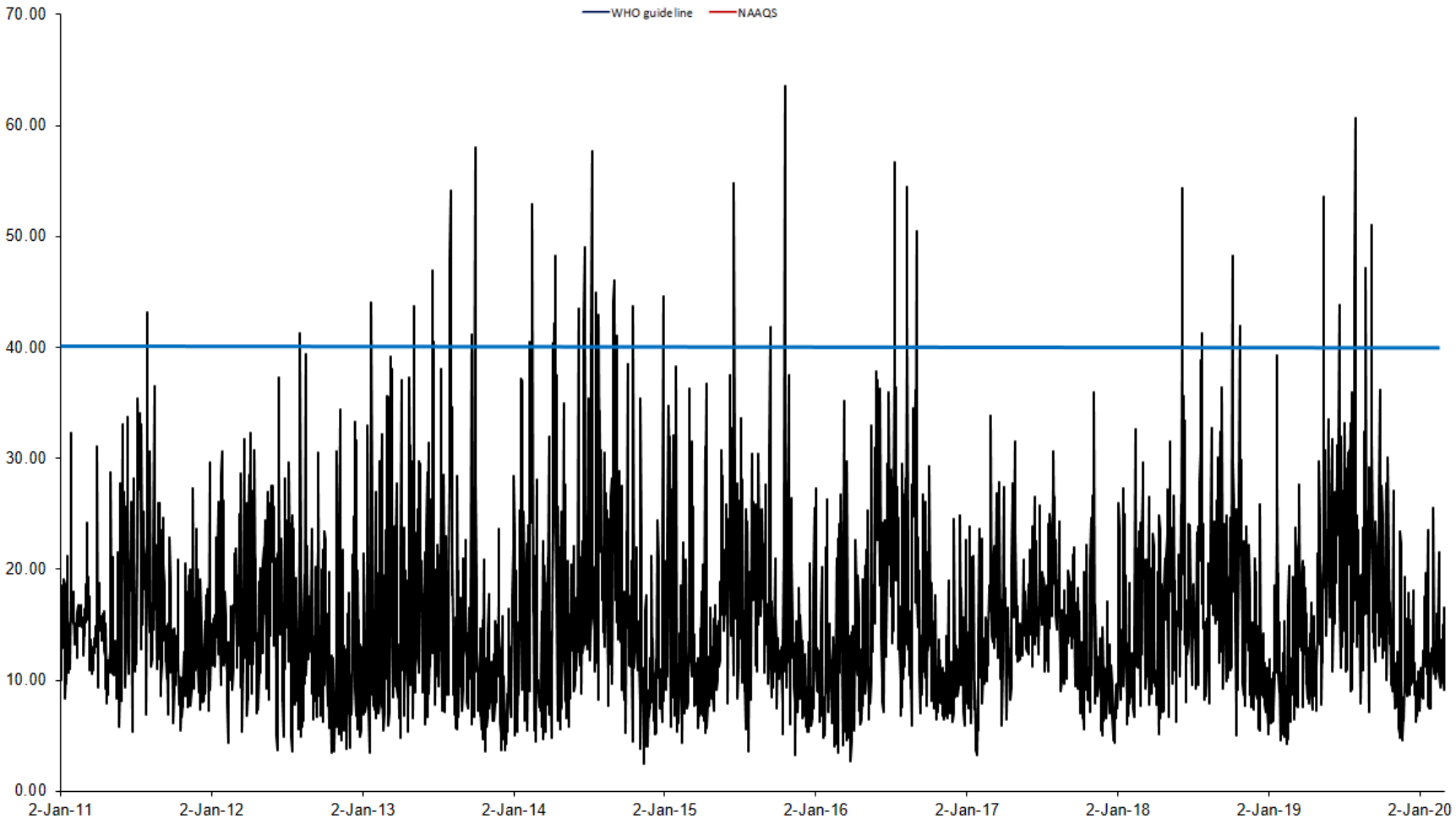


Figure 6.1: Time-series of SO_2 levels in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

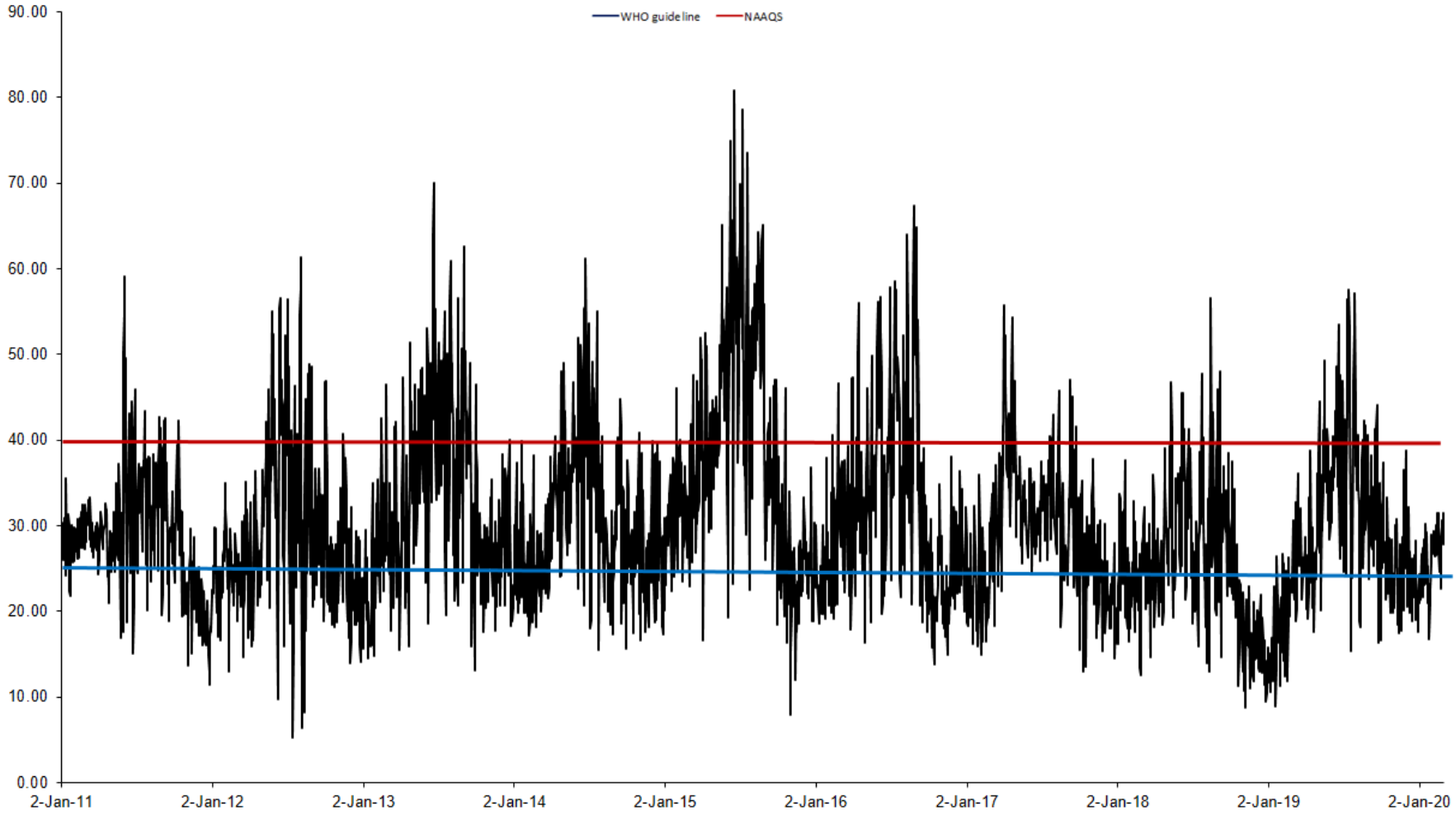


Figure 6.2: Time-series of NO_2 levels in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

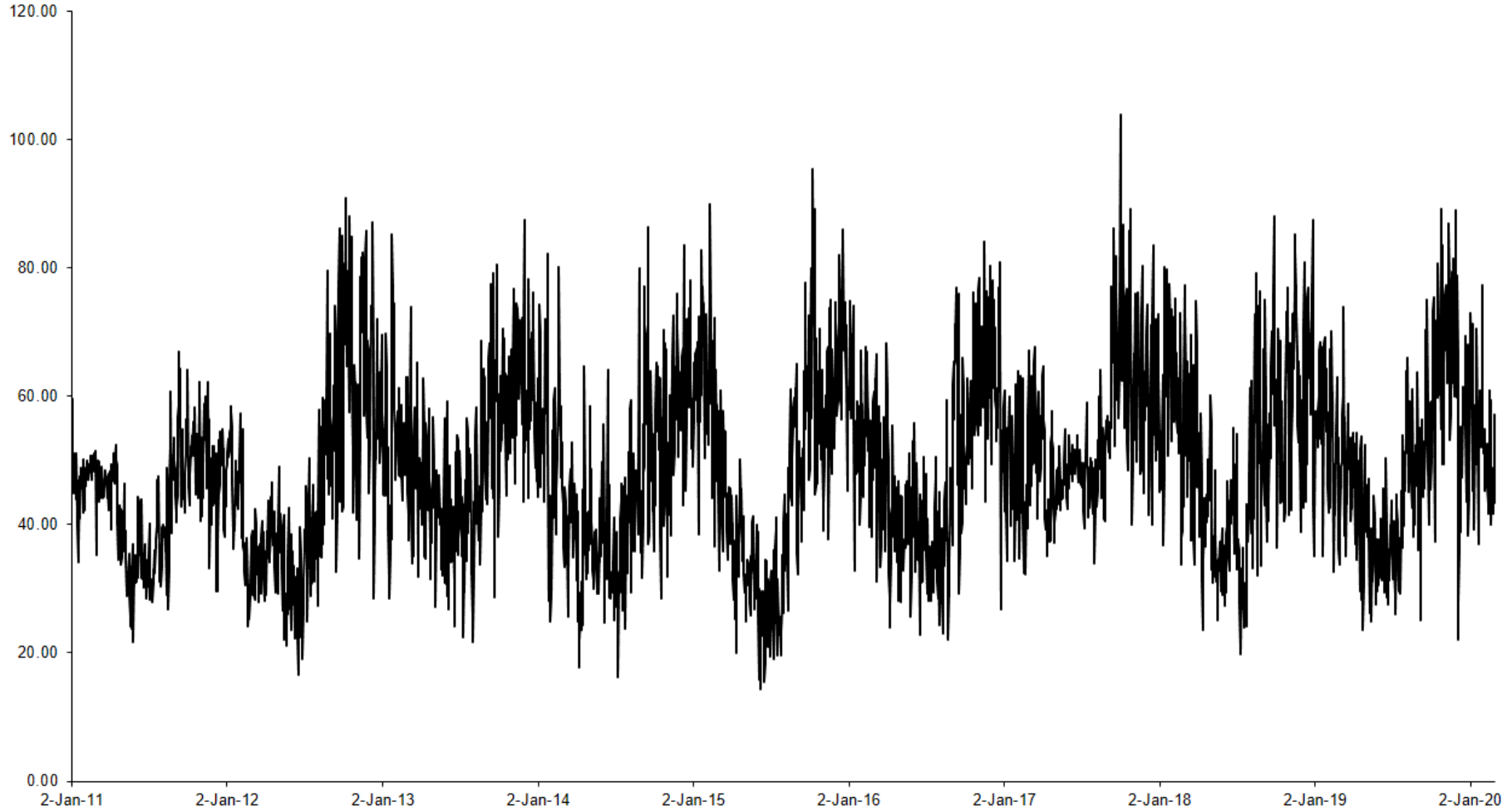


Figure 6.3: Time-series of O_3 levels in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

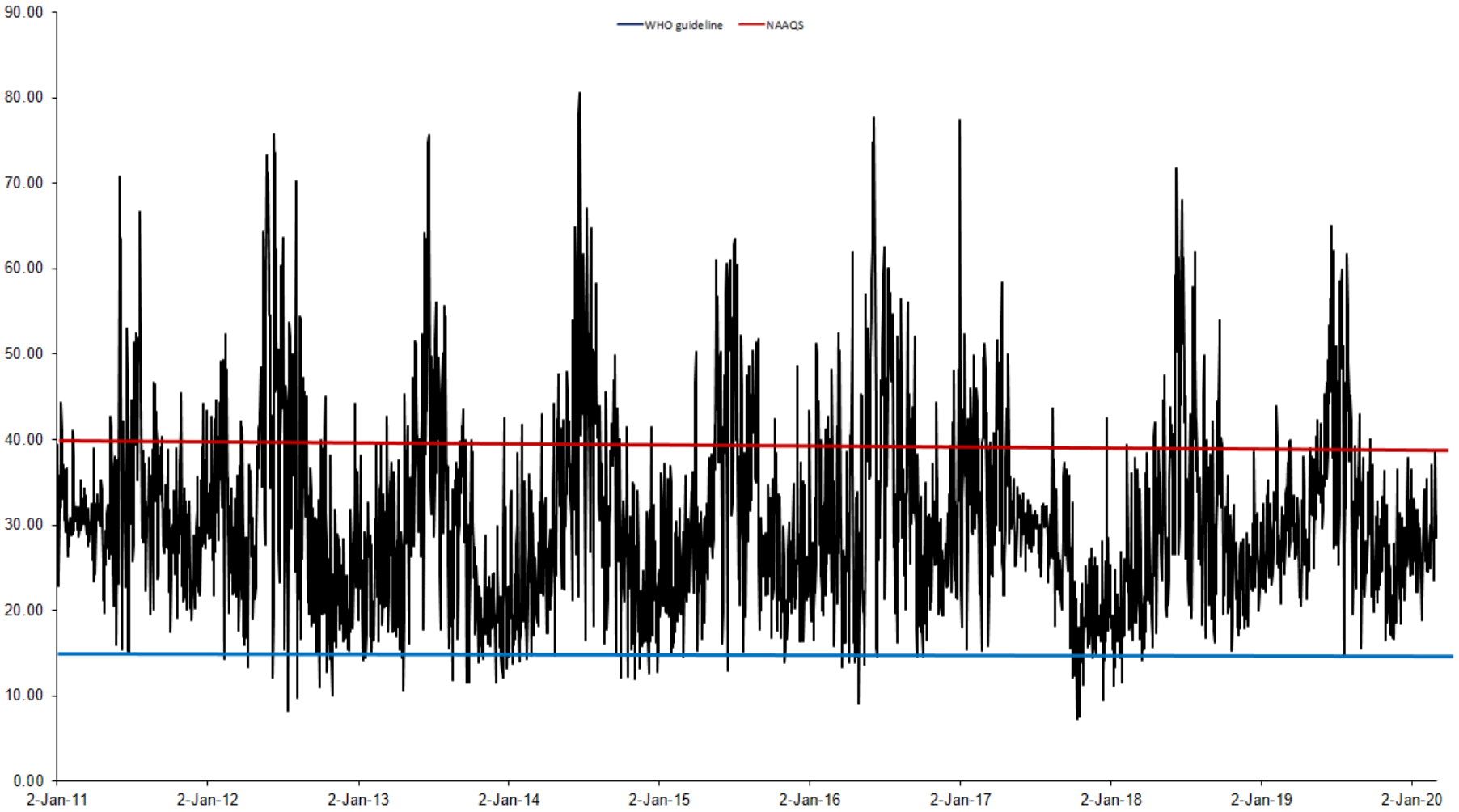


Figure 6.4: Time-series of $PM_{2.5}$ levels in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

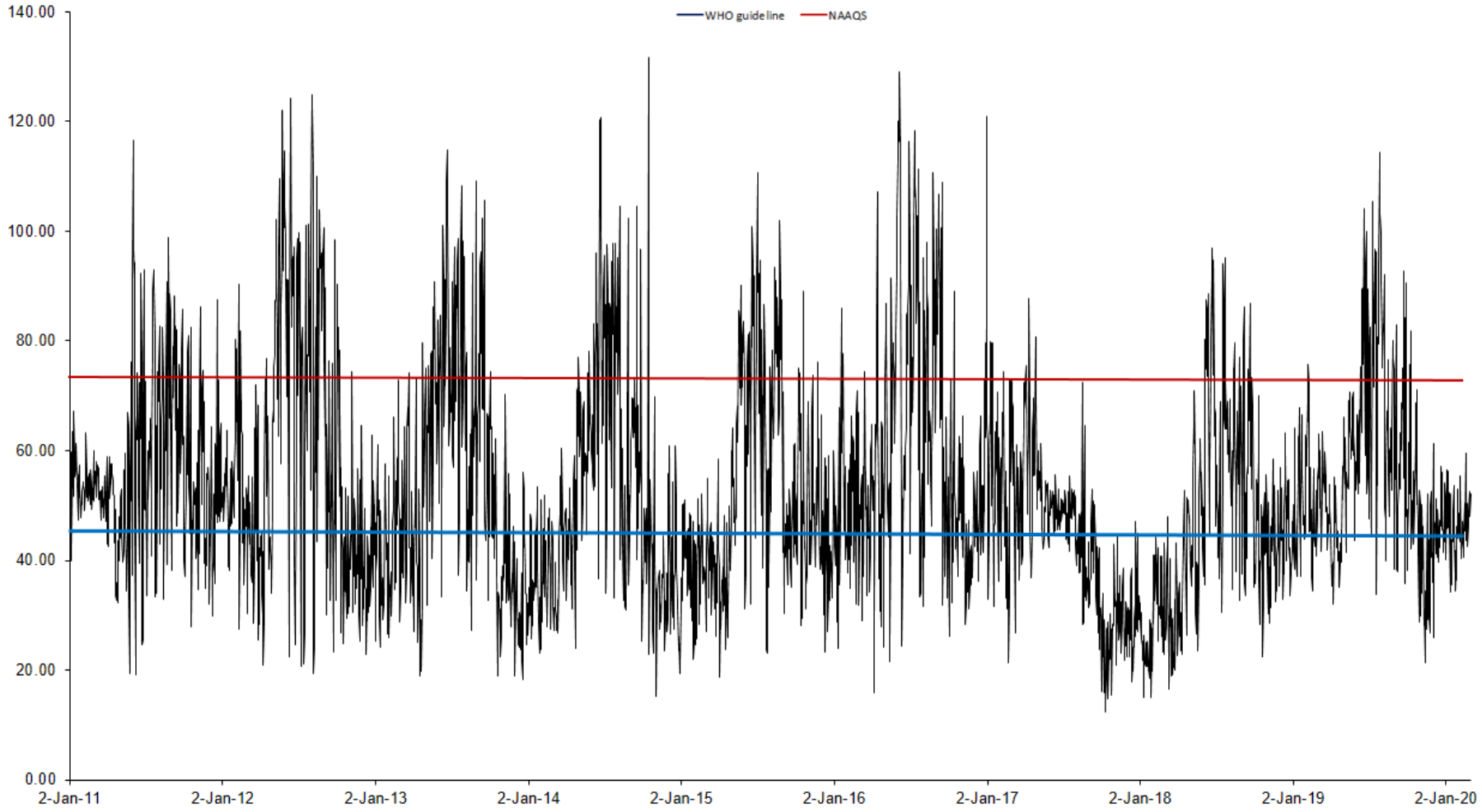


Figure 6.5: Time-series of PM_{10} levels in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

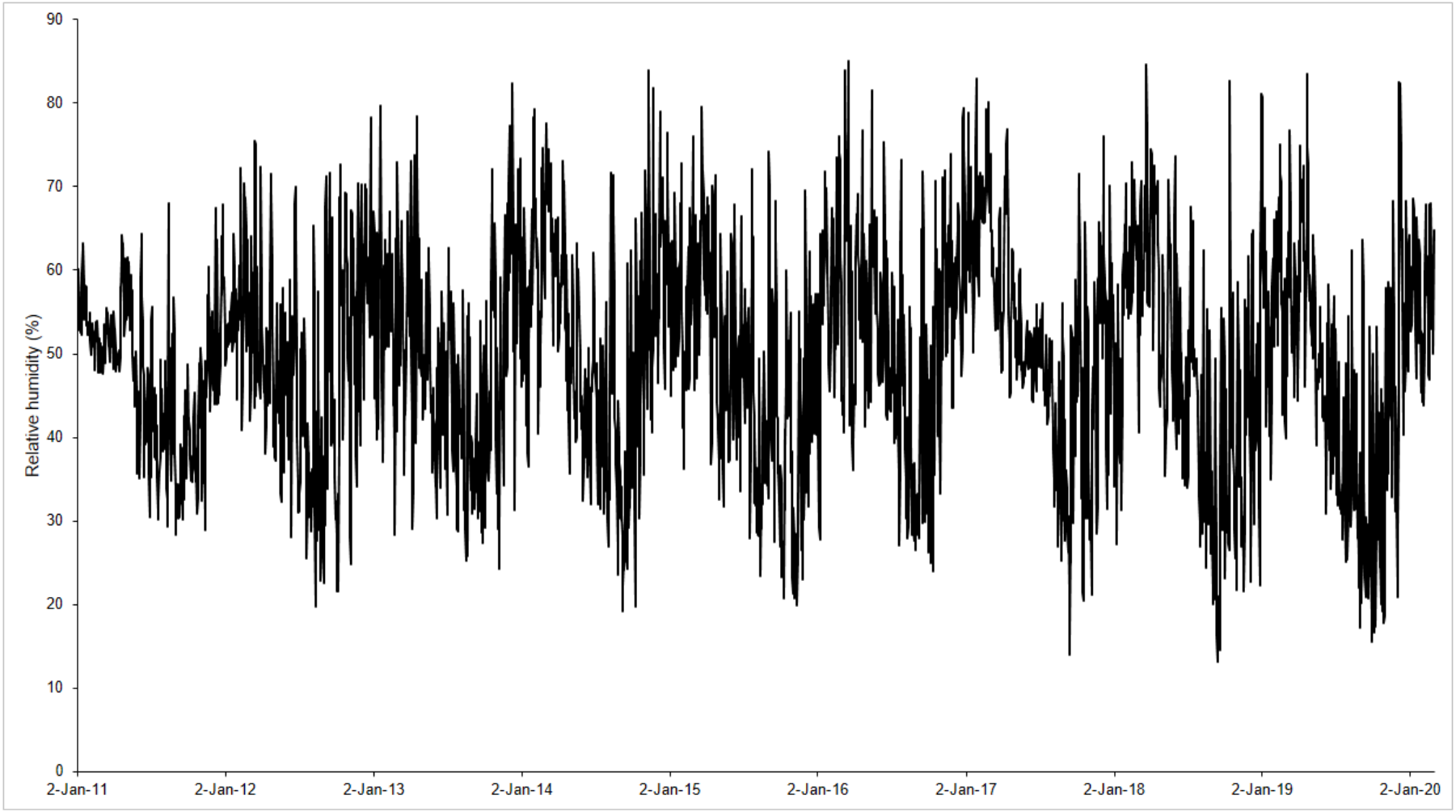


Figure 6.6: Time-series of relative humidity in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

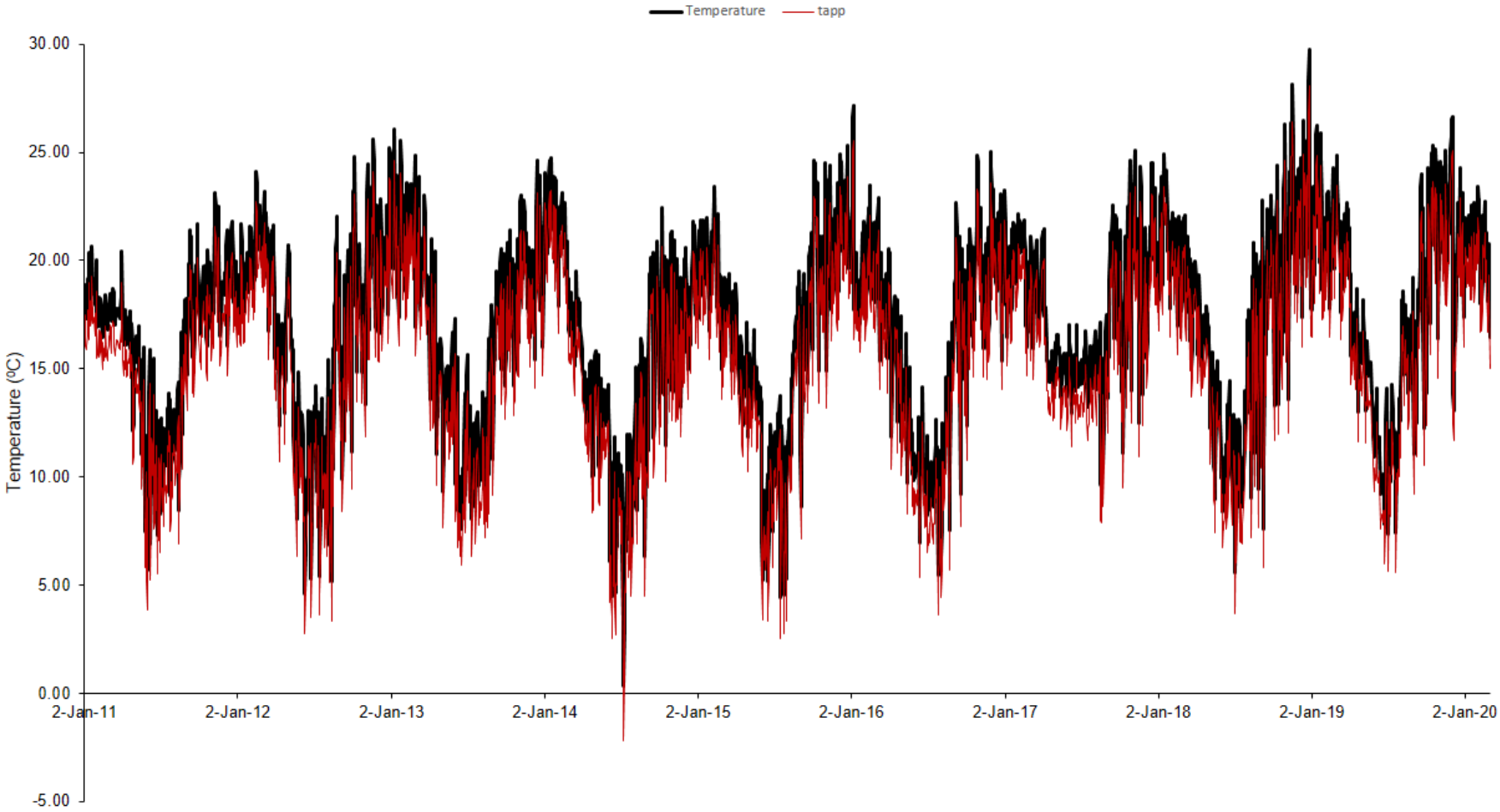


Figure 6.7: Time-series of temperature and apparent temperature in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

Table 6.2 shows the correlations between air pollutants and the meteorological data. There is a strong positive correlation between PM_{10} and $PM_{2.5}$. Strong positive correlations can also be seen between PM_{10} and NO_2, and $PM_{2.5}$ and NO_2. There are moderate positive correlations between PM_{10} and SO_2, and $PM_{2.5}$ and $SO_{2.5}$. Moderate negative correlations can be seen with O_3 and NO_2, as well as O_3 and SO_2.

Table 6.2: Spearman correlation coefficients between air pollution and weather variables in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

	SO_2	NO_2	O_3	$PM_{2.5}$	PM_{10}	RH	Temp
NO_2	0.52						
O_3	-0.34	-0.34					
$PM_{2.5}$	0.41	0.67	-0.30				
PM_{10}	0.37	0.64	-0.24	0.86			
RH	-0.22	-0.27	-0.12	-0.14	-0.33		
Temp	-0.31	-0.41	0.69	-0.35	-0.32	0.01	
Tapp	-0.31	-0.41	0.68	-0.36	-0.33	0.03	0.99

Table 6.3 shows the minimum and maximum concentration levels for the five air pollutants of interest. The maximum concentration for SO_2 was 63.6 $\mu g/m^3$ in October and the minimum concentration 2.4 $\mu g/m^3$ in November. The maximum concentration for NO_2 was 80.8 $\mu g/m^3$ in June and the minimum concentration was 5.3 $\mu g/m^3$ in the month of July. The maximum concentration for O_3 was 103.9 $\mu g/m^3$ in October. The minimum concentration for O_3 was 14.3 $\mu g/m^3$ in the month of June. The maximum concentration for $PM_{2.5}$ was 80.6 $\mu g/m^3$ in June and the minimum concentration was 7.2 $\mu g/m^3$ in October. The maximum concentration for PM_{10} 131.7$\mu g/m^3$ in October and minimum concentration for PM_{10} was 12.5 $\mu g/m^3$ in October. The Kruskal Wallis tests showed significant differences in months for all five air pollutants (p-value < 0.001).

Table 6.3: Mean and range of daily air pollutant levels by month in VTAPA, South Africa, 2 January 2011 – 29 February 2020.

	SO_2 ($\mu g.m^{-3}$)			NO_2 ($\mu g.m^{-3}$)			O_3 ($\mu g.m^{-3}$)			$PM_{2.5}$ ($\mu g.m^{-3}$)			PM_{10} ($\mu g.m^{-3}$)		
Month	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
January	14.1	3.4	44.1	24.2	8.9	46.1	54.7	24.9	85.2	27.5	11.2	52.3	44.6	14.9	85.9
February	13.6	3.51	52.9	26.6	11.9	46.6	50.6	24.1	89.8	29.1	14.3	52.4	47.2	21.2	90.4
March	15.1	2.7	39.2	28.8	14.6	55.7	46.8	25.6	77.4	28.8	13.3	52.5	45.7	16.5	74.4
April	14.9	4.8	48.3	31.4	15.9	55.9	41.4	17.8	64.7	28.3	9.1	61.9	46.3	15.9	107.1
May	16.9	5.7	53.5	35.6	16.9	65.1	37.3	21.2	60.1	36.2	13.6	73.3	61.8	19.4	121.1
June	19.5	3.7	54.8	38.9	9.8	80.8	36.8	14.3	64.2	43.2	12.1	80.6	71.4	19.2	129.1
July	20.2	3.5	60.6	37.5	5.3	78.5	36.8	16.1	59.1	39.4	8.2	67.1	68.5	20.6	118.4
August	18.4	4.9	54.5	35.6	6.4	67.4	47.5	21.8	79.9	34.1	9.7	70.3	65.0	19.4	124.8
September	16.6	5.1	58.0	31.8	13.2	64.8	55.6	25.0	87.9	30.1	10.9	53.9	58.0	23.4	108.9
October	13.9	3.4	63.6	26.2	11.4	46.9	59.9	28.4	103.9	25.3	7.2	45.4	45.7	12.5	131.7
November	11.4	2.4	35.9	23.8	7.9	40.8	61.4	31.8	89.0	22.6	11.5	44.3	41.6	15.2	86.1
December	10.9	3.7	44.5	23.3	9.4	40.0	59.0	22.1	87.5	25.9	9.5	77.5	42.2	17.7	121.0

Table 6.4 shows the concentration levels of each pollutant per day in the week. SO_2 mean concentration levels varied from of 15.1 μg/m^3 to 16.0 μg/m^3. The mean SO_2 levels did not vary during the week (p=0.70). O_3 mean concentration levels varied from of 48.5 μg/m^3 to 49.1 μg/m^3. The mean O_3 levels did not vary during the week (p=0.63).

The mean concentration levels of NO_2 (p=0.001), $PM_{2.5}$ (p=0.001), and PM_{10} (p=0.001) varied during the week. Each showing higher mean concentration levels during work days (Monday to Friday) and lower mean concentration levels during the weekends (Saturday and Sunday).

Table 6.4: Mean and range of daily air pollutant levels by day of the week in VTAPA, South Africa, 2 January 2011 – 29 February 2020.

	SO_2 ($\mu g.m^{-3}$)			NO_2 ($\mu g.m^{-3}$)			O_3 ($\mu g.m^{-3}$)			$PM_{2.5}$ ($\mu g.m^{-3}$)			PM_{10} ($\mu g.m^{-3}$)		
Day of week	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**	**Mean**	**Min**	**Max**
Sunday	15.4	3.9	56.7	29.5	7.9	72.7	49.4	18.7	103.9	29.9	10.6	71.2	50.8	15.1	120.1
Monday	16.0	3.6	63.56	31.1	6.4	73.5	48.6	14.2	93.3	31.3	9.7	77.7	53.9	15.2	129.1
Tuesday	15.5	3.5	54.5	30.7	8.8	74.9	48.5	17.8	89.8	31.1	7.2	66.2	54.1	12.5	114.1
Wednesday	15.5	2.4	57.7	31.2	9.4	66.5	49.1	15.5	89.2	31.4	9.5	73.5	55.3	14.7	131.7
Thursday	15.4	3.2	54.8	31.5	10.4	80.8	48.7	15.8	89.0	31.5	11.2	74.8	55.0	15.5	124.8
Friday	15.1	3.2	48.2	29.8	9.8	65.6	49.1	19.6	95.3	31.0	9.1	78.2	52.7	20.6	123.0
Saturday	15.2	2.7	51.0	27.9	5.3	64.2	49.9	16.5	90.9	29.6	8.2	80.6	49.7	15.9	121.0

Abbreviations: PM_{10}: particulate matter with an aerodynamic diameter of less than 10 μm; NO_2: nitrogen dioxide, SO_2: sulphur dioxide

Due to the nature of the method used and the high correlation between PM_{10} and $PM_{2.5}$, the seven air pollutant mixtures explored were:

- PM_{10}, NO_2, and SO_2 (mixture 1)
- $PM_{2.5}$, NO_2, and SO_2 (mixture 2)
- PM_{10}, NO_2, and O_3 (mixture 3)
- $PM_{2.5}$, NO_2, and O_3 (mixture 4)
- PM_{10}, SO_2, and O_3 (mixture 5)
- $PM_{2.5}$, SO_2, and O_3 (mixture 6)
- O_3, NO_2, and SO_2 (mixture 7)

Table 6.5 presents the distribution of non-referent and referent days by month for PM_{10}, NO_2, and SO_2 (mixture 1). Table 6.6 shows the distribution of non-referent and referent days by week. Most of the non-referent days were observed in January (282 days, 8.99%), which was in summer. Out of 209 referent days, most occurred in December (51 days, 24.4%), which was also in summer. The most non-referent days fell on Tuesdays (457 days, 14.57%), while most referent days fell on Sundays (43 days, 20.57%).

Table 6.5: Distribution of non-referent and referent days by month in VTAPA, South Africa, 2 January 2011 – 29 February 2020, PM_{10}, NO_2 and SO_2 (mixture 1).

	Non-referent days (n = 3137 days)		Referent days (n = 209 days)	
Month	**Frequency**	**Percent**	**Frequency**	**Percent**
January	282	8.99	27	12.92
February	269	8.58	14	6.70
March	265	8.45	14	6.70
April	254	8.10	16	7.66
May	276	8.80	3	1.44
June	262	8.35	8	3.83
July	277	8.83	2	0.96
August	274	8.73	5	2.39
September	268	8.54	2	0.96
October	256	8.16	23	11.00
November	226	7.20	44	21.05
December	228	7.27	51	24.40

Table 6.6: Distribution of non-referent days and referent days by day of the week in VTAPA, South Africa, 2 January 2011 – 29 February 2020, PM_{10}, NO_2 and SO_2 (mixture 1).

	Non-referent days (n = 3137 days)		Referent days (n = 209 days)	
Day	**Frequency**	**Percent**	**Frequency**	**Percent**
Sunday	435	13.87	43	20.57
Monday	456	14.54	22	10.53
Tuesday	457	14.57	21	10.05
Wednesday	438	13.96	40	19.14
Thursday	457	14.57	21	10.05
Friday	456	14.54	22	10.53
Saturday	438	13.96	40	19.14

Table 6.7 presents the distribution of non-referent days and referent days by month, and table 6.8 shows the distribution of non-referent and referent days by week, for $PM_{2.5}$, NO_2, and SO_2 (mixture 2). Most of the non-referent days were observed in January (280 days, 8.95%), i.e. in summer. Most of the referent days, out of 216, were observed in November (54 days, 25%), i.e. in spring. Table 6.8 shows that majority of the non-referent days fell on Mondays (457 days, 14.6%), while majority of referent days fell on Sundays (47 days, 21.76%).

Table 6.7: Distribution of non-referent and referent days by month in VTAPA, South Africa, 2 January 2011 – 29 February 2020, $PM_{2.5}$, NO_2 and SO_2 (mixture 2).

	Non-referent days (n = 3130 days)		Referent days (n = 216 days)	
Month	**Frequency**	**Percent**	**Frequency**	**Percent**
January	280	8.95	29	13.43
February	273	8.72	10	4.63
March	266	8.50	13	6.02
April	255	8.15	15	6.94
May	277	8.85	2	0.93
June	263	8.40	7	3.24
July	276	8.82	3	1.39
August	270	8.63	9	4.17
September	267	8.53	3	1.39
October	251	8.02	28	12.96
November	216	6.90	54	25.00
December	236	7.54	43	19.91

Table 6.8: Distribution of non-referent days and referent days by day of the week in VTAPA, South Africa, 2 January 2011 – 29 February 2020, $PM_{2.5}$, NO_2 and SO_2 (mixture 2).

	Non-referent days (n = 3130 days)		Referent days (n = 216 days)	
Day	**Frequency**	**Percent**	**Frequency**	**Percent**
Sunday	431	13.77	47	21.76
Monday	457	14.60	21	9.72
Tuesday	454	14.50	24	11.11
Wednesday	441	14.09	37	17.13
Thursday	454	14.50	24	11.11
Friday	451	14.41	27	12.50
Saturday	442	14.12	36	16.67

Table 6.9 presents the distribution of non-referent days and referent days by month for PM_{10}, NO_2, and O_3 (mixture 3). Most non-referent days were observed in January (300 days, 9.18%), i.e. in summer. Most of the 79 referent days were observed in April (16 days, 20.25%), i.e. in autumn. Table 6.10 shows that the majority of non-referent days fell on Tuesdays (471 days, 14.42%). The majority of referent days fell on Saturdays (17 days, 21.52%).

Table 6.9: Distribution of non-referent and referent days by month in VTAPA, South Africa, 2 January 2011 – 29 February 2020, PM_{10}, NO_2 and O_3 (mixture 3).

	Non-referent days (n = 3267 days)		Referent days (n = 79 days)	
Month	**Frequency**	**Percent**	**Frequency**	**Percent**
January	300	9.18	9	11.39
February	275	8.42	8	10.13
March	264	8.08	15	18.99
April	254	7.77	16	20.25
May	273	8.36	6	7.59
June	261	7.99	9	11.39
July	272	8.33	7	8.86
August	277	8.48	2	2.53
September	269	8.23	1	1.27
October	278	8.51	1	1.27
November	269	8.23	1	1.27
December	275	8.42	4	5.06

Table 6.10: Distribution of non-referent days and referent days by day of the week, in VTAPA, South Africa, 2 January 2011 – 29 February 2020, PM_{10}, NO_2 and O_3 (mixture 3).

	Non-referent days (n = 3267 days)		Referent days (n = 79 days)	
Day	**Frequency**	**Percent**	**Frequency**	**Percent**
Sunday	465	14.23	13	16.46
Monday	469	14.36	9	11.39
Tuesday	471	14.42	7	8.86
Wednesday	468	14.33	10	12.66
Thursday	469	14.36	9	11.39
Friday	464	14.20	14	17.72
Saturday	461	14.11	17	21.52

Table 6.11 presents the distribution of non-referent days and referent days by month for $PM_{2.5}$, NO_2, and O_3 (mixture 4). Most of non-referent days were observed in January (300 days, 9.16%), i.e. in summer. Most of the 70 referent days were observed in March and April (13 days, 18.57%, each), i.e. in autumn. Table 6.12 shows that majority of the non-referent days fell on Tuesdays (253 days, 14.7%), while the majority of referent days fell on Sundays (13 days, 18.57%).

Table 6.11: Distribution of non-referent and referent days by month in VTAPA, South Africa, 2 January 2011 – 29 February 2020, $PM_{2.5}$, NO_2 and O_3 (mixture 4).

	Non-referent days (n = 3276 days)		Referent days (n = 70 days)	
Month	**Frequency**	**Percent**	**Frequency**	**Percent**
January	300	9.16	9	12.86
February	277	8.46	6	8.57
March	266	8.12	13	18.57
April	257	7.84	13	18.57
May	275	8.39	4	5.71
June	262	8.00	8	11.43
July	271	8.27	8	11.43
August	277	8.46	2	2.86
September	268	8.18	2	2.86
October	278	8.49	1	1.43
November	269	8.21	1	1.43
December	276	8.42	3	4.29

Table 6.12: Distribution of non-referent days and referent days by day of the week in VTAPA, South Africa, 2 January 2011 – 29 February 2020, $PM_{2.5}$, NO_2 and O_3 (mixture 4).

	Non-referent days (n = 3276 days)		Referent days (n = 70 days)	
Day	**Frequency**	**Percent**	**Frequency**	**Percent**
Sunday	465	14.19	13	18.57
Monday	469	14.32	9	12.86
Tuesday	472	14.41	6	8.57
Wednesday	467	14.26	11	15.71
Thursday	471	14.38	7	10.00
Friday	466	14.22	12	17.14
Saturday	466	14.22	12	17.14

Table 6.13 presents the distribution of non-referent days and referent days by month for PM_{10}, SO_2, and O_3 (mixture 5). Most non-referent days were observed in January (307 days, 9.3%), i.e. in summer. Most of the 46 referent days were observed in April (12 days, 26.09%), i.e. in autumn. Table 6.14 shows that the majority of non-referent days fell on Thursdays (473 days, 14.33%), while the majority of referent days fell on Wednesdays (9 days, 19.57%).

Table 6.13: Distribution of non-referent and referent days by month in VTAPA, South Africa, 2 January 2011 – 29 February 2020, PM_{10}, SO_2 and O_3 (mixture 5).

	Non-referent days (n = 3300 days)		Referent days (n = 46 days)	
Month	**Frequency**	**Percent**	**Frequency**	**Percent**
January	307	9.30	2	4.35
February	282	8.55	1	2.17
March	268	8.12	11	23.91
April	258	7.82	12	26.09
May	272	8.24	7	15.22
June	263	7.97	7	15.22
July	276	8.36	3	6.52
August	278	8.42	1	2.17
September	270	8.18	1	2.17
October	279	8.45	1	2.17
November	269	8.15	2	4.35
December	278	8.42	1	2.17

Table 6.14: Distribution of non-referent days and referent days by day of the week in VTAPA, South Africa, 2 January 2011 – 29 February 2020, PM_{10}, SO_2 and O_3 (mixture 5).

	Non-referent days (n = 3300 days)		Referent days (n = 46 days)	
Day	**Frequency**	**Percent**	**Frequency**	**Percent**
Sunday	471	14.27	7	15.22
Monday	475	14.39	3	6.52
Tuesday	471	14.27	7	15.22
Wednesday	469	14.21	9	19.57
Thursday	473	14.33	5	10.87
Friday	470	14.24	8	17.39
Saturday	471	14.27	7	15.22

Table 6.15 presents the distribution of non-referent days and referent days by month for $PM_{2.5}$, SO_2, and O_3 (mixture 6). Most non-referent days were observed in January (159 days, 9.3%), i.e. in summer. Most of the 41 referent days were observed in March (11 days, 24.39%), i.e. in autumn. Table 6.16 shows that majority of the non-referent days fell on Sundays and Mondays (474 days, 14.34%, each), while majority of referent days fell on Wednesdays (10 days, 24.39%).

Table 6.15: Distribution of non-referent and referent days by month in VTAPA, South Africa, 2 January 2011 – 29 February 2020, $PM_{2.5}$, SO_2 and O_3 (mixture 6).

	Non-referent days (n = 3305 days)		Referent days (n = 41 days)	
Month	**Frequency**	**Percent**	**Frequency**	**Percent**
January	308	9.32	1	2.44
February	282	8.53	1	2.44
March	269	8.14	10	24.39
April	261	7.90	9	21.95
May	272	8.23	7	17.07
June	264	7.99	6	14.63
July	275	8.32	4	9.76
August	277	8.38	2	4.88
September	270	8.17	1	2.44
October	279	8.44	1	2.44
November	269	8.14	1	2.44
December	279	8.44		

Table 6.16: Distribution of non-referent days and referent days by day of the week in VTAPA, South Africa, 2 January 2011 – 29 February 2020, $PM_{2.5}$, SO_2 and O_3 (mixture 6).

	Non-referent days (n = 3305 days)		Referent days (n = 41 days)	
Day	**Frequency**	**Percent**	**Frequency**	**Percent**
Sunday	474	14.34	4	9.76
Monday	474	14.34	4	9.76
Tuesday	474	14.34	4	9.76
Wednesday	468	14.16	10	24.39
Thursday	471	14.25	7	17.07
Friday	471	14.25	7	17.07
Saturday	473	14.31	5	12.20

Table 6.17 presents the distribution of non-referent days and referent days by month for O_3, NO_2, and SO_2 (mixture 7). Most of non-referent days were observed in January (306 days, 9.25%), i.e. in summer. Most of the 38 referent days were observed in March, June, and December (8 days, 21.05%, each), i.e. in autumn, winter, and summer, respectively. Table 6.18 shows the majority of the non-referent days fell on Thursdays (476 days, 14.39%). The majority of referent days fell on both Fridays and Saturdays (8 days, 21.05%, each).

Table 6.17: Distribution of non-referent and referent days by month in VTAPA, South Africa, 2 January 2011 – 29 February 2020, O_3, NO_2, and SO_2 (mixture 7).

	Non-referent days (n = 3308 days)		Referent days (n = 38 days)	
Month	**Frequency**	**Percent**	**Frequency**	**Percent**
January	306	9.25	3	7.89
February	277	8.37	6	15.79
March	271	8.19	8	21.05
April	267	8.07	3	7.89
May	276	8.34	3	7.89
June	262	7.92	8	21.05
July	277	8.37	2	5.26
August	277	8.37	2	5.26
September	270	8.16	3	7.89
October	279	8.43	3	7.89
November	270	8.16	6	15.79
December	276	8.34	8	21.05

Table 6.18: Distribution of non-referent days and referent days by day of the week in VTAPA, South Africa, 2 January 2011 – 29 February 2020, O_3, NO_2, and SO_2 (mixture 7).

	Non-referent days (n = 3308 days)		Referent days (n = 38 days)	
Day	**Frequency**	**Percent**	**Frequency**	**Percent**
Sunday	473	14.30	5	13.16
Monday	473	14.30	5	13.16
Tuesday	474	14.33	4	10.53
Wednesday	472	14.27	6	15.79
Thursday	476	14.39	2	5.26
Friday	470	14.21	8	21.05
Saturday	470	14.21	8	21.05

Table 6.18 shows the majority of the non-referent days fell on Thursdays (476 days, 14.39%). The majority of referent days fell on both Fridays and Saturdays (8 days, 21.05%, each).

Table 6.19 reports on the frequency of each of the 64 day types used in the regression models for PM_{10}, NO_2, and SO_2 (mixture 1). Day type 444 was the most frequently observed day type (338 days; 10.1%), i.e. when all three air pollutants had levels in the highest quartile. Day type 414 was the least frequently observed, where only one out of the 3 346 days (0.03%) had PM_{10} levels in the highest quartile, NO_2 levels in the lowest quartile, and SO_2 levels in the highest quartile.

Table 6.19: Frequency of each day type (64 day types) in VTAPA, South Africa, 2 January 2011 – 29 February 2020, PM_{10}, NO_2 and SO_2 (mixture 1).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	209	6.25	**211**	111	3.32
112	104	3.11	**212**	75	2.24
113	47	1.40	**213**	37	1.11
114	40	1.20	**214**	18	0.54
121	108	3.23	**221**	87	2.60
122	67	2.00	**222**	83	2.48
123	50	1.49	**223**	59	1.76
124	52	1.55	**224**	47	1.40
131	44	1.32	**231**	38	1.14
132	31	0.93	**232**	59	1.76
133	19	0.57	**233**	73	2.18
134	26	0.78	**234**	62	1.85
141	14	0.42	**241**	10	0.30
142	9	0.27	**242**	27	0.81
143	7	0.21	**243**	26	0.78
144	9	0.27	**244**	25	0.75
311	63	1.88	**411**	21	0.63
312	42	1.26	**412**	14	0.42
313	34	1.02	**413**	9	0.27
314	11	0.33	**414**	1	0.03
321	54	1.61	**421**	12	0.36
322	65	1.94	**422**	37	1.11
323	61	1.82	**423**	22	0.66
324	27	0.81	**424**	6	0.18
331	31	0.93	**431**	13	0.39
332	80	2.39	**432**	49	1.46
333	143	4.27	**433**	53	1.58
334	74	2.21	**434**	42	1.26
341	13	0.39	**441**	8	0.24
342	33	0.99	**442**	62	1.85
343	48	1.43	**443**	149	4.45
344	58	1.73	**444**	338	10.10

First digit refers to PM_{10} levels, second digit refers to NO_2 levels and third digit refers to SO_2 levels. 1 indicates when an air pollutant had levels in the lowest quartile, 2 indicates when an air pollutant had levels in the second quartile, 3 indicates when an air pollutant had levels in the third quartile, 4 indicates when an air pollutant had levels in the highest quartile, e.g. 123 means that PM_{10} levels were in the lowest quartile, NO_2 levels were in the second quartile and SO_2 levels were in the third quartile.

Table 6.20 reports on the frequency of each of the 64 day types used in the regression models for $PM_{2.5}$, NO_2, and SO_2 (mixture 2). Day type 444 was the most frequently observed day type (339 days; 10.13%), i.e. when all three air pollutants had levels in the highest quartile. Day type 143 was the least frequently observed, where only three out of the 3 346 days (0.09%) had $PM_{2.5}$ levels in the lowest quartile, NO_2 levels in the highest quartile, and SO_2 levels in the third quartile.

Table 6.20: Frequency of each day type (64 day types) in VTAPA, South Africa, 2 January 2011 – 29 February 2020, $PM_{2.5}$, NO_2 and SO_2 (mixture 2).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	216	6.46	**211**	118	3.53
112	109	3.26	**212**	76	2.27
113	52	1.55	**213**	46	1.37
114	44	1.32	**214**	16	0.48
121	113	3.38	**221**	82	2.45
122	65	1.94	**222**	79	2.36
123	45	1.34	**223**	70	2.09
124	47	1.40	**224**	44	1.32
131	45	1.34	**231**	43	1.29
132	30	0.90	**232**	61	1.82
133	15	0.45	**233**	73	2.18
134	30	0.90	**234**	45	1.34
141	7	0.21	**241**	16	0.48
142	8	0.24	**242**	27	0.81
143	3	0.09	**243**	22	0.66
144	7	0.21	**244**	19	0.57
311	47	1.40	**411**	23	0.69
312	37	1.11	**412**	13	0.39
313	21	0.63	**413**	8	0.24
314	6	0.18	**414**	4	0.12
321	50	1.49	**421**	16	0.48
322	78	2.33	**422**	30	0.90
323	60	1.79	**423**	17	0.51
324	33	0.99	**424**	8	0.24
331	27	0.81	**431**	11	0.33
332	91	2.72	**432**	37	1.11
333	132	3.95	**433**	68	2.03
334	69	2.06	**434**	60	1.79
341	15	0.45	**441**	7	0.21
342	39	1.17	**442**	57	1.70
343	67	2.00	**443**	138	4.12
344	65	1.94	**444**	339	10.13

First digit refers to $PM_{2.5}$ levels, second refers to NO_2 levels, third refers to SO_2 levels. 1 indicates when an air pollutant had levels in the lowest quartile, 2 indicates when an air pollutant had levels in the second quartile, 3 indicates when an air pollutant had levels in the third quartile, 4 indicates when an air pollutant had levels in the highest quartile, e.g. 123 means that $PM_{2.5}$ levels were in the lowest quartile, NO_2 levels were in the second quartile and SO_2 levels were in the third quartile.

Table 6.21 reports on the frequency of each of the 64 day types used in the regression models for PM_{10}, NO_2, and O_3 (mixture 3). Day type 441 was the most frequently observed day type (299 days; 8.94%), when PM_{10} and NO_2 had levels in the highest quartile, and O_3 was in the lowest quartile. Day type 411 was the least frequently observed, where only three out of 3 346 days (0.09%) had PM_{10} levels in the highest quartile, and NO_2 and O_3 levels in the lowest quartile.

Table 6.21: Frequency of each day type (64 day types) in VTAPA, South Africa, 2 January 2011 – 29 February 2020, PM_{10}, NO_2 and O_3 (mixture 3).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	79	2.36	**211**	27	0.81
112	75	2.24	**212**	53	1.58
113	101	3.02	**213**	62	1.85
114	145	4.33	**214**	99	2.96
121	36	1.08	**221**	55	1.64
122	56	1.67	**222**	74	2.21
123	78	2.33	**223**	79	2.36
124	107	3.20	**224**	68	2.03
131	21	0.63	**231**	42	1.26
132	21	0.63	**232**	68	2.03
133	36	1.08	**233**	69	2.06
134	42	1.26	**234**	53	1.58
141	14	0.42	**241**	23	0.69
142	12	0.36	**242**	23	0.69
143	6	0.18	**243**	16	0.48
144	7	0.21	**244**	26	0.78
311	12	0.36	**411**	3	0.09
312	39	1.17	**412**	9	0.27
313	50	1.49	**413**	15	0.45
314	49	1.46	**414**	18	0.54
321	27	0.81	**421**	12	0.36
322	75	2.24	**422**	18	0.54
323	61	1.82	**423**	24	0.72
324	44	1.32	**424**	23	0.69
331	87	2.60	**431**	37	1.11
332	92	2.75	**432**	42	1.26
333	101	3.02	**433**	48	1.43
334	48	1.43	**434**	30	0.90
341	62	1.85	**441**	299	8.94
342	52	1.55	**442**	128	3.83
343	21	0.63	**443**	70	2.09
344	17	0.51	**444**	60	1.79

First digit refers to PM_{10} levels, second refers to NO_2 levels, third refers to O_3 levels.1 indicates when an air pollutant had levels in the lowest quartile, 2 indicates when an air pollutant had levels in the second quartile, 3 indicates when an air pollutant had levels in the third quartile, 4 indicates when an air pollutant had levels in the highest quartile, e.g. 123 means that PM_{10} levels were in the lowest quartile, NO_2 levels were in the second quartile and O_3 levels were in the third quartile.

Table 6.22 reports on the frequency of each of the 64 day types used in the regression models for $PM_{2.5}$, NO_2, and O_3 (mixture 4). Day type 441 was the most frequently observed day type (299 days; 8.94%), when $PM_{2.5}$ and NO_2 had levels in the highest quartile and NO_2 levels in the lowest quartile. Day types 411 and 143 were the least frequently observed, with only four out of the 3 346 days (0.12%).

Table 6.22: Frequency of each day type (64 day types) in VTAPA, South Africa, 2 January 2011 – 29 February 2020, $PM_{2.5}$, NO_2 and O_3 (mixture 4).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	70	2.09	**211**	38	1.14
112	78	2.33	**212**	57	1.70
113	106	3.17	**213**	69	2.06
114	167	4.99	**214**	92	2.75
121	25	0.75	**221**	51	1.52
122	57	1.70	**222**	67	2.00
123	81	2.42	**223**	85	2.54
124	107	3.20	**224**	72	2.15
131	22	0.66	**231**	37	1.11
132	22	0.66	**232**	63	1.88
133	34	1.02	**233**	75	2.24
134	40	1.20	**234**	57	1.70
141	9	0.27	**241**	28	0.84
142	5	0.15	**242**	26	0.78
143	4	0.12	**243**	16	0.48
144	7	0.21	**244**	14	0.42
311	9	0.27	**411**	4	0.12
312	22	0.66	**412**	19	0.57
313	40	1.20	**413**	13	0.39
314	40	1.20	**414**	12	0.36
321	42	1.26	**421**	12	0.36
322	81	2.42	**422**	18	0.54
323	54	1.61	**423**	22	0.66
324	44	1.32	**424**	19	0.57
331	59	1.76	**431**	69	2.06
332	77	2.30	**432**	61	1.82
333	88	2.63	**433**	57	1.70
334	40	1.20	**434**	36	1.08
341	57	1.70	**441**	304	9.09
342	60	1.79	**442**	124	3.71
343	32	0.96	**443**	61	1.82
344	37	1.11	**444**	52	1.55

First digit refers to $PM_{2.5}$ levels, second digit refers to NO_2 levels and third digit refers to O_3 levels 1 indicates when an air pollutant had levels in the lowest quartile, 2 indicates when an air pollutant had levels in the second quartile, 3 indicates when an air pollutant had levels in the third quartile, 4 indicates when an air pollutant had levels in the highest quartile, e.g. 123 means that $PM_{2.5}$ levels were in the lowest quartile, NO_2 levels were in the second quartile and O_3 levels were in the third quartile.

Table 6.23 reports on the frequency of each of the 64 day types used in the regression models for PM_{10}, SO_2, and O_3 (mixture 5). Day type 441 was the most frequently observed day type (221 days; 6.6%), when PM_{10} and SO_2 had levels in the highest quartile and O_3 levels in the lowest quartile. Day type 411 was the least frequently observed, where only eleven out of the 3 346 days (0.33%) had PM_{10} levels in the highest quartile, and SO_2 and O_3 had levels in the lowest quartile.

Table 6.23: Frequency of each day type (64 day types) in VTAPA, South Africa, 2 January 2011 – 29 February 2020, PM_{10}, SO_2 and O_3 (mixture 5).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	46	1.37	**211**	33	0.99
112	62	1.85	**212**	48	1.43
113	86	2.57	**213**	62	1.85
114	181	5.41	**214**	103	3.08
121	37	1.11	**221**	36	1.08
122	42	1.26	**222**	69	2.06
123	62	1.85	**223**	67	2.00
124	70	2.09	**224**	72	2.15
131	34	1.02	**231**	32	0.96
132	19	0.57	**232**	54	1.61
133	45	1.34	**233**	60	1.79
134	25	0.75	**234**	49	1.46
141	33	0.99	**241**	46	1.37
142	41	1.23	**242**	47	1.40
143	28	0.84	**243**	37	1.11
144	25	0.75	**244**	22	0.66
311	16	0.48	**411**	11	0.33
312	35	1.05	**412**	13	0.39
313	53	1.58	**413**	14	0.42
314	57	1.70	**414**	16	0.48
321	42	1.26	**421**	40	1.20
322	66	1.97	**422**	42	1.26
323	66	1.97	**423**	36	1.08
324	46	1.37	**424**	44	1.32
331	63	1.88	**431**	79	2.36
332	94	2.81	**432**	56	1.67
333	93	2.78	**433**	58	1.73
334	36	1.08	**434**	40	1.20
341	67	2.00	**441**	221	6.60
342	63	1.88	**442**	86	2.57
343	21	0.63	**443**	49	1.46
344	19	0.57	**444**	31	0.93

First digit refers to PM_{10} levels, second refers to SO_2 levels, third refers to O_3 levels. 1 indicates when an air pollutant had levels in the lowest quartile, 2 indicates when an air pollutant had levels in the second quartile, 3 indicates when an air pollutant had levels in the third quartile, 4 indicates when an air pollutant had levels in the highest quartile, e.g. 123 means that PM_{10} levels were in the lowest quartile, SO_2 levels were in the second quartile and O_3 levels were in the third quartile.

Table 6.24 reports on the frequency of each of the 64 day types used in the regression models for $PM_{2.5}$, SO_2, and O_3 (mixture 6). Day type 441 was the most frequently observed day type (299 days; 8.94%), when $PM_{2.5}$ and NO_2 had levels in the highest quartile and SO_2 levels in the lowest quartile. Day type 411 was the least frequently observed, where only three out of the 3 346 days (0.09%) had PM_{10} levels in the highest quartile, and NO_2 and SO_2 had levels in the lowest quartile.

Table 6.24: Frequency of each day type (64 day types) in VTAPA, South Africa, 2 January 2011 – 29 February 2020, $PM_{2.5}$, SO_2 and O_3 (mixture 6).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	41	1.23	**211**	35	1.05
112	54	1.61	**212**	52	1.55
113	88	2.63	**213**	74	2.21
114	198	5.92	**214**	98	2.93
121	30	0.90	**221**	38	1.14
122	39	1.17	**222**	66	1.97
123	64	1.91	**223**	73	2.18
124	79	2.36	**224**	66	1.97
131	27	0.81	**231**	42	1.26
132	23	0.69	**232**	49	1.46
133	44	1.32	**233**	68	2.03
134	21	0.63	**234**	52	1.55
141	29	0.87	**241**	44	1.32
142	44	1.32	**242**	38	1.14
143	32	0.96	**243**	26	0.78
144	23	0.69	**244**	16	0.48
311	35	1.05	**411**	7	0.21
312	52	1.55	**412**	15	0.45
313	74	2.21	**413**	18	0.54
314	98	2.93	**414**	17	0.51
321	38	1.14	**421**	44	1.32
322	66	1.97	**422**	37	1.11
323	73	2.18	**423**	25	0.75
324	66	1.97	**424**	31	0.93
331	42	1.26	**431**	94	2.81
332	49	1.46	**432**	63	1.88
333	68	2.03	**433**	46	1.37
334	52	1.55	**434**	28	0.84
341	44	1.32	**441**	236	7.05
342	38	1.14	**442**	90	2.69
343	26	0.78	**443**	50	1.49
344	16	0.48	**444**	35	1.05

First digit refers to $PM_{2.5}$ levels, second refers to SO_2 levels, third refers to O_3 levels. 1 indicates when an air pollutant had levels in the lowest quartile, 2 indicates when an air pollutant had levels in the second quartile, 3 indicates when an air pollutant had levels in the third quartile, 4 indicates when an air pollutant had levels in the highest quartile, e.g. 123 means that $PM_{2.5}$ levels were in the lowest quartile, SO_2 levels were in the second quartile and O_3 levels were in the third quartile.

Table 6.25 reports on the frequency of each of the 64 day types used in the regression models for O_3, NO_2, and SO_2 (mixture 7). Day type 411 was the most frequently observed day type (197 days; 5.89%), when O_3 and NO_2 had levels in the highest quartile and SO_2 levels in the lowest quartile. Day type 141 was the least frequently observed, where only eleven out of the 3 346 days (0.33%) had O_3 levels in the lowest quartile, NO_2 levels in the highest quartile, and SO_2 levels in the lowest quartile.

Table 6.25: Frequency of each day type (64 day types) in VTAPA, South Africa, 2 January 2011 – 29 February 2020, O_3, NO_2 and SO_2 (mixture 7).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	38	1.14	**211**	62	1.85
112	39	1.17	**212**	45	1.34
113	26	0.78	**213**	38	1.14
114	18	0.54	**214**	31	0.93
121	27	0.81	**221**	52	1.55
122	34	1.02	**222**	78	2.33
123	34	1.02	**223**	49	1.46
124	35	1.05	**224**	44	1.32
131	30	0.90	**231**	23	0.69
132	37	1.11	**232**	59	1.76
133	49	1.46	**233**	80	2.39
134	71	2.12	**234**	61	1.82
141	11	0.33	**241**	21	0.63
142	45	1.34	**242**	37	1.11
143	99	2.96	**243**	56	1.67
144	243	7.26	**244**	101	3.02
311	107	3.20	**411**	197	5.89
312	73	2.18	**412**	78	2.33
313	35	1.05	**413**	28	0.84
314	13	0.39	**414**	8	0.24
321	63	1.88	**421**	119	3.56
322	75	2.24	**422**	65	1.94
323	74	2.21	**423**	35	1.05
324	30	0.90	**424**	23	0.69
331	37	1.11	**431**	36	1.08
332	64	1.91	**432**	59	1.76
333	110	3.29	**433**	49	1.46
334	43	1.29	**434**	29	0.87
341	8	0.24	**441**	5	0.15
342	19	0.57	**442**	30	0.90
343	37	1.11	**443**	38	1.14
344	49	1.46	**444**	37	1.11

First digit refers to O_3 levels, second refers to NO_2 levels, third refers to SO_2 levels.1 indicates when an air pollutant had levels in the lowest quartile, 2 indicates when an air pollutant had levels in the second quartile, 3 indicates when an air pollutant had levels in the third quartile, 4 indicates when an air pollutant had levels in the highest quartile, e.g. 123 means that O_3 levels were in the lowest quartile, NO_2 levels were in the second quartile and SO_2 levels were in the third quartile.

6.2. HOSPITAL ADMISSIONS FOR RESPIRATORY AND CARDIOVASCULAR DISEASE

Table 6.26 shows the descriptive statistics of the daily number of respiratory disease (RD) and cardiovascular disease (CVD) hospital admissions. The admissions are set

at a two-day cumulative lag, omitting 1 January 2011, producing a total of 3 346 days in the study.

Table 6.26: Summary statistics of the daily number of RD and CVD hospitalisations in VTAPA, South Africa, 2 January 2011 - 29 February 2020.

Variable	Mean	Min	Max
Respiratory disease			
All ages and both sexes (n=54 822)	16.4	1	55
0-14 year olds (n=28 124)	2.6	1	16
15-64 year olds (n=20 066)	8.3	1	35
≥65 year olds (n=6637)	8.1	1	30
Females (n= 27 749)	8.5	1	37
Males (n= 27 085)	6.1	1	27
Cardiovascular disease			
All ages and both sexes (n=22 507)	6.8	1	29
0-14 year olds (n=405)	1.2	1	3
15-64 year olds (n=13 916)	4.3	1	17
≥65 year olds (n=8199)	2.9	1	13
Females (n= 11 475)	3.7	1	16
Males (n= 11 047)	3.6	1	14

Abbreviations: N-total number of days, Miss- missing values, Min-minimum, Max-Maximum.

In total, 54 822 RD hospital admissions were recorded at the two hospitals. The maximum daily number of RD admissions was 55 and the average 16.4. The total number of females admitted was 27 749 (50.6%), and males admitted for RD reached 27 085 (49.4%). The 0-14 years age group showed the majority (51.3%) of the RD admissions, this was followed by the 15-64 year old age group (36.6%), and lastly, the elderly (≥ 65 years) (12.1%).

In total, 22 507 CVD hospital admissions were recorded at the two hospitals. The maximum daily number of CVD admissions was 29 and the average 6.8. The total number of females admitted was 11 475 (50.95%), and males admitted for CVD reached 11 047 (49.05%). The 15-64 years age group showed the majority (61.8%) of the CVD admissions. The elderly (≥ 65 years) accounted for 36.4% of the CVD hospital admissions.

For the purpose of the study, the combined hospital admissions (all ages and both sexes) for both RD and CVD were investigated.

Figure 6.8 illustrates the time-series of the daily number of RD hospital admissions. Clear seasonal trends were observed, with more RD admissions during the colder months (May to August) than during the warmer months (September to April). Figure 6.9 shows the time-series of the daily number of CVD hospital admissions. There were no definitive seasonal trends observed with CVD hospital admissions from 2 January 2011 to 29 February 2020.

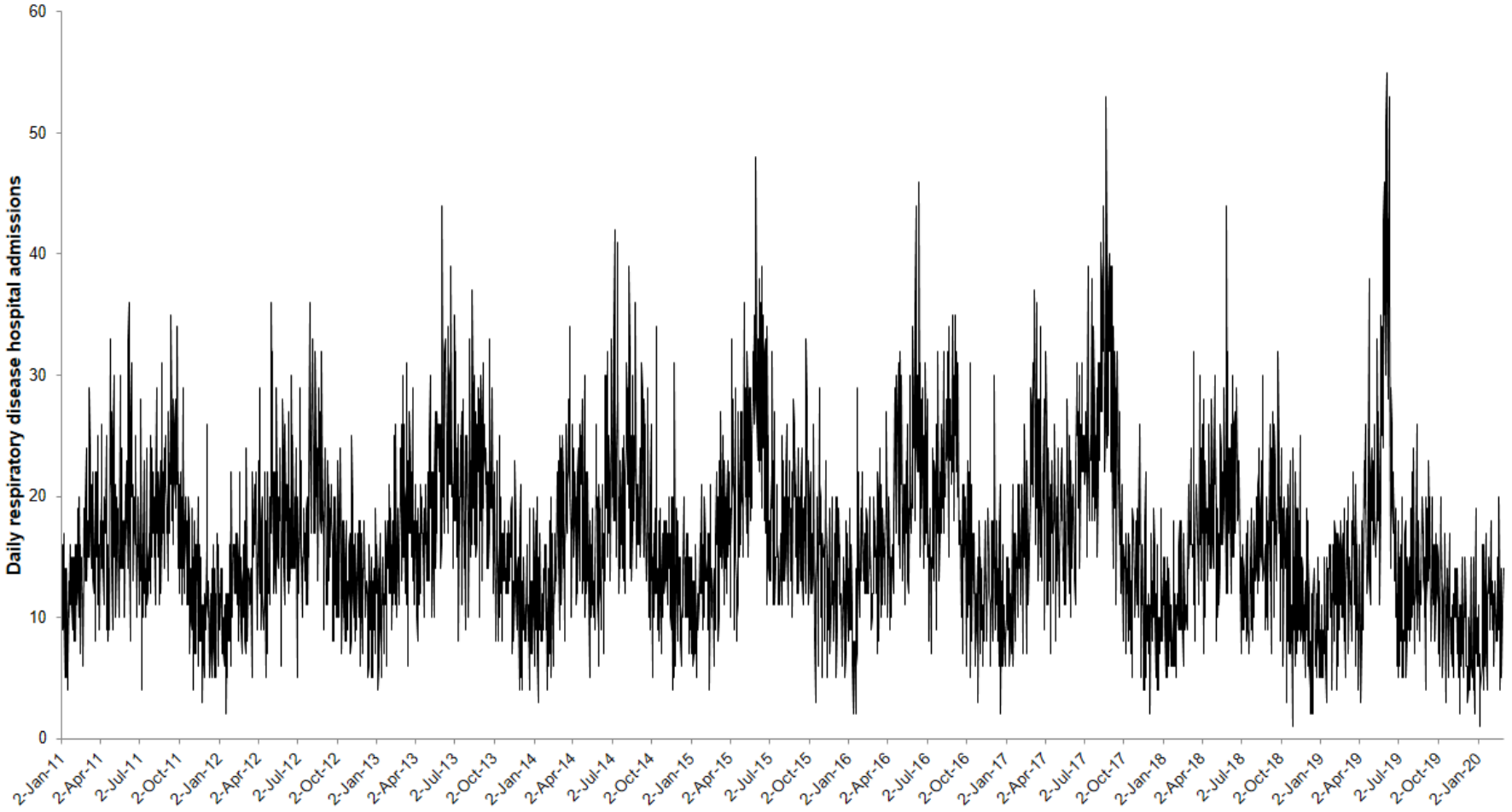


Figure 6.8: Time-series of the daily number of respiratory disease hospital in VTAPA, South Africa, 2 January 2011 – 29 February 2020 (3346 days).

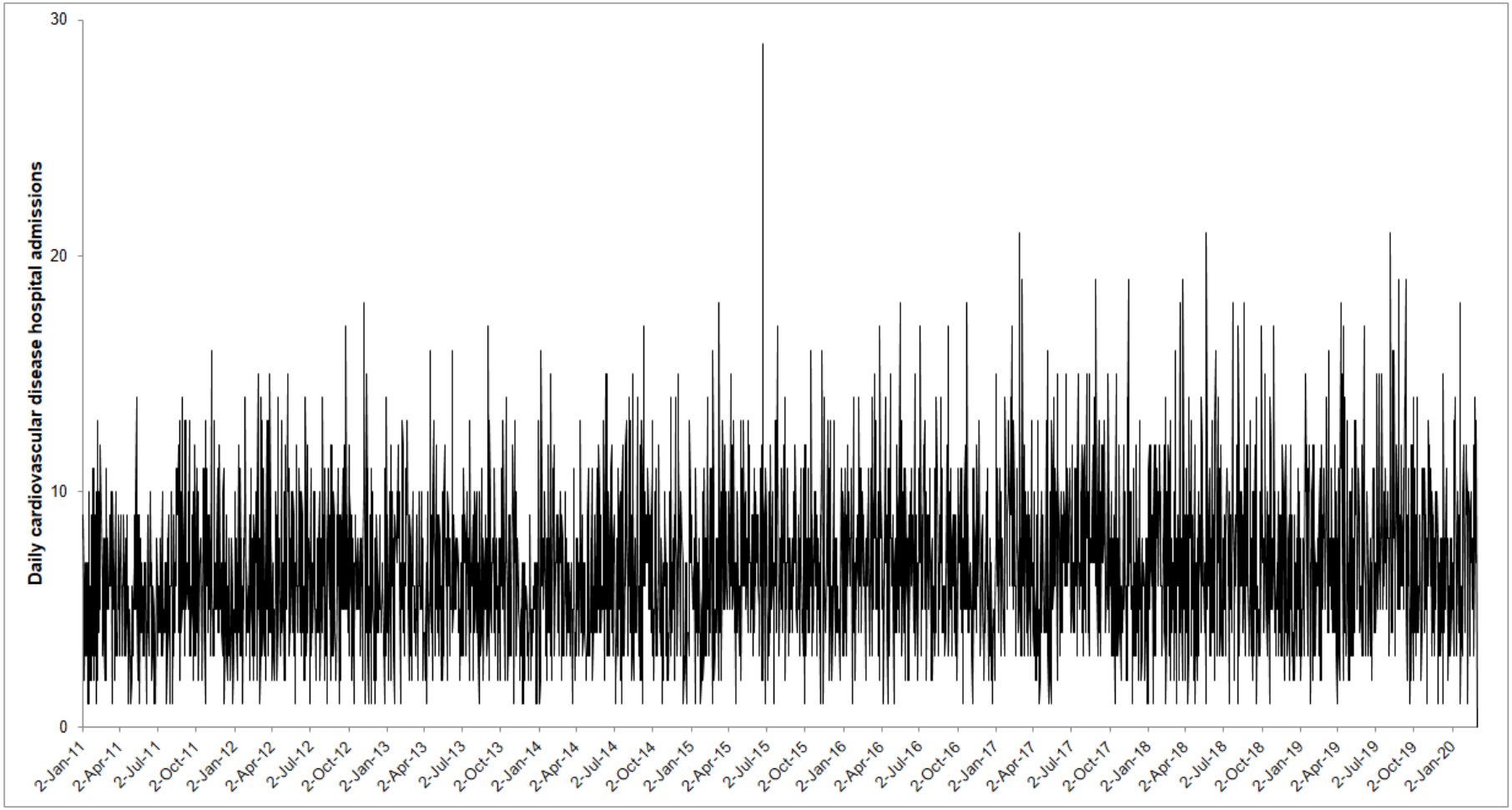


Figure 6.9: Time-series of the daily number of cardiovascular disease hospital admissions in VTAPA, South Africa, 2 January 2011 – 29 February 2020 (3346 days).

6.2.1. FREQUENCY OF RESPIRATORY HOSPITAL ADMISSIONS PER DAYTYPE

Table 6.27 reports the frequency of RD hospital admissions by the 64 day types (all ages and sexes combined) for PM_{10}, NO_2, and SO_2 (mixture 1). There were 54 822 total RD hospital admissions, from which 209 referent days showed 2 709 (4.94%) RD hospital admissions. Day type 444 was recorded as the most frequently observed day type (338 days; 10.1%), i.e. when all three air pollutants had levels in the highest quartile. The most RD hospital admissions also occurred for this day type (6 834; 12.47%). The least hospital admissions occurred at the 414 day type with only six (0.01%) RD hospital admissions. Day type 414 was the least frequently observed, where one out of the 3 346 days (0.3%) had PM_{10} levels in the highest quartile, NO_2 had levels in the lowest quartile, and SO_2 had levels in the highest quartile.

Table 6.27: Frequency of respiratory disease hospital admissions by day types for all ages and sexes combined, PM_{10}, NO_2 and SO_2 (mixture 1).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	2709	4.94	**211**	1396	2.55
112	1421	2.59	**212**	962	1.75
113	697	1.27	**213**	517	0.94
114	536	0.98	**214**	247	0.45
121	1538	2.81	**221**	1245	2.27
122	1013	1.85	**222**	1189	2.17
123	769	1.40	**223**	967	1.76
124	845	1.54	**224**	792	1.44
131	656	1.20	**231**	651	1.19
132	429	0.78	**232**	941	1.72
133	348	0.63	**233**	1378	2.51
134	375	0.68	**234**	1012	1.85
141	250	0.46	**241**	160	0.29
142	109	0.20	**242**	479	0.87
143	105	0.19	**243**	564	1.03
144	172	0.31	**244**	457	0.83
311	738	1.35	**411**	284	0.52
312	556	1.01	**412**	242	0.44
313	434	0.79	**413**	140	0.26
314	130	0.24	**414**	6	0.01
321	802	1.46	**421**	174	0.32
322	991	1.81	**422**	602	1.10
323	934	1.70	**423**	317	0.58
324	424	0.77	**424**	74	0.13
331	465	0.85	**431**	287	0.52
332	1304	2.38	**432**	977	1.78
333	2509	4.58	**433**	961	1.75
334	1523	2.78	**434**	763	1.39
341	217	0.40	**441**	127	0.23
342	625	1.14	**442**	1242	2.27
343	1021	1.86	**443**	2919	5.32
344	1271	2.32	**444**	6834	12.47

Table 6.28 reports the frequency of RD hospital admissions by the 64 day types (all ages and sexes combined) for $PM_{2.5}$, NO_2, and SO_2 (mixture 2). From the 54 822 total RD hospital admissions, there were 216 referent days where 2 795 (5.1%) RD hospital admissions occurred. Day type 444 was recorded as the most frequently observed day type (339 out of the 3 346 days; 10.13%), i.e. when all three air pollutants had levels in the highest quartile. For this day type the most RD hospital admissions also occurred (6 972; 12.72%). The least hospital admissions occurred for the 143 day type with 69 (0.13%) RD hospital admissions. Day type 143 was the least frequently observed (3 out of the 3 346 days; 0.09%), when $PM_{2.5}$ had levels in the lowest quartile, NO_2 had levels in the highest quartile, and SO_2 had levels in the third quartile.

Table 6.28: Frequency of respiratory disease hospital admissions by day types for all ages and sexes combined, $PM_{2.5}$, NO_2 and SO_2 (mixture 2).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	2795	5.10	**211**	1458	2.66
112	1485	2.71	**212**	1043	1.90
113	778	1.42	**213**	628	1.15
114	647	1.18	**214**	162	0.30
121	1606	2.93	**221**	1251	2.28
122	885	1.61	**222**	1218	2.22
123	663	1.21	**223**	1145	2.09
124	718	1.31	**224**	736	1.34
131	718	1.31	**231**	694	1.27
132	410	0.75	**232**	1011	1.84
133	221	0.40	**233**	1335	2.44
134	427	0.78	**234**	692	1.26
141	95	0.17	**241**	301	0.55
142	131	0.24	**242**	478	0.87
143	69	0.13	**243**	381	0.69
144	107	0.20	**244**	362	0.66
311	549	1.00	**411**	325	0.59
312	486	0.89	**412**	167	0.30
313	246	0.45	**413**	136	0.25
314	65	0.12	**414**	45	0.08
321	701	1.28	**421**	201	0.37
322	1211	2.21	**422**	481	0.88
323	926	1.69	**423**	253	0.46
324	530	0.97	**424**	151	0.28
331	447	0.82	**431**	200	0.36
332	1582	2.89	**432**	648	1.18
333	2370	4.32	**433**	1270	2.32
334	1376	2.51	**434**	1178	2.15
341	260	0.47	**441**	98	0.18
342	735	1.34	**442**	1111	2.03
343	1374	2.51	**443**	2785	5.08
344	1293	2.36	**444**	6972	12.72

Table 6.29 reports the frequency of RD hospital admissions by the 64 day types (all ages and sexes combined) for PM_{10}, NO_2, and O_3 (mixture 3). From the 54 822 total RD hospital admissions, there were 79 referent days where 1 140 (2.08%) RD hospital admissions occurred. Day type 441 was recorded as the most frequently observed day type (299; 8.94%), when PM_{10} and NO_2 were in the highest quartile and O_3 was in the lowest quartile. For this day type the most RD hospital admissions also occurred (6 028, 11%). The least hospital admissions occurred for the 411 day type, with 53 (0.1%) RD hospital admissions. Day type 441 was the least frequently observed (3 days; 0.09%), when PM_{10} had levels in the highest quartile, NO_2 had levels in the lowest quartile, and O_3 had levels in the lowest quartile.

Table 6.29: Frequency of respiratory disease hospital admissions by day types for all ages and sexes combined, PM_{10}, NO_2 and O_3 (mixture 3).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	1140	2.08	**211**	406	0.74
112	1073	1.96	**212**	718	1.31
113	1371	2.50	**213**	796	1.45
114	1779	3.25	**214**	1202	2.19
121	602	1.10	**221**	918	1.67
122	963	1.76	**222**	1171	2.14
123	1271	2.32	**223**	1167	2.13
124	1329	2.42	**224**	937	1.71
131	358	0.65	**231**	753	1.37
132	321	0.59	**232**	1335	2.44
133	570	1.04	**233**	1160	2.12
134	559	1.02	**234**	734	1.34
141	252	0.46	**241**	480	0.88
142	238	0.43	**242**	419	0.76
143	68	0.12	**243**	329	0.60
144	78	0.14	**244**	432	0.79
311	165	0.30	**411**	53	0.10
312	530	0.97	**412**	126	0.23
313	624	1.14	**413**	210	0.38
314	539	0.98	**414**	283	0.52
321	440	0.80	**421**	207	0.38
322	1119	2.04	**422**	261	0.48
323	937	1.71	**423**	368	0.67
324	655	1.19	**424**	331	0.60
331	1699	3.10	**431**	709	1.29
332	1677	3.06	**432**	824	1.50
333	1740	3.17	**433**	929	1.69
334	685	1.25	**434**	526	0.96
341	1325	2.42	**441**	6028	11.00
342	1027	1.87	**442**	2453	4.47
343	408	0.74	**443**	1385	2.53
344	374	0.68	**444**	1256	2.29

Table 6.30 reports the frequency of RD hospital admissions by the 64 day types (all ages and sexes combined) for $PM_{2.5}$, NO_2, and O_3 (mixture 4). From the 54 822 total RD hospital admissions, there were 70 referent days where 1 069 (1.95%) RD hospital admissions occurred. Day type 441 was recorded as the most frequently observed day type (304 days; 9.09%), when PM_{10} and NO_2 were in the highest quartile and O_3 was in the lowest quartile. For this day type the most RD hospital admissions also occurred (6 203; 11.31%). The least hospital admissions occurred at the 143 day type with 53 (0.1%) RD hospital admissions. Day type 143 was the least frequently observed (4 days; 0.12%), when PM_{10} had levels in the lowest quartile, NO_2 had levels in the highest quartile, and O_3 had levels in the third quartile.

Table 6.30: Frequency of respiratory disease hospital admissions by day types for all ages and sexes combined, $PM_{2.5}$, NO_2 and O_3 (mixture 4).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	1069	1.95	**211**	514	0.94
112	1130	2.06	**212**	781	1.42
113	1428	2.60	**213**	894	1.63
114	2078	3.79	**214**	1102	2.01
121	422	0.77	**221**	847	1.55
122	913	1.67	**222**	1118	2.04
123	1165	2.13	**223**	1436	2.62
124	1372	2.50	**224**	949	1.73
131	377	0.69	**231**	675	1.23
132	389	0.71	**232**	1179	2.15
133	451	0.82	**233**	1362	2.48
134	526	0.96	**234**	797	1.45
141	154	0.28	**241**	563	1.03
142	110	0.20	**242**	493	0.90
143	53	0.10	**243**	251	0.46
144	85	0.16	**244**	215	0.39
311	117	0.21	**411**	64	0.12
312	266	0.49	**412**	270	0.49
313	511	0.93	**413**	168	0.31
314	452	0.82	**414**	171	0.31
321	695	1.27	**421**	203	0.37
322	1218	2.22	**422**	265	0.48
323	816	1.49	**423**	326	0.59
324	639	1.17	**424**	292	0.53
331	1171	2.14	**431**	1296	2.36
332	1376	2.51	**432**	1213	2.21
333	1484	2.71	**433**	1102	2.01
334	557	1.02	**434**	624	1.14
341	1165	2.13	**441**	6203	11.31
342	1155	2.11	**442**	2379	4.34
343	632	1.15	**443**	1254	2.29
344	710	1.30	**444**	1130	2.06

Table 6.31 reports the frequency of RD hospital admissions by the 64 day types (all ages and sexes combined) for PM_{10}, SO_2, and O_3 (mixture 5). From the total 54 822 total RD hospital admissions, there were 46 referent days where 758 (1.38%) RD hospital admissions occurred. Day type 441 was recorded as the most frequently observed day type (221 days; 6.6%), when PM_{10} and SO_2 had levels in the highest quartile and O_3 was in the lowest quartile. For this day type the most RD hospital admissions occurred (4 507; 8.22%). The least hospital admissions occurred for the 412 day type with 184 (0.34%) RD hospital admissions. Day type 412 occurred 13 out of the 3 346 days (0.39%), when PM_{10} had levels in the highest quartile, SO_2 had levels in the lowest quartile, and O_3 had levels in the second quartile. However, the day type

to occur the least number of times was 411, which occurred 11 out of the 3 346 days (0.33%).

Table 6.31: Frequency of respiratory disease hospital admissions by day types for all ages and sexes combined, PM_{10}, SO_2 and O_3 (mixture 5).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	758	1.38	**211**	529	0.96
112	1122	2.05	**212**	728	1.33
113	1086	1.98	**213**	906	1.65
114	2187	3.99	**214**	1289	2.35
121	532	0.97	**221**	574	1.05
122	617	1.13	**222**	1084	1.98
123	930	1.70	**223**	930	1.70
124	893	1.63	**224**	983	1.79
131	510	0.93	**231**	642	1.17
132	275	0.50	**232**	983	1.79
133	827	1.51	**233**	1067	1.95
134	307	0.56	**234**	734	1.34
141	552	1.01	**241**	812	1.48
142	581	1.06	**242**	848	1.55
143	437	0.80	**243**	549	1.00
144	358	0.65	**244**	299	0.55
311	248	0.45	**411**	187	0.34
312	518	0.94	**412**	184	0.34
313	803	1.46	**413**	227	0.41
314	653	1.19	**414**	274	0.50
321	723	1.32	**421**	750	1.37
322	1068	1.95	**422**	832	1.52
323	1024	1.87	**423**	679	1.24
324	661	1.21	**424**	802	1.46
331	1234	2.25	**431**	1553	2.83
332	1583	2.89	**432**	1011	1.84
333	1491	2.72	**433**	1031	1.88
334	590	1.08	**434**	742	1.35
341	1424	2.60	**441**	4507	8.22
342	1184	2.16	**442**	1637	2.99
343	391	0.71	**443**	955	1.74
344	349	0.64	**444**	578	1.05

Table 6.32 reports the frequency of RD hospital admissions by the 64 day types (all ages and sexes combined) for $PM_{2.5}$, SO_2, and O_3 (mixture 6). From the total of 54 822 RD hospital admissions, there were 41 referent days where 689 (1.26%) RD hospital admissions occurred. Day type 441 was recorded as the most frequently observed day type (236 days; 7.05%), when $PM_{2.5}$ and SO_2 had levels in the highest quartile and O_3 levels in the lowest quartile. For this day type, the most RD hospital admissions occurred (4 860; 8.87%). The least hospital admissions occurred for the 411 day type with 125 (0.23%) RD hospital admissions. Day type 411 occurred 7 out of the 3 346

days (0.21%), when $PM_{2.5}$ had levels in the highest quartile, SO_2 had levels in the lowest quartile, and O_3 had levels in the lowest quartile.

Table 6.32: Frequency of respiratory disease hospital admissions by day types for all ages and sexes combined, $PM_{2.5}$, SO_2 and O_3 (mixture 6).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	689	1.26	**211**	524	0.96
112	914	1.67	**212**	894	1.63
113	1140	2.08	**213**	1125	2.05
114	2471	4.51	**214**	1161	2.12
121	439	0.80	**221**	592	1.08
122	566	1.03	**222**	1070	1.95
123	879	1.60	**223**	1166	2.13
124	1027	1.87	**224**	922	1.68
131	437	0.80	**231**	763	1.39
132	367	0.67	**232**	830	1.51
133	675	1.23	**233**	1217	2.22
134	252	0.46	**234**	679	1.24
141	486	0.89	**241**	789	1.44
142	663	1.21	**242**	595	1.09
143	452	0.82	**243**	362	0.66
144	298	0.54	**244**	206	0.38
311	384	0.70	**411**	125	0.23
312	553	1.01	**412**	191	0.35
313	509	0.93	**413**	248	0.45
314	511	0.93	**414**	260	0.47
321	805	1.47	**421**	743	1.36
322	1256	2.29	**422**	709	1.29
323	1110	2.02	**423**	408	0.74
324	843	1.54	**424**	547	1.00
331	865	1.58	**431**	1874	3.42
332	1524	2.78	**432**	1131	2.06
333	1648	3.01	**433**	876	1.60
334	879	1.60	**434**	563	1.03
341	1160	2.12	**441**	4860	8.87
342	1245	2.27	**442**	1747	3.19
343	477	0.87	**443**	1041	1.90
344	382	0.70	**444**	698	1.27

Table 6.33 reports the frequency of RD hospital admissions by the 64 day types (all ages and sexes combined) for O_3, NO_2, and SO_2 (mixture 7). From the total of 54 822 RD hospital admissions, there were 38 referent days where 576 (1.05%) RD hospital admissions occurred. Day type 144 was recorded as the most frequently observed day type (243 days; 7.26%), which had O_3 levels in the lowest quartile and NO_2 and SO_2 had levels in the highest levels. For this day type the most RD hospital admissions occurred (5 062; 9.23%). The least hospital admissions occurred for the 441 day type with 72 (0.13%) RD hospital admissions. Day type 441 was observed in 5 out of the 3

346 days (0.15%), when O_3 and NO_2 had levels in the highest quartiles and NO_2 had levels in the lowest quartile.

Table 6.33: Frequency of respiratory disease hospital admissions by day types for all ages and sexes combined, O_3, NO_2 and SO_2 (mixture 7).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	576	1.05	**211**	894	1.63
112	520	0.95	**212**	622	1.13
113	391	0.71	**213**	528	0.96
114	277	0.51	**214**	403	0.74
121	423	0.77	**221**	858	1.57
122	573	1.05	**222**	1160	2.12
123	559	1.02	**223**	790	1.44
124	612	1.12	**224**	706	1.29
131	514	0.94	**231**	451	0.82
132	676	1.23	**232**	1058	1.93
133	985	1.80	**233**	1483	2.71
134	1344	2.45	**234**	1165	2.13
141	209	0.38	**241**	349	0.64
142	810	1.48	**242**	761	1.39
143	2004	3.66	**243**	1051	1.92
144	5062	9.23	**244**	1976	3.60
311	1346	2.46	**411**	2311	4.22
312	1028	1.88	**412**	1011	1.84
313	477	0.87	**413**	392	0.72
314	150	0.27	**414**	89	0.16
321	960	1.75	**421**	1518	2.77
322	1153	2.10	**422**	909	1.66
323	1173	2.14	**423**	465	0.85
324	457	0.83	**424**	360	0.66
331	592	1.08	**431**	502	0.92
332	1041	1.90	**432**	876	1.60
333	2042	3.72	**433**	686	1.25
334	724	1.32	**434**	440	0.80
341	124	0.23	**441**	72	0.13
342	341	0.62	**442**	543	0.99
343	724	1.32	**443**	830	1.51
344	1001	1.83	**444**	695	1.27

6.2.2. FREQUENCY OF CARDIOVASCULAR HOSPITAL ADMISSIONS PER DAYTYPE

Table 6.34 reports the frequency of CVD hospital admissions by the 64 day types (for all ages and sexes combined) for air pollutant PM_{10}, NO_2, and SO_2 (mixture 1). There was a total of 22 507 CVD hospital admissions. There were 209 referent days where 1 269 (5.63%) CVD hospital admissions occurred. Day type 444 was recorded as the most frequently observed day type (338 days; 10.1%), i.e. when all three air pollutants

had levels in the highest quartile. At this day type the most CVD hospital admissions also occurred 2 338 (10.38%). The least hospital admissions occurred for the 414 day type with only 5 (0.02%) CVD hospital admissions, when PM_{10} and SO_2 were in the highest quartile and NO_2 was in the lowest quartile.

Table 6.34: Frequency of cardiovascular disease hospital admissions by day types for all ages and sexes combined for PM_{10}, NO_2 and SO_2 (mixture 1).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	1269	5.63	**211**	698	3.10
112	648	2.88	**212**	535	2.38
113	334	1.48	**213**	252	1.12
114	272	1.21	**214**	100	0.44
121	729	3.24	**221**	604	2.68
122	471	2.09	**222**	505	2.24
123	360	1.60	**223**	404	1.79
124	347	1.54	**224**	273	1.21
131	244	1.08	**231**	263	1.17
132	199	0.88	**232**	394	1.75
133	119	0.53	**233**	520	2.31
134	180	0.80	**234**	425	1.89
141	87	0.39	**241**	68	0.30
142	42	0.19	**242**	193	0.86
143	65	0.29	**243**	206	0.91
144	61	0.27	**244**	141	0.63
311	453	2.01	**411**	143	0.63
312	277	1.23	**412**	80	0.36
313	237	1.05	**413**	63	0.28
314	91	0.40	**414**	5	0.02
321	386	1.71	**421**	79	0.35
322	405	1.80	**422**	263	1.17
323	402	1.79	**423**	157	0.70
324	202	0.90	**424**	33	0.15
331	196	0.87	**431**	110	0.49
332	469	2.08	**432**	322	1.43
333	989	4.39	**433**	356	1.58
334	488	2.17	**434**	314	1.39
341	53	0.24	**441**	55	0.24
342	262	1.16	**442**	465	2.06
343	354	1.57	**443**	1021	4.53
344	431	1.91	**444**	2338	10.38

Table 6.35 reports the frequency of CVD hospital admissions by the 64 day types (for all ages and sexes combined) for air pollutant $PM_{2.5}$, NO_2, and SO_2 (mixture 2). From the 22 507 total CVD hospital admissions, there were 216 referent days where 1 328 (5.9%) CVD hospital admissions occurred. Day type 444 was recorded as the most frequently observed day type (339 days; 10.13%), i.e. when all three air pollutants had levels in the highest quartile. For this day type, the most CVD hospital admissions

occurred (2 348; 10.43%). The least hospital admissions occurred for day type 414, with only 22 (0.10%) CVD hospital admissions, when $PM_{2.5}$ and SO_2 were in the highest quartiles and NO_2 was in the lowest quartile.

Table 6.35: Frequency of cardiovascular disease hospital admissions by day types for all ages and sexes combined for $PM_{2.5}$, NO_2 and SO_2 (mixture 2).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	1328	5.90	**211**	741	3.29
112	695	3.09	**212**	518	2.30
113	382	1.70	**213**	305	1.35
114	300	1.33	**214**	95	0.42
121	742	3.29	**221**	566	2.51
122	437	1.94	**222**	515	2.29
123	301	1.34	**223**	504	2.24
124	302	1.34	**224**	295	1.31
131	284	1.26	**231**	230	1.02
132	187	0.83	**232**	398	1.77
133	78	0.35	**233**	519	2.30
134	216	0.96	**234**	312	1.39
141	39	0.17	**241**	104	0.46
142	31	0.14	**242**	198	0.88
143	33	0.15	**243**	148	0.66
144	38	0.17	**244**	113	0.50
311	336	1.49	**411**	158	0.70
312	260	1.15	**412**	67	0.30
313	148	0.66	**413**	51	0.23
314	51	0.23	**414**	22	0.10
321	396	1.76	**421**	94	0.42
322	519	2.30	**422**	173	0.77
323	389	1.73	**423**	129	0.57
324	190	0.84	**424**	68	0.30
331	212	0.94	**431**	87	0.39
332	573	2.54	**432**	226	1.00
333	930	4.13	**433**	457	2.03
334	434	1.93	**434**	445	1.98
341	78	0.35	**441**	42	0.19
342	294	1.31	**442**	439	1.95
343	521	2.31	**443**	944	4.19
344	472	2.10	**444**	2348	10.43

Table 6.36 reports the frequency of CVD hospital admissions by the 64 day types (all ages and sexes combined) for air pollutant PM_{10}, NO_2, and O_3 (mixture 3). From the 22 507 total CVD hospital admissions, there were 79 referent days where 482 (2.14%) CVD hospital admissions occurred. Day type 441 was recorded as the most frequently observed day type (299 days; 8.94%), when PM_{10} and NO_2 were in the highest quartile and O_3 was in the lowest quartile. For this day type the most CVD hospital admissions occurred (2151; 9.55%). The least hospital admissions occurred day type 411, with

only 13 (0.06%) CVD hospital admissions. Day type 441 was the least frequently observed (3 days; 0.09%), when PM_{10} had levels in the highest quartile, NO_2 had levels in the lowest quartile and O_3 had levels in the lowest quartile.

Table 6.36: Frequency of cardiovascular disease hospital admissions by day types for all ages and sexes combined for PM_{10}, NO_2 and O_3 (mixture 3).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	482	2.14	**211**	167	0.74
112	493	2.19	**212**	318	1.41
113	608	2.70	**213**	433	1.92
114	940	4.17	**214**	667	2.96
121	230	1.02	**221**	334	1.48
122	397	1.76	**222**	510	2.26
123	578	2.57	**223**	517	2.30
124	702	3.12	**224**	425	1.89
131	129	0.57	**231**	285	1.27
132	137	0.61	**232**	486	2.16
133	235	1.04	**233**	468	2.08
134	241	1.07	**234**	363	1.61
141	98	0.44	**241**	182	0.81
142	76	0.34	**242**	138	0.61
143	36	0.16	**243**	118	0.52
144	45	0.20	**244**	170	0.75
311	80	0.36	**411**	13	0.06
312	274	1.22	**412**	56	0.25
313	351	1.56	**413**	110	0.49
314	353	1.57	**414**	112	0.50
321	176	0.78	**421**	69	0.31
322	495	2.20	**422**	117	0.52
323	403	1.79	**423**	152	0.67
324	321	1.43	**424**	194	0.86
331	534	2.37	**431**	237	1.05
332	641	2.85	**432**	370	1.64
333	678	3.01	**433**	297	1.32
334	289	1.28	**434**	198	0.88
341	462	2.05	**441**	2151	9.55
342	371	1.65	**442**	840	3.73
343	135	0.60	**443**	464	2.06
344	132	0.59	**444**	424	1.88

Table 6.37 reports the frequency of CVD hospital admissions by the 64 day types (for all ages and sexes combined) for air pollutant $PM_{2.5}$, NO_2, and O_3 (mixture 4). From the total 22 507 CVD hospital admissions, there were 70 referent days where 414 (1.84%) CVD hospital admissions occurred. Day type 441 was recorded as the most frequently observed day type (304 days; 9.09%), when PM_{10} and NO_2 were in the highest quartile and O_3 was in the lowest quartile. For this day type the most CVD hospital admissions occurred (2 188; 9.72%). The least hospital admissions occurred for day type 411, with

22 (0.1%) CVD hospital admissions when PM_{10} levels were in the highest quartile, and NO_2 and O_3 had levels in the lowest quartile.

Table 6.37: Frequency of cardiovascular disease hospital admissions by day types for all ages and sexes combined for $PM_{2.5}$, NO_2 and O_3 (mixture 4).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	414	1.84	**211**	247	1.10
112	547	2.43	**212**	321	1.43
113	669	2.97	**213**	463	2.06
114	1075	4.77	**214**	628	2.79
121	163	0.72	**221**	303	1.35
122	392	1.74	**222**	474	2.10
123	542	2.41	**223**	587	2.61
124	685	3.04	**224**	516	2.29
131	143	0.63	**231**	238	1.06
132	162	0.72	**232**	427	1.90
133	210	0.93	**233**	513	2.28
134	236	1.05	**234**	383	1.70
141	60	0.27	**241**	206	0.91
142	24	0.11	**242**	160	0.71
143	30	0.13	**243**	91	0.40
144	27	0.12	**244**	106	0.47
311	59	0.26	**411**	22	0.10
312	151	0.67	**412**	122	0.54
313	291	1.29	**413**	79	0.35
314	294	1.31	**414**	75	0.33
321	277	1.23	**421**	66	0.29
322	528	2.34	**422**	125	0.56
323	374	1.66	**423**	147	0.65
324	315	1.40	**424**	126	0.56
331	363	1.61	**431**	441	1.96
332	542	2.41	**432**	503	2.23
333	575	2.55	**433**	380	1.69
334	239	1.06	**434**	233	1.03
341	439	1.95	**441**	2188	9.72
342	436	1.94	**442**	805	3.57
343	225	1.00	**443**	407	1.81
344	265	1.18	**444**	373	1.66

Table 6.38 reports the frequency of CVD hospital admissions by the 64 day types (all ages and sexes combined) for air pollutant PM_{10}, SO_2, and O_3 (mixture 5). From the total 22 507 CVD hospital admissions, there were 46 referent days where 267 (1.19%) CVD hospital admissions occurred. Day type 441 was recorded as the most frequently observed day type (221 days; 6.6%), when PM_{10} and SO_2 had levels in the highest quartile and O_3 levels in the lowest quartile. For this day type the most CVD hospital admissions also occurred (1 588; 7.05%). The least hospital admissions occurred at the 411 day type with 73 (0.32%) CVD hospital admissions, when PM_{10} levels were in

the highest quartile, SO_2 levels in the lowest quartile, and O_3 levels in the second quartile.

Table 6.38: Frequency of cardiovascular disease hospital admissions by day types for all ages and sexes combined for PM_{10}, SO_2 and O_3 (mixture 5).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	267	1.19	**211**	223	0.99
112	389	1.73	**212**	327	1.45
113	507	2.25	**213**	425	1.89
114	1166	5.18	**214**	658	2.92
121	257	1.14	**221**	258	1.15
122	265	1.18	**222**	439	1.95
123	392	1.74	**223**	433	1.92
124	446	1.98	**224**	497	2.21
131	207	0.92	**231**	222	0.99
132	152	0.67	**232**	376	1.67
133	337	1.50	**233**	455	2.02
134	182	0.81	**234**	329	1.46
141	208	0.92	**241**	265	1.18
142	297	1.32	**242**	310	1.38
143	221	0.98	**243**	223	0.99
144	134	0.60	**244**	141	0.63
311	96	0.43	**411**	73	0.32
312	252	1.12	**412**	110	0.49
313	334	1.48	**413**	101	0.45
314	406	1.80	**414**	103	0.46
321	262	1.16	**421**	283	1.26
322	449	1.99	**422**	279	1.24
323	416	1.85	**423**	243	1.08
324	286	1.27	**424**	325	1.44
331	442	1.96	**431**	526	2.34
332	629	2.79	**432**	404	1.79
333	656	2.91	**433**	395	1.75
334	255	1.13	**434**	272	1.21
341	452	2.01	**441**	1588	7.05
342	451	2.00	**442**	590	2.62
343	161	0.71	**443**	284	1.26
344	148	0.66	**444**	228	1.01

Table 6.39 reports the frequency of CVD hospital admissions by the 64 day types (for all ages and sexes combined) for air pollutant $PM_{2.5}$, SO_2, and O_3 (mixture 6). From the total 22 507 CVD hospital admissions, there were 41 referent days where 235 (1.04%) CVD hospital admissions occurred. Day type 441 was recorded as the most frequently observed day type (236 days; 7.05%), when $PM_{2.5}$ and SO_2 had levels in the highest quartile and O_3 was in the lowest quartile. For this day type the most CVD hospital admissions also occurred (1 700; 7.55%). The least hospital admissions occurred for day type 411 with 35 (0.16%) CVD hospital admissions. Day type 411 occurred on 7 of

the 3 346 days (0.21%), when $PM_{2.5}$ had levels in the highest quartile, SO_2 levels in the lowest quartile, and O_3 levels in the lowest quartile.

Table 6.39: Frequency of cardiovascular disease hospital admissions by day types for all ages and sexes combined for $PM_{2.5}$, SO_2 and O_3 (mixture 6).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	235	1.04	**211**	229	1.02
112	343	1.52	**212**	312	1.39
113	523	2.32	**213**	469	2.08
114	1292	5.74	**214**	631	2.80
121	204	0.91	**221**	258	1.15
122	269	1.19	**222**	445	1.98
123	391	1.74	**223**	446	1.98
124	486	2.16	**224**	480	2.13
131	177	0.79	**231**	269	1.19
132	161	0.71	**232**	333	1.48
133	330	1.47	**233**	502	2.23
134	126	0.56	**234**	372	1.65
141	172	0.76	**241**	274	1.22
142	335	1.49	**242**	251	1.11
143	227	1.01	**243**	183	0.81
144	122	0.54	**244**	107	0.48
311	160	0.71	**411**	35	0.16
312	301	1.34	**412**	122	0.54
313	253	1.12	**413**	122	0.54
314	308	1.37	**414**	102	0.45
321	290	1.29	**421**	308	1.37
322	464	2.06	**422**	254	1.13
323	506	2.25	**423**	141	0.63
324	386	1.71	**424**	202	0.90
331	333	1.48	**431**	618	2.74
332	662	2.94	**432**	405	1.80
333	659	2.93	**433**	352	1.56
334	334	1.48	**434**	206	0.91
341	367	1.63	**441**	1700	7.55
342	439	1.95	**442**	623	2.77
343	171	0.76	**443**	308	1.37
344	170	0.75	**444**	252	1.12

Table 6.40 reports the frequency of CVD hospital admissions by the 64 day types (all ages and sexes combined), for air pollution O_3, NO_2, and SO_2 (mixture 7). From the total 22 507 CVD hospital admissions, there were 38 referent days where 219 (0.97%) CVD hospital admissions occurred. Day type 144 was recorded as the most frequently observed day type (243 days; 7.26%), when O_3 had levels in the lowest quartile, and NO_2 and SO_2 had levels in the highest quartile. The most CVD hospital admissions occurred for this day type (1 748; 7.76%). The least hospital admissions occurred for day type 341 with 29 (0.13%) CVD hospital admissions. When O_3 levels were in the

third quartile, NO_2 had levels in the highest quartile, and SO_2 had levels in the lowest quartile.

Table 6.40: Frequency of cardiovascular disease hospital admissions by day types for all ages and sexes combined for O_3, NO_2 and SO_2 (mixture 7).

Day type	Frequency	Percent	Day type	Frequency	Percent
111	219	0.97	**211**	401	1.78
112	269	1.19	**212**	273	1.21
113	152	0.67	**213**	236	1.05
114	102	0.45	**214**	231	1.03
121	177	0.79	**221**	375	1.67
122	220	0.98	**222**	529	2.35
123	199	0.88	**223**	341	1.51
124	213	0.95	**224**	274	1.22
131	198	0.88	**231**	168	0.75
132	209	0.93	**232**	371	1.65
133	328	1.46	**233**	593	2.63
134	450	2.00	**234**	502	2.23
141	65	0.29	**241**	134	0.60
142	362	1.61	**242**	259	1.15
143	718	3.19	**243**	391	1.74
144	1748	7.76	**244**	641	2.85
311	658	2.92	**411**	1285	5.71
312	477	2.12	**412**	521	2.31
313	282	1.25	**413**	216	0.96
314	85	0.38	**414**	50	0.22
321	439	1.95	**421**	807	3.58
322	473	2.10	**422**	422	1.87
323	519	2.30	**423**	264	1.17
324	219	0.97	**424**	149	0.66
331	241	1.07	**431**	206	0.91
332	401	1.78	**432**	403	1.79
333	763	3.39	**433**	300	1.33
334	273	1.21	**434**	182	0.81
341	29	0.13	**441**	35	0.16
342	133	0.59	**442**	208	0.92
343	279	1.24	**443**	258	1.15
344	312	1.39	**444**	270	1.20

6.3. ASSOCIATION OF JOINT EFFECTS OF AIR POLLUTION ON RESPIRATORY DISEASE HOSPITAL ADMISSIONS

6.3.1. PM_{10}, NO_2 and SO_2 (MIXTURE 1)

Figure 6.10 shows the classification and regression tree of RD hospital admissions all ages and both sexes combined for mixture PM_{10}, NO_2 and SO_2. Four terminal nodes were identified. The levels of PM_{10}, NO_2, and SO_2 increase from left to right, e.g. node 3 includes mixtures (day types) with higher NO_2 levels compared to node 2.

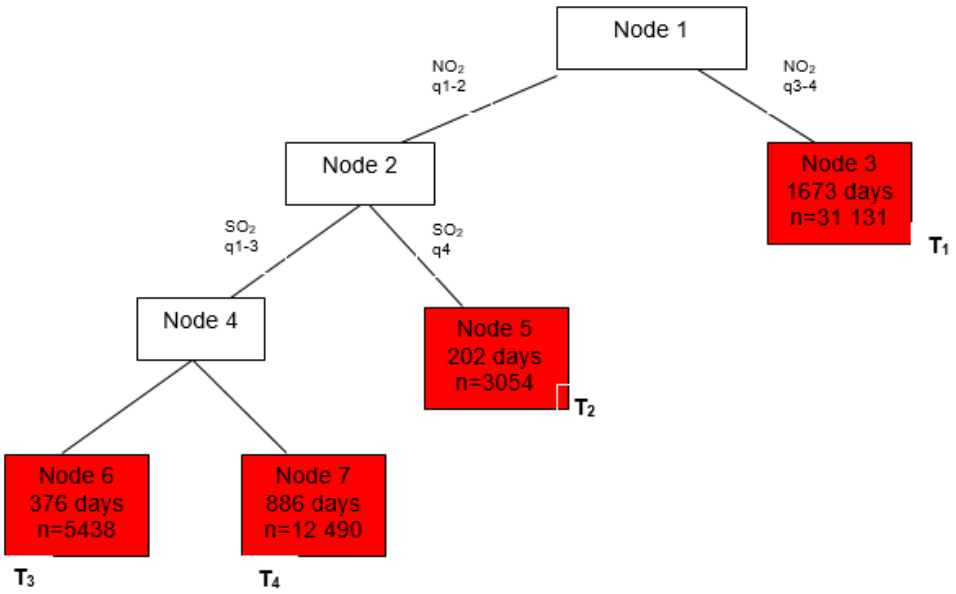


Figure 6.10: Classification and regression tree of respiratory disease hospital admissions modelled for all ages and both sexes combined (mixture PM_{10}, NO_2 and SO_2). n= number of respiratory hospital admissions in the terminal node

Table 6.41 reports on the joint effects of PM_{10}, NO_2, and SO_2 on RD hospital admissions, obtained in the adjusted regression models. Terminal node 1 was found to significantly increase RD hospitalisation by a rate ratio (RR) of 1.04 (95% CI: 1.01, 1.08). This mix included PM_{10} and SO_2 in all four quartiles, while NO_2 was in the two highest quartiles.

Table 6.41: Joint effects of PM_{10}, NO_2 and SO_2 on respiratory disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**PM_{10}**	**NO_2**	**SO_2**
Referent group	209	2709	1.0	1	1	1
T1	1673	31 131	**1.04(1.01-1.08)**	1-4	3-4	1-4
T2	202	3054	1.03(0.98-1.09)	1-4	1-2	4
T3	376	5438	0.98(0.95-1.01)	1	1-2	1-3
T4	886	12 490	0.99(0.92-1.07)	2-4	1-2	1-3

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of RD hospital admissions in the terminal node, adds up to 54 822
Bold: Significant p<0.05

6.3.2. $PM_{2.5}$, NO_2 AND SO_2 (MIXTURE 2)

Figure 6.11 depicts the classification and regression tree of RD hospital admissions for all ages and both sexes combined for mixture 2, $PM_{2.5}$, NO_2, and SO_2. Three terminal nodes were identified.

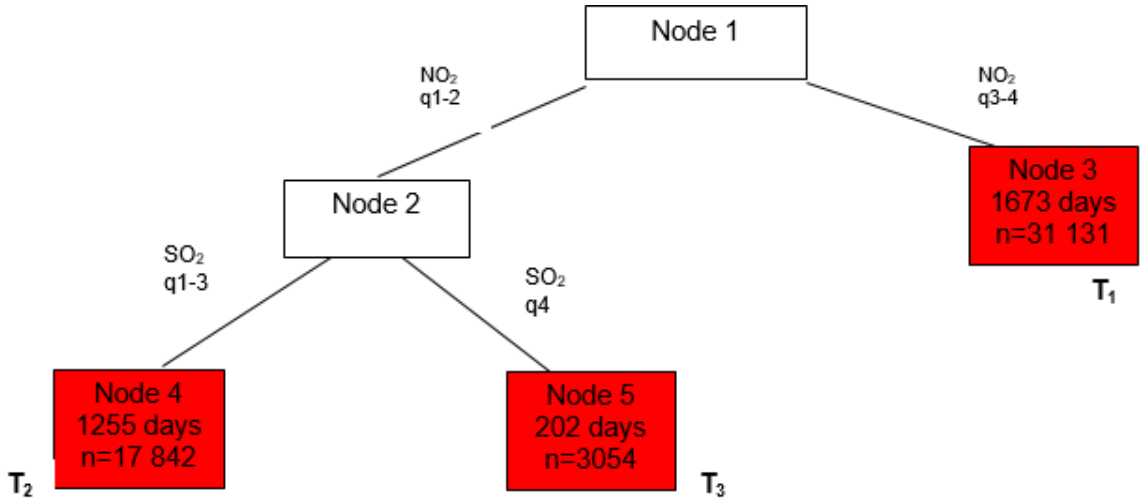


Figure 6.11: Classification and regression tree of respiratory disease hospital admissions modelled for all ages and both sexes combined, (mixture $PM_{2.5}$, NO_2 and SO_2). n= number of respiratory hospital admissions in the terminal node

Table 6.42 illustrates the joint effects of $PM_{2.5}$, NO_2, and SO_2 on RD hospital admissions, obtained in the adjusted regression models. Terminal node 1 was found to significantly increase RD hospitalisation by an RR of 1.04 (95% CI: 1.01, 1.08). This mix included $PM_{2.5}$ and SO_2 in all four quartiles, while NO_2 was in the two highest quartiles. This mix included $PM_{2.5}$ in all four quartiles, while NO_2 was in the two lowest quartiles, and SO_2 in the first three quartiles.

Table 6.42: Joint effects of $PM_{2.5}$, NO_2 and SO_2 on respiratory disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**$PM_{2.5}$**	**NO_2**	**SO_2**
Referent group	216	2795	1.0	1	1	1
T1	1673	31 131	**1.04 (1.01–1.08)**	1-4	3-4	1-4
T2	1255	17 842	0.97 (0.94–1.00)	1-4	1-2	1-3
T3	202	3054	1.03 (0.98–1.09)	2-4	1-2	4

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of RD hospital admissions in the terminal node, adds up to 54 822
Bold: Significant p<0.05

6.3.3. PM_{10}, NO_2 AND O_3 (MIXTURE 3)

Figure 6.12 depicts the classification and regression tree of RD hospital admissions for all ages and both sexes combined for mixture PM_{10}, NO_2, and O_3, where six terminal nodes were identified.

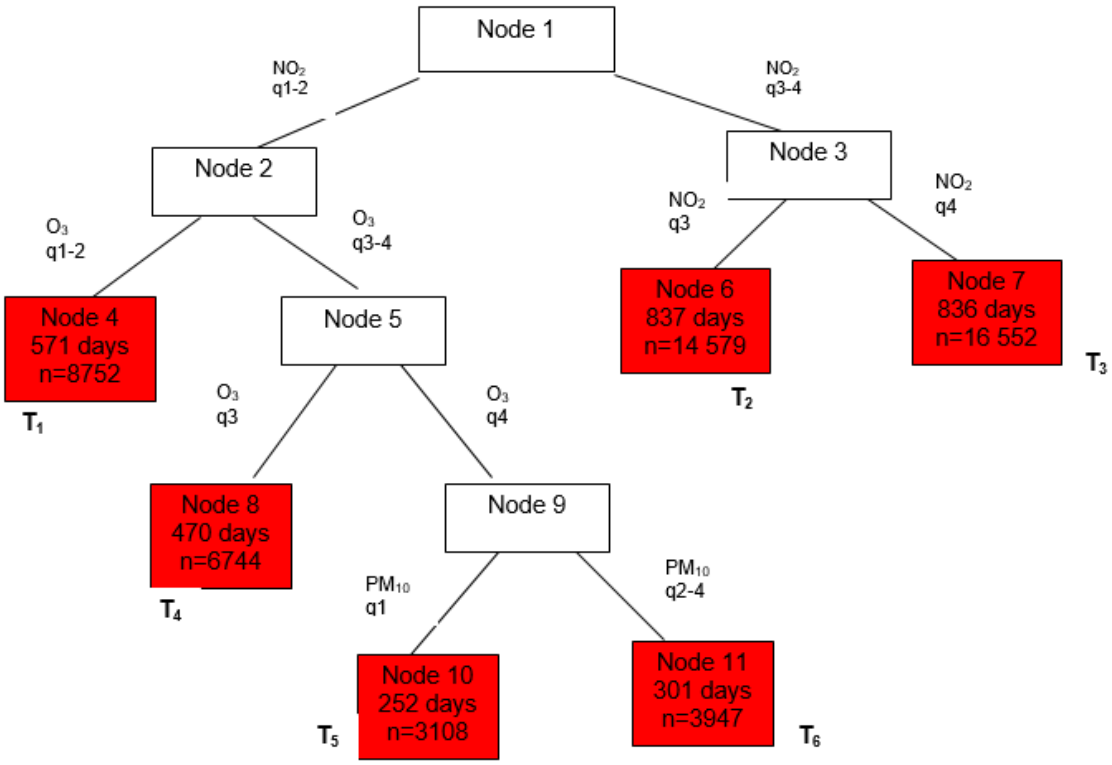


Figure 6.12: Classification and regression tree of respiratory disease hospital admissions modelled for all ages and both sexes combined (mixture PM_{10}, NO_2 and O_3). n= number of respiratory hospital admissions in the terminal node

Table 6.43 reports the joint effects of PM_{10}, NO_2, and O_3 on RD hospital admissions, obtained in the adjusted regression models. Only one of the six identified nodes were found to be significant. Terminal node 1 mixture significantly affected RD hospitalisation by an RR of 0.94 (95% CI: 0.88, 0.99) respectively. Terminal node 1 had NO_2 and O_3 in the lower quartiles and PM_{10} in all four quartiles. Exposure to mixture 1 suggest that the risk of RD hospitalisation was lower in comparison to exposure to PM_{10}, NO_2 and O_3 in the referent group.

Table 6.43: Joint effects of PM_{10}, NO_2 and O_3 on respiratory disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal Node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**PM_{10}**	**NO_2**	**O_3**
Referent Group	79	1140	1.0	1	1	1
T1	571	8752	0.96 (0.93–1.00**)**	1-4	1-2	1-2
T2	837	14579	1.01 (0.98–1.04)	1-4	3	1-4
T3	836	16552	1.03 (0.99–1.06)	1-4	4	1-4
T4	470	6744	1.02 (0.98-1.06)	1-4	1-2	3
T5	252	3108	**0.94 (0.88-0.99)**	1	1-2	4
T6	301	3947	1.01 (0.96-1.06)	2-4	1-2	4

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of RD hospital admissions in the terminal node, adds up to 54 822
Bold: Significant p<0.05

6.3.4. $PM_{2.5}$, NO_2 AND O_3 (MIXTURE 4)

Figure 6.13 depicts the classification and regression tree of RD hospital admissions for all ages and both sexes combined for mixture $PM_{2.5}$, NO_2, and O_3, where six terminal nodes were identified.

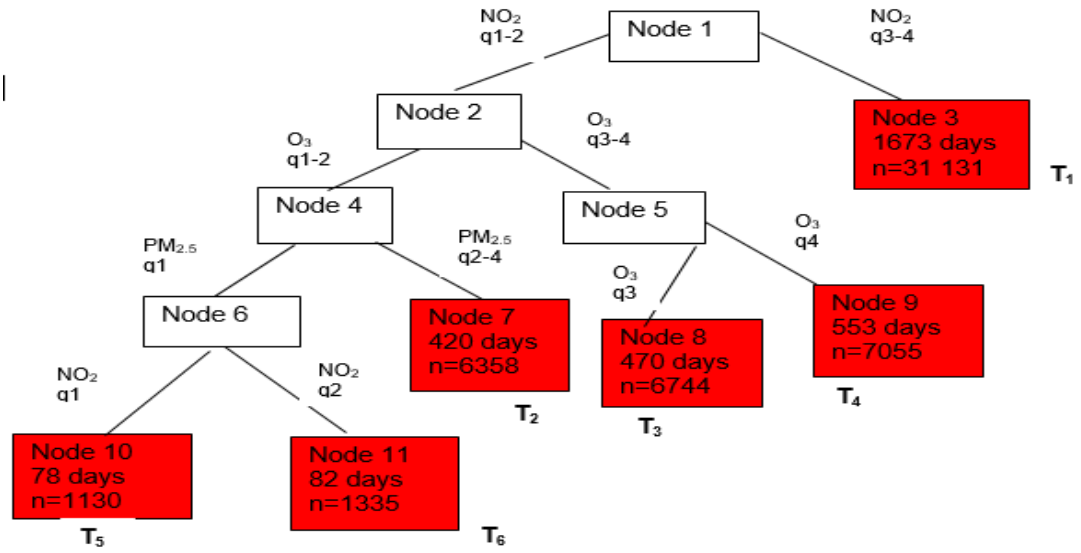


Figure 6.13: Classification and regression tree of respiratory disease hospital admissions modelled for all ages and both sexes combined (mixture $PM_{2.5}$, NO_2 and O_3). n= number of respiratory hospital admissions in the terminal node

Table 6.44 reports on the joint effects of $PM_{2.5}$, NO_2, and O_3 on RD hospital admissions, obtained in the adjusted regression models. Of the six terminal nodes identified, only one was found to significantly increase RD hospital admissions. Terminal node mixture 1 showed $PM_{2.5}$ and O_3 in all quartiles, while NO_2 was in the higher quartiles. Exposure to terminal node 1 showed to significantly increase RD hospital admission with an RR of 1.04 (95% CI: 1.01, 1.08).

Table 6.44: Joint effects of $PM_{2.5}$, NO_2 and O_3 on respiratory disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**$PM_{2.5}$**	**NO_2**	**O_3**
Referent group	70	1069	1.0	1	1	1
T1	1673	31131	**1.04 (1.01–1.08)**	1-4	3-4	1-4
T2	420	6358	0.96 (0.92–1.00)	2-4	1-2	1-2
T3	470	6744	1.02 (0.98–1.06)	1-4	1-2	3
T4	553	7055	0.98 (0.94-1.02)	1-4	1-2	4
T5	78	1130	94 (0.86-1.02)	1	1	1-2
T6	82	1335	1.00 (0.93–1.09)	1	2	1-2

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of RD hospital admissions in the terminal node, adds up to 54 822
Bold: Significant p<0.05

6.3.5. PM_{10}, SO_2 AND O_3 (MIXTURE 5)

Figure 6.14 depicts the classification and regression tree of RD hospital admissions for all ages and both sexes combined for mixture PM_{10}, SO_2, and O_3. There were seven terminal nodes identified, of which only two showed to be significant.

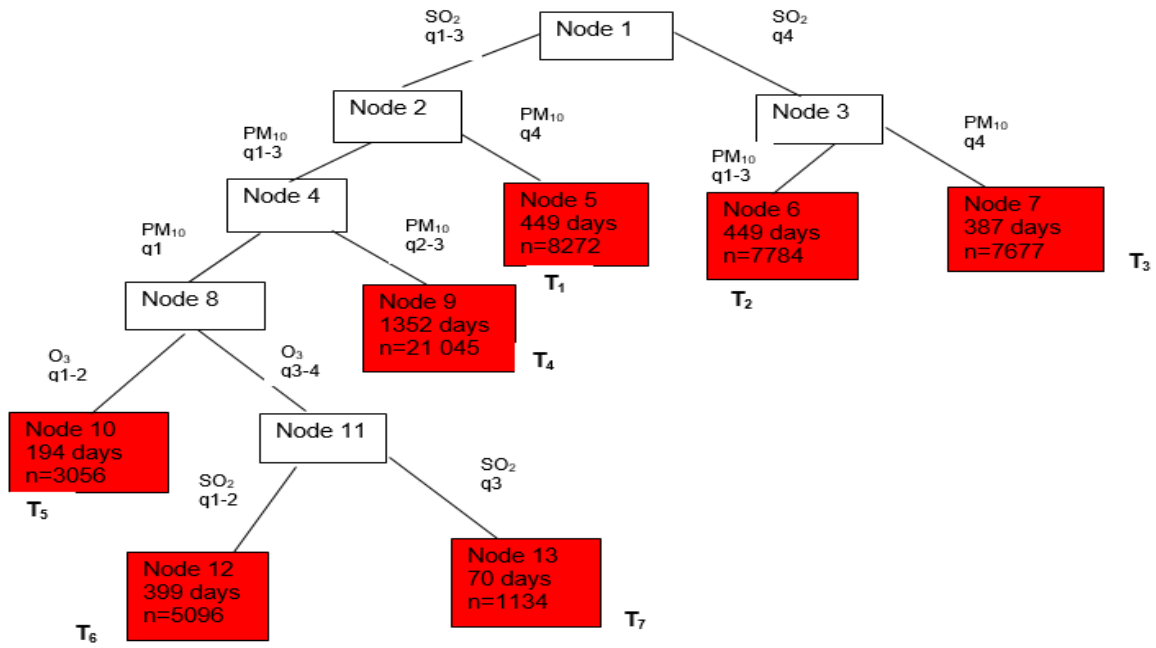


Figure 6.14: Classification and regression tree of respiratory disease hospital admissions modelled for all ages and both sexes combined (mixture PM_{10}, SO_2 and O_3). n= number of respiratory hospital admissions in the terminal node

Table 6.45 reports on the joint effects of PM_{10}, SO_2, and O_3 on RD hospital admissions, obtained in the adjusted regression models. No mixtures were found to be significant.

Table 6.45: Joint effects of PM_{10}, SO_2 and O_3 on respiratory disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**PM_{10}**	**SO_2**	**O_3**
Referent group	46	758	1.0	1	1	1
T1	449	8272	1.01 (0.97–1.05)	4	1-3	1-4
T2	387	7677	1.02 (0.99–1.06)	1-3	4	1-4
T3	449	7784	0.99 (0.96–1.02)	4	4	1-4
T4	1352	1352	0.98 (0.96–1.01)	2-3	1-3	1-4
T5	194	3056	0.94 (0.89–0.99)	1	1-3	1-2
T6	399	5096	0.95 (0.9-1.00)	1	1-2	3-4
T7	70	1134	1.08 (0.99–1.18)	1	3	3-4

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of RD hospital admissions in the terminal node, adds up to 54 822
Bold: Significant p<0.05

6.3.6. $PM_{2.5}$, SO_2 and O_3 (MIXTURE 6)

Figure 6.15 depicts the classification and regression tree of RD hospital admissions for all ages and both sexes combined. Seven terminal nodes were identified and none were found to be significant.

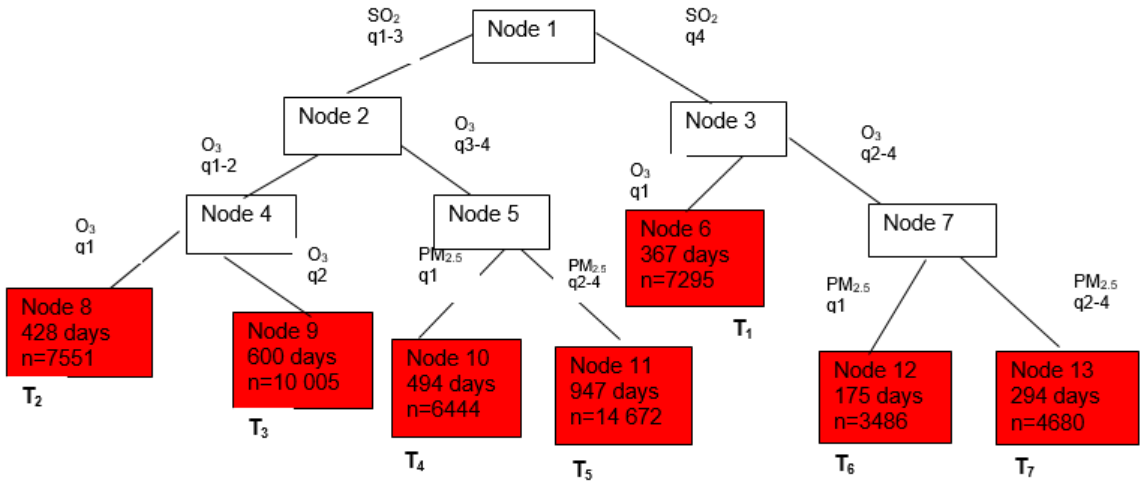


Figure 6.15: Classification and regression tree of respiratory disease hospital admissions modelled for all ages and both sexes combined ($PM_{2.5}$, SO_2 and O_3). n= number of respiratory hospital admissions in the terminal node

Table 6.46 reports the joint effects of $PM_{2.5}$, SO_2, and O_3 on RD hospital admissions, obtained in the adjusted regression models. No mixtures were found to significantly increase RD hospital admissions.

Table 6.46. Joint effects of $PM_{2.5}$, SO_2 and O_3 on respiratory disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**$PM_{2.5}$**	**SO_2**	**O_3**
Referent group	41	689	1.0	1	1	1
T1	367	7295	1.04 (1.00–1.08)	1-4	4	1
T2	428	7551	0.97 (0.93–1.01)	1-4	1-3	1
T3	600	10005	0.97 (0.93–1.00)	1-4	1-3	2
T4	494	6444	0.97 (0.93–1.02)	1	1-3	3-4
T5	947	14672	1.02 (0.99–1.05)	2-4	1-3	3-4
T6	175	3486	1.02 (0.97–1.07)	1	4	2-4
T7	294	4680	1.05 (0.99–1.10)	2-4	4	2-4

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of RD hospital admissions in the terminal node, adds up to 54 822
Bold: Significant p<0.05

6.3.7. O_3, NO_2 AND SO_2 (MIXTURE 7)

Figure 6.16 depicts the classification and regression tree of RD hospital admissions for all ages and both sexes combined for mixture O_3, NO_2, and SO_2. Six terminal nodes were identified and two of the six mixtures were significant.

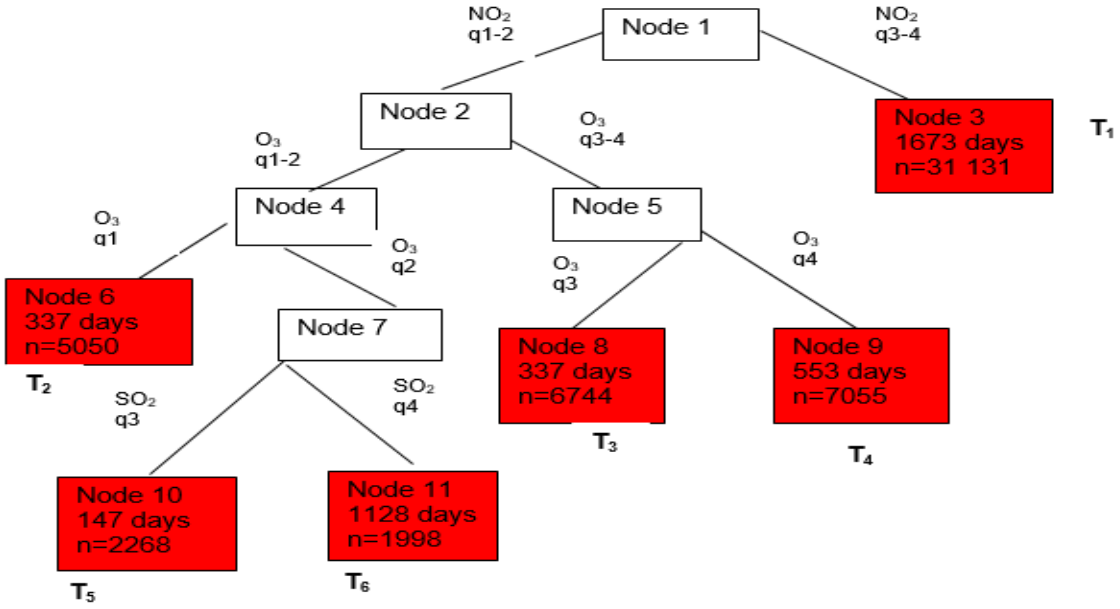


Figure 6.16: Classification and regression tree of respiratory disease hospital admissions modelled for all ages and both sexes combined (mixture O_3, NO_2 and SO_2).
n= number of respiratory hospital admissions in the terminal node

Table 6.47 reports the joint effects of O_3, NO_2, and SO_2 on RD hospital admissions, obtained in the adjusted regression models. Terminal node mixtures 1 and 2 were found to significantly increase RD hospital admissions with RRs of 1.04 (95% CI: 1.01, 1.08) and 0.93 (95% CI: 0.89, 0.97), respectively. Node 1 showed O_3 and SO_2 in all four quartiles and NO_2 in the higher quartiles. Exposure to this terminal node mixture showed an increased risk of RD hospitalisation. Terminal node mixture 2 showed all three pollutants in the two lower quartiles. Results suggest that RD hospital admissions were lower when exposed to terminal node mixture 2 when comparing exposure to pollutants in the referent group.

Table 6.47: Joint effects of O_3, NO_2 and SO_2 on respiratory disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**O_3**	**NO_2**	**SO_2**
Referent group	38	576	1.0	1	1	1
T1	1673	31131	**1.04 (1.01–1.08)**	1-4	3-4	1-4
T2	337	5050	**0.93 (0.89-0.97)**	1-2	1-2	1-2
T3	337	6744	1.02 (0.98–1.06)	3	1-2	1-4
T4	553	7055	0.97 (0.93–1.01)	4	1-2	1-4
T5	147	2268	0.97 (0.92–1.04)	1-2	1-2	3
T6	128	1998	1.06 (0.99–1.13)	1-2	1-2	4

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of RD hospital admissions in the terminal node, adds up to 54 822
Bold: Significant p<0.05

6.4. ASSOCIATION OF JOINT EFFECTS OF AIR POLLUTION ON CARDIOVASCULAR DISEASE HOSPITAL ADMISSIONS

6.4.1. PM_{10}, NO_2 and SO_2 (MIXTURE 1)

Figure 6.17 depicts the classification and regression tree of CVD hospital admissions for all ages and both sexes combined. Only one terminal node was identified.

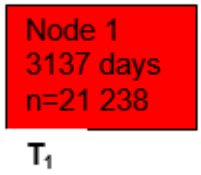


Figure 6.17: Classification and regression tree of cardiovascular disease hospital admissions modelled for all ages and both sexes combined (mixture PM_{10}, NO_2 and SO_2). n= number of cardiovascular hospital admissions in the terminal node

Table 6.48 reports on the joint effects of PM_{10}, NO_2, and SO_2 on CVD hospital admissions, obtained in the adjusted regression models. The joint effects of PM_{10}, NO_2, and SO_2 did not significantly increase CVD hospital admissions in patients.

Table 6.48: Joint effects of PM_{10}, NO_2 and SO_2 on cardiovascular disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models.

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal Node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**PM_{10}**	**NO_2**	**SO_2**
Referent Group	209	1269	1.0	1	1	1
T1	3137	21 238	1.01 (0.93–1.09)	1-4	1-4	1-4

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of CVD hospital admissions in the terminal node, adds up to 22 507
Bold: Significant p<0.05

6.4.2. $PM_{2.5}$, NO_2 AND SO_2 (MIXTURE 2)

Figure 6.18 depicts the classification and regression tree of CVD hospital admissions for all ages and both sexes combined. Ten terminal nodes were identified and only two were found to be significant.

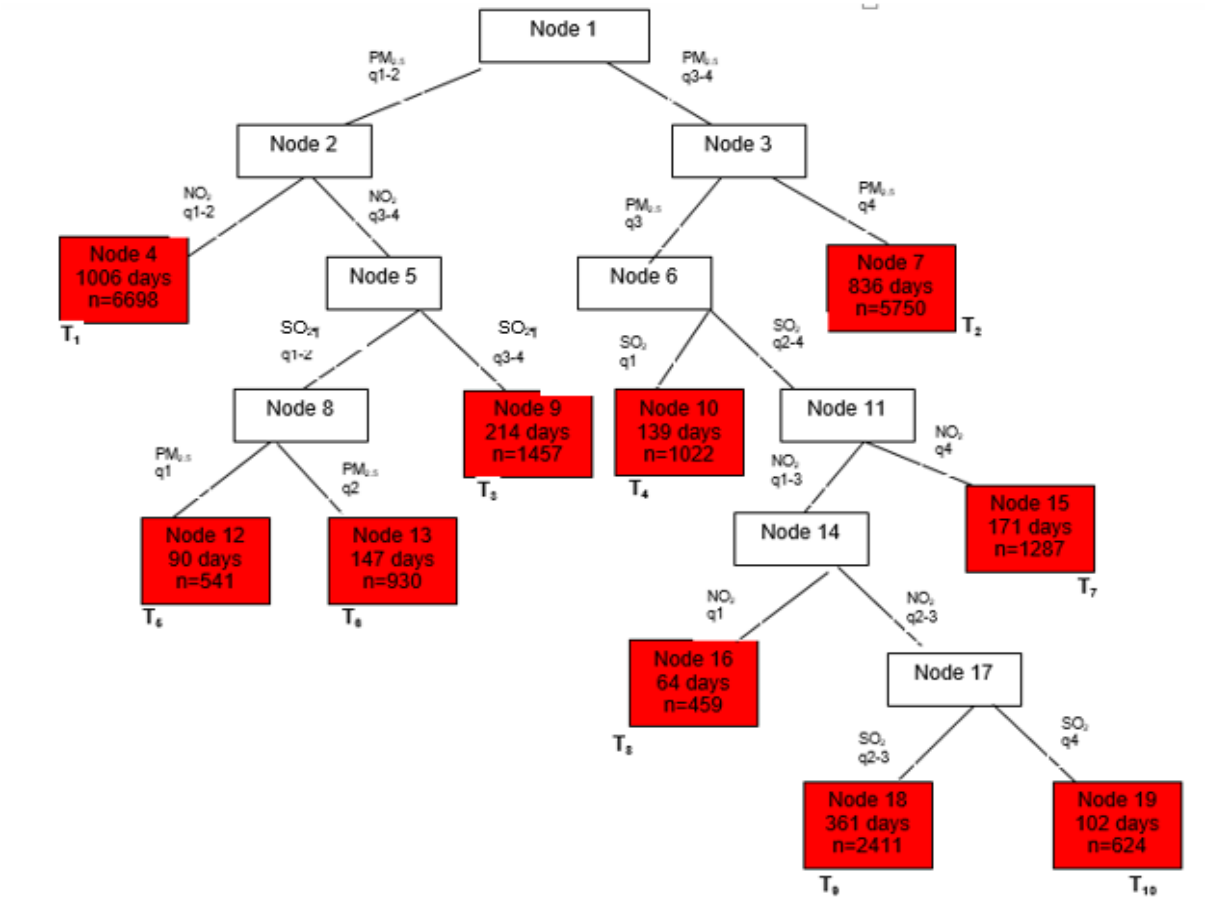


Figure 6.18: Classification and regression tree of cardiovascular disease hospital admissions modelled for all ages and both sexes combined ($PM_{2.5}$, NO_2 and SO_2).
n= number of cardiovascular hospital admissions in the terminal node

Table 6.49 reports on the joint effects of $PM_{2.5}$, NO_2, and SO_2 on CVD hospital admissions, obtained in the adjusted regression models. Terminal node 4 and 10 were found to be significant with the RRs of 1.11 (95% CI: 1.02, 1.20) and 0.86 (95% CI: 0.78, 0.96), respectively. Terminal node mixture 3 showed $PM_{2.5}$ in the third quartile, NO_2 in all four quartiles, and SO_2 in the lowest quartile. Exposure to terminal node mixture 4 suggested an increased risk of CVD hospital admission. Terminal node mixture 10 also showed $PM_{2.5}$ in the third quartile, NO_2 in the second and third quartiles, and SO_2 in the highest quartile. Exposure to terminal node mixture 10 suggests that the risk of CVD hospitalisation was lower than when patients were exposed to $PM_{2.5}$, NO_2, and SO_2 in the referent group.

Table 6.49: Joint effects of $PM_{2.5}$, NO_2 and SO_2 on cardiovascular disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models.

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**$PM_{2.5}$**	**NO_2**	**SO_2**
Referent group	216	1329	1.0	1	1	1
T1	1006	6698	1.00 (0.96–1.05)	1-2	1-2	1-4
T2	836	5750	1.00 (0.95–1.05)	4	1-4	1-4
T3	214	1457	0.98 (0.91-1.05)	1-2	3-4	3-4
T4	139	1022	**1.11 (1.02-1.20)**	3	1-4	1
T5	90	541	0.92 (0.82–1.03)	1	3-4	1-2
T6	147	930	0.93 (0.85-1.02)	2	3-4	1-2
T7	171	1287	1.08 (1.00-1.17)	3	4	2-4
T8	64	459	1.07 (0.94-1.22)	3-4	1	2-4
T9	361	2411	1.02 (0.96-1.08)	3	2-3	2-3
T10	102	624	**0.86 (0.78–0.96)**	3	2-3	4

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of CVD hospital admissions in the terminal node, adds up to 22 507
Bold: Significant p<0.05

6.4.3. PM_{10}, NO_2 AND O_3 (MIXTURE 3)

Figure 6.19 depicts the classification and regression tree of CVD hospital admissions for all ages and both sexes combined. Eight terminal nodes were found.

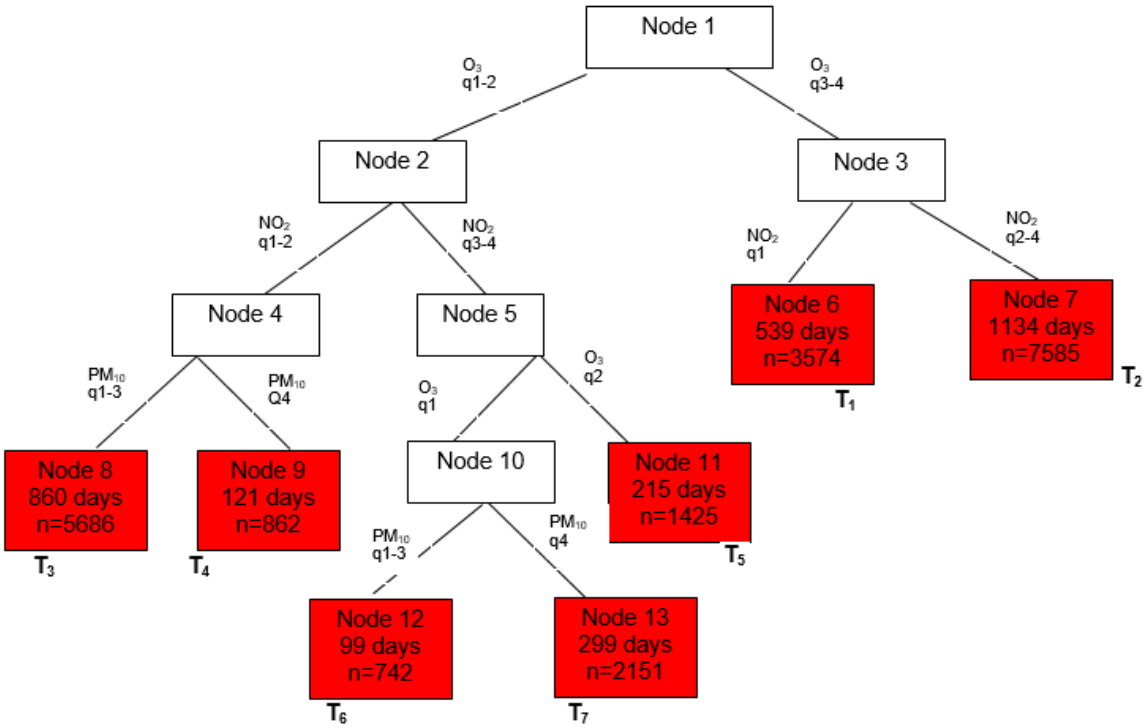


Figure 6.19: Classification and regression tree of cardiovascular disease hospital admissions modelled for all ages and both sexes combined ($PM_{2.5}$, NO_2 and O_3).
n= number of cardiovascular hospital admissions in the terminal node

Table 6.50 reports the joint effects of PM_{10}, NO_2, and O_3 on CVD hospital admissions, obtained in the adjusted regression models. Terminal nodes 2 and 8 were found to be significant with the highest RRs of 0.94 (95% CI: 0.90, 0.99) and 1.15 (95% CI: 1.04, 1.29) respectively. Terminal node 2 showed PM_{10} in all four quartiles, while NO_2 and O_3 were in the third and fourth quartiles. Terminal node 8 had PM_{10} in the second and third quartiles, NO_2 in the highest quartile, and O_3 in the lowest quartile. Exposure to terminal node mixture 2 suggest that CVD hospital admissions were lower in relation to the PM_{10}, NO_2, and O_3 exposure of the referent group. However, there was in increased risk of CVD hospital admissions when patients were exposed to terminal node mixture 8.

Table 6.50: Joint effects of PM_{10}, NO_2 and O_3 on cardiovascular disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**PM_{10}**	**NO_2**	**O_3**
Referent group	79	482	1.0	1	1	1
T1	235	1560	1.0 (0.5–2.3)	1	1-4	1-2
T2	539	3574	1.0 (0.7–1.5)	1-4	1-2	3-4
T3	1134	7585	**0.94 (0.90–0.99)**	1-4	3-4	3-4
T4	651	4300	1.0 (0.7-1.4)	2-3	2-3	1-2
T5	121	862	1.1 (0.8-1.4)	4	2-3	1-2
T6	203	1349	1.0 (0.7-1.6)	2-4	4	2
T7	85	644	1.08 (1.00-1.16)	2-3	4	1
T8	299	2151	**1.15 (1.04-1.29)**	4	4	1

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of CVD hospital admissions in the terminal node, adds up to 22 507
Bold: Significant p<0.05

6.4.4. $PM_{2.5}$, NO_2 AND O_3 (MIXTURE 4)

Figure 6.20 depicts the classification and regression tree of CVD hospital admissions for all ages and both sexes combined. Nine terminal nodes were identified, terminal nodes 2 and 7 were found to be significant.

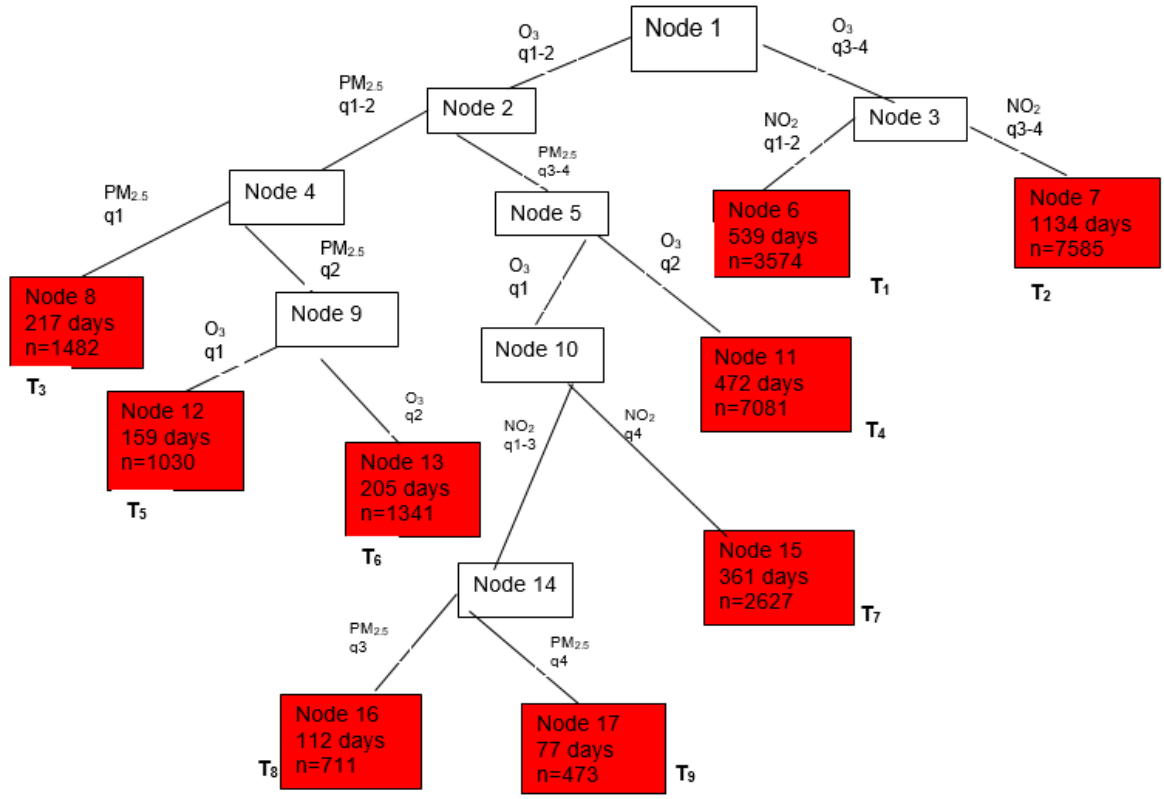


Figure 6.20: Classification and regression tree of cardiovascular disease hospital admissions modelled for all ages and both sexes combined ($PM_{2.5}$, NO_2 and O_3).
n= number of cardiovascular hospital admissions in the terminal node

Table 6.51 reports the joint effects of $PM_{2.5}$, NO_2, and O_3 on CVD hospital admissions, obtained in the adjusted regression models. Terminal 7 had $PM_{2.5}$ and NO_2 in the higher quartiles. Exposure to terminal node 7 significantly increased CVD hospital admissions for patients, with an RR of 1.13 (95% CI: 1.05, 1.21).

Table 6.51: Joint effects of $PM_{2.5}$, NO_2 and O_3 on cardiovascular disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**$PM_{2.5}$**	**NO_2**	**O_3**
Referent group	70	414	1.0	1	1	1
T1	539	3574	1.04 (0.98–1.10)	1-4	1	3-4
T2	1134	7585	0.94 (0.90–0.99)	1-4	2-4	3-4
T3	217	1482	0.95 (0.88–1.03)	1	1-4	1-2
T4	472	7081	1.02 (0.97–1.07)	3-4	1-4	2
T5	159	1030	0.93 (0.86–1.01)	2	1-4	1
T6	205	1341	0.98 (0.92–1.05)	2	1-4	2
T7	361	2627	**1.13 (1.05–1.21)**	3-4	4	1
T8	112	711	0.97 (0.90–1.04)	3	1-3	1
T9	77	473	0.97 (0.86–1.09)	4	1-3	1

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of CVD hospital admissions in the terminal node, adds up to 22 507
Bold: Significant p<0.05

6.4.5. PM_{10}, SO_2 AND O_3 (MIXTURE 5)

Figure 6.21 depicts the classification and regression tree of CVD hospital admissions for all ages and both sexes combined. Six terminal nodes were identified.

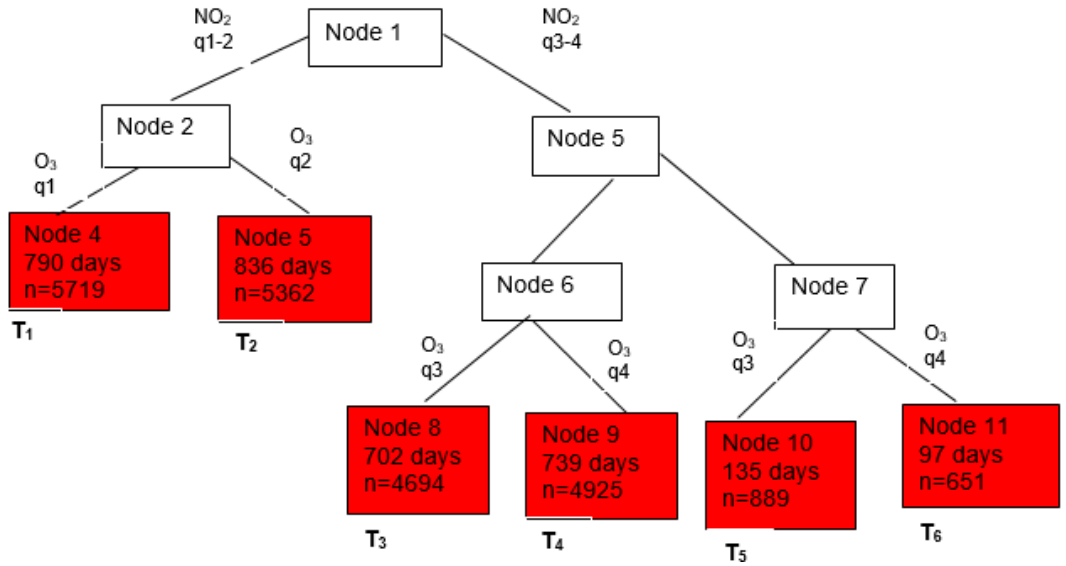


Figure 6.21: Classification and regression tree of cardiovascular disease hospital admissions modelled for all ages and both sexes combined (mixture PM_{10}, SO_2 and O_3). n= number of cardiovascular hospital admissions in the terminal node

Table 6.52 reports on the joint effects of PM_{10}, SO_2, and O_3 on CVD hospital admissions, obtained in the adjusted regression models. None of the mixtures in the terminal nodes significantly increased CVD hospital admissions for patients.

Table 6.52: Joint effects of PM_{10}, SO_2 and O_3 on cardiovascular disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**PM_{10}**	**SO_2**	**O_3**
Referent group	46	267	1.0	1	1	1
T1	790	5362	1.04 (0.99–1.11)	1-4	1-4	1
T2	837	5719	1.01 (0.97–1.06)	1-4	1-4	2
T3	702	4694	0.97 (0.93–1.01)	1-4	1-3	3
T4	739	4925	1.05 (0.99–1.11)	1-4	1-3	4
T5	135	889	0.92 (0.84–1.00)	1-4	4	3
T6	97	651	0.93 (0.83–1.03)	1-4	4	4

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of CVD hospital admissions in the terminal node, adds up to 22 507
Bold: Significant p<0.05

6.4.6. $PM_{2.5}$, NO_2 and O_3 (MIXTURE 6)

Figure 6.22 depicts the classification and regression tree of CVD hospital admissions for all ages and both sexes combined. Six terminal nodes were identified.

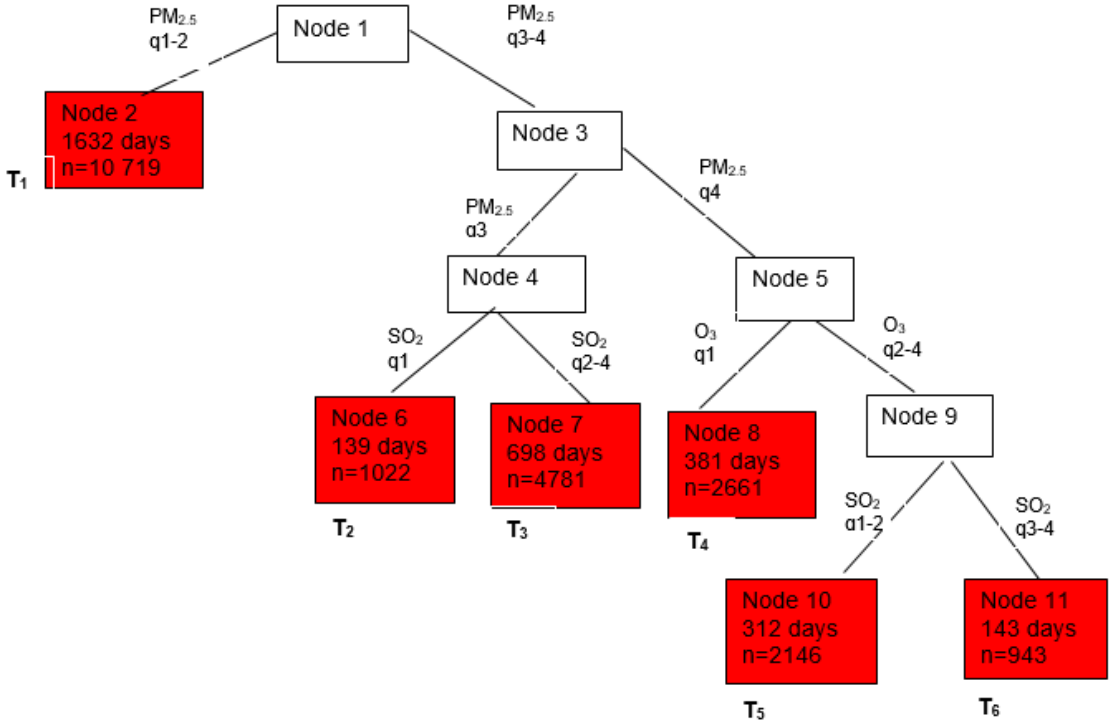


Figure 6.22: Classification and regression tree of cardiovascular disease hospital admissions modelled for all ages and both sexes combined (mixture $PM_{2.5}$, SO_2 and O_3). n= number of cardiovascular hospital admissions in the terminal node

Table 6.53 reports on the joint effects of $PM_{2.5}$, SO_2, and O_3 on CVD hospital admissions, obtained in the adjusted regression models. Terminal node mixture 2 showed to significantly increase CVD hospital admissions in patients with an RR of 1.11 (95% CI 1.02-1.20). $PM_{2.5}$ was in the third quartile, SO_2 in the lowest quartile, and O_3 in all four quartiles.

Table 6.53: Joint effects of $PM_{2.5}$, SO_2 and O_3 on cardiovascular disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models.

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**$PM_{2.5}$**	**SO_2**	**O_3**
Referent group	41	235	1.0	1	1	1
T1	1632	10719	0.97 (0.93–1.01)	1-2	1-4	1-4
T2	139	1022	**1.11 (1.02–1.20)**	3	1	1-4
T3	698	4781	1.02 (0.97–1.06)	3	2-4	1-4
T4	381	2661	1.06 (1.00–1.13)	4	1—4	1
T5	312	2146	0.95 (0.89–1.01)	4	1-2	2-4
T6	143	943	0.98 (0.90–1.08)	4	3-4	2-4

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of CVD hospital admissions in the terminal node, adds up to 22 507
Bold: Significant p<0.05

6.4.7. O_3, NO_2 AND SO_2 (MIXTURE 7)

Figure 6.23 depicts the classification and regression tree of CVD hospital admissions for all ages and both sexes combined. Ten terminal nodes were identified.

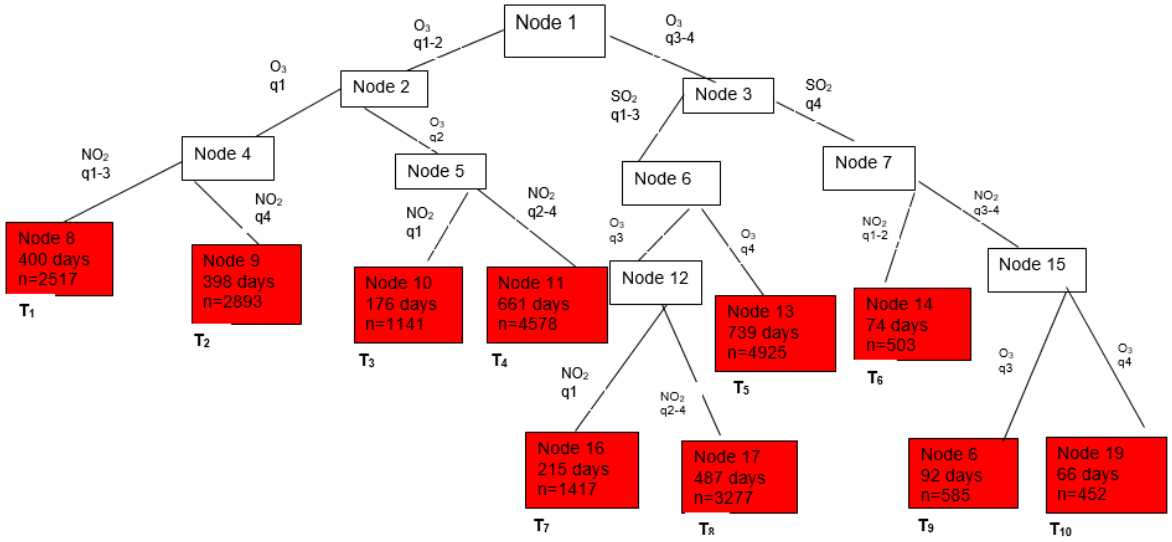


Figure 6.23: Classification and regression tree of cardiovascular disease hospital admissions modelled for all ages and both sexes combined (mixture O_3, NO_2 and SO_2). n= number of cardiovascular hospital admissions in the terminal node

Table 6.54 reports the joint effects of O_3, NO_2, and SO_2 on CVD hospital admissions, obtained in the adjusted regression models. Terminal nodes 2 and 9 were found to be significant with RRs of 1.14 (95% CI: 1.06, 1.22) and 0.87 (95% CI: 0.78, 0.97), respectively. Terminal node 2 mixture showed to significantly increase the CVD hospital admissions in patients. This mix showed O_3 and NO_2 in all four quartiles and SO_2 in the higher quartiles. Terminal node 9 showed O_3 in the third quartile, NO_2 in the third and fourth quartiles, and SO_2 was in the fourth quartile. Patients were at higher risk for CVD hospitalisation from exposure to the referent group than from exposure to mixture 9.

Table 6.54: Joint effects of O_3, NO_2 and SO_2 on cardiovascular disease hospital admissions for all ages and both sexes combined in VTAPA, South Africa during 2 January 2011 to 29 February 2020, obtained in the adjusted regression models.

				Type of days in the terminal node mixture (quartile number indicated)		
Terminal node	**N[a]**	**Number admissions[b]**	**Rate ratio (95% CI)**	**O_3**	**NO_2**	**SO_2**
Referent group	38	219	1.0	1	1	1
T1	400	2517	0.94 (0.89–1.01)	1	1-3	1-4
T2	398	2893	**1.14 (1.06–1.22)**	1	4	1-4
T3	176	1141	0.95 (0.88–1.03)	2	1	1-4
T4	661	4578	1.03 (0.98–1.08)	2	2-4	1-4
T5	739	4925	1.05 (0.99–1.11)	4	1-4	1-3
T6	74	503	0.97 (0.86–1.09)	3-4	1-2	4
T7	215	1417	1.00 (0.93–1.08)	3	1	1-3
T8	487	3277	0.96 (0.91–1.01)	4	2-4	1-3
T9	92	585	**0.87 (0.78–0.97)**	3	3-4	4
T10	66	452	0.95 (0.84–1.08)	4	3-4	4

[a] Days in the terminal node, adds up to 3346 days.
[b] Number of CVD hospital admissions in the terminal node, adds up to 22 507
Bold: Significant p<0.05

6.5. DISCUSSION

The aim of this part of the project was to determine the joint effects of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} on respiratory disease and cardiovascular disease hospital admissions in Vereeniging and Vanderbijlpark, Gauteng, using CART analysis statistical modelling.

6.5.1. DESCRIPTIVE STATISTICS OF AIR POLLUTION

The PM_{10}, $PM_{2.5}$, NO_2, and SO_2 levels showed a pattern of higher concentration levels in colder months (May to August). This concurs with other research that has found that concentration levels of pollutants such as particulate matter and black carbon increase in colder months.[1-4] Temperature inversion can be related to this trend in air pollutants; in colder months a layer is developed that traps air pollutants, whereas air pollutants are released easier from the atmosphere during warmer months.[5] Temperature inversion is a meteorological phenomenon where temperature profiles in the atmosphere deviate from the norm, causing air pollutants to rise and disperse in warmer conditions, but causes a 'trap' in cooler temperatures that keeps air pollutants closer to the ground.[6] However, O_3 showed a pattern opposite to that of the other four air pollutants, where there were higher concentration levels in warmer months (September to April). This showed to be a constant observation in other studies

conducted in the VTAPA and similar areas in South Africa.[1-3,7] During warmer months there is stronger radiation and the higher temperatures provide favourable conditions for photochemical reactions and O_3 production.[8-9]

SO_2 and O_3 did not significantly vary throughout the week, although PM_{10}, $PM_{2.5}$, and NO_2 showed to have varied in concentration level during the week. The highest means were seen during Monday to Friday, with PM_{10} and NO_2 having the highest concentration levels on Wednesday and Tuesday, respectively. The lowest concentration levels, however, were recorded on weekends. Since the normal working week is from Monday to Friday, the increase in industrial activities during this time is a likely cause for the increase in air pollution.[10-12]

The concentration levels may have not been fully representative, with occasional peaks that may have occurred outside what is possibly the norm. The main reason for this could be the imputations explained in Chapter 5. However the results do reflect relatively high air pollution concentration levels over the ten-year study period. In the VTAPA, possible sources of $PM_{2.5}$ include industry, coal burning, wood and biomass burning, secondary aerosols, and vehicles.[4] Additionally the concentration levels were higher than WHO guidelines in comparison to the NAAQS. Although the South African standards seem lenient compared to the WHO guidelines, it has been argued that the WHO guidelines are unattainable in places like South Africa due to multiple strong and varied natural air pollution sources such as dust and biomass burning.[13]

Seven air pollution mixtures were explored in this study. PM_{10} and $PM_{2.5}$ were not placed in the same air pollution mixtures, because they were too strongly correlated ($PM_{2.5}$ is a subset of PM_{10}). Although SO_2, NO_2, and O_3 have an inverse relationship, not many studies have explored the combination of the three pollutants and their association with RD and CVD hospital admissions. Hence, the inclusion of the different air pollution mixtures allowed an exploration of the interactive nature of air pollution;[14] to observe the health effects of only one pollutant or one mixture would limit possible estimates. The mixture of air pollutants can worsen health effects. PM and O_3 have been found to have a synergistic effect and can increase risk or acute respiratory inflammation and other respiratory diseases.[15] The mixture of air pollutants can also have additive and potentiation qualities, this can double the effect of one air pollutant or make one pollutant more toxic.[16] The joint effects of air pollutant mixtures, like the

combination of PM_{10}, NO_2, and SO_2, have shown to have a detrimental influence on hospital admission.[17-18]

In addition to the correlation among the air pollutants, external factors such as temperature, specifically Tapp, can influence the number and frequency of hospital admissions.[19] Air pollutants such as PM in the presence of high temperatures can intensify or increase the likelihood of conditions like migraines.[20] Thus, the exploration of potential air pollution mixtures and possible joint effects of multiple air pollutants on RD and CVD hospital admissions are critical. Research that investigates the health effects of exposure to ambient air pollutants in Africa is still fairly new.[21-23] Thus, the multiple air pollution combinations create a basis to further investigation.

6.5.2. THE ASSOCIATION OF AIR POLLUTION MIXTURE AND RESPIRATORY DISEASE HOSPITAL ADMISSIONS

Evidence supports that short-term effects of ambient air pollution exposure can increase respiratory hospital admissions.[24-26] Such exposure is often associated with multiple respiratory conditions and the aggravation of existing respiratory disease, such as chronic obstructive pulmonary disease (COPD).[27-28]

The results of the study showed that patients exposed to the different mixtures of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} could experience an increased the risk RD hospitalisation by at least 1.0 compared to those exposed to air pollutants on the days when pollutants were in the lowest quartiles. When $PM_{2.5}$ or PM_{10} were at low or high concentrations, and at higher concentrations of NO_2, patients were at an increased risk of RD hospitalisation. A study found that short-term exposure to air pollutants such as $PM_{2.5}$, NO_2 and SO_2 is associated with an increased risk of developing respiratory disease by factors of 2.5% (95% CI: 1.6 - 3.4%), 4.2% (95% CI: 2.5 - 6.0%), and 2.1% (95% CI: 0.7 - 3.5%).[29] Findings from a Cape Town study that investigated PM_{10}, NO_2, and SO_2 showed and increase risk of RD hospital admission at an RR 1.3 (95% CI: 1.1 - 1.5) and 1.2 (95% CI: 1.1 - 1.) when NO_2 was in all four quartiles.[30] The RRs for air pollutants investigated as mixtures seem to be lower than when the air pollutants were individually investigated. A previous study conducted in the VTAPA investigated the effects of Tapp and air pollutants on RD hospital admissions. The latter study found that in the presence of warm-medium Tapp, exposure to NO_2, SO_2, PM_{10}, and $PM_{2.5}$ was 7.8% (95% CI 4.9 - 10.8%), 6.5% (95% CI 3.4- 9.6%), 3.3% (95% CI 1.7- 5.0%), and 4.2% (95% CI 1.5 - 6.9%), respectively.[19] O_3, when in the presence of high Tapp, significantly

increased RD hospital admission by 5.2% (95% CI 1.6 - 9.0%).[19] The risk of RD hospital admission from single pollutant and multiple pollutant exposure can vary.

Combined exposure of PM_{10} and NO_2, and $PM_{2.5}$ and NO_2, have led to a continued decline in of respiratory health, which leaves individuals susceptible to respiratory disease and complications[31-32] Combinations such as PM_{10}, NO_2, and SO_2 have shown to increase hospital admissions in children. Exposure to such combinations can intensify respiratory illnesses, such as wheezing and asthma, that has the potential to increase absenteeism in schools.[24,33-34]

Short-term exposures to $PM_{2.5}$, NO_2, and O_3 may increase the risk of asthma and mortality in the elderly.[35-36] PM and O_3 have shown to have a synergetic effect that seems to lead to the deterioration of respiratory function.[37] Results from this study showed an increased risk of RD hospital admission when exposed to $PM_{2.5}$, NO_2, and O_3, however, there was a lower risk of RD hospital admissions when patients were exposed to PM_{10}, NO_2, and O_3.

A study that investigated exposure to air pollution and emergency department visits for respiratory diseases also showed a significant RR under 1.[38] This could suggest that a combination of PM_{10}, NO_2, and O_3, even at low concentration levels, can increase the risk of RD hospital admission. The elderly population is more susceptible, due to the increased airway hyper-responsiveness when exposed to air pollution.[39] Studies have shown that exposure to ambient air pollutants can affect subgroups other than children and the elderly. Some studies suggests that females may be more susceptible to RD hospital admissions because of their particular physiological constitution.[40-41] However, the latter studies used different lags and compared different population groups.

The uncommon combination of O_3, NO_2, and SO_2, was also explored where two terminal node mixtures were found significant with RRs of 1.04 (95% CI: 1.01, 1.08) and 0.93 (95% CI: 0.89, 0.97). In one terminal node mixture, all three pollutants were in the lower concentration levels. In the other mixture O_3 and SO_2 were in both low and high concentrations, while NO_2 showed higher concentration levels. In spite of the negative inverse relationships SO_2 and NO_2 have with O_3, this mixture showed to increase risk of RD hospital admission. A Chinese study showed short-term exposure to O_3 is associated with lengths of hospital stays in children, however, inconsistencies were found in other pollutants.[42] Another study showed that short-term exposure to O_3,

NO_2, and SO_2 can cause an increase in asthma emergency department visits and hospital admissions, but there was no significant difference in the admission of different subgroups.[43] Continuous research on multiple air pollutant mixtures that include O_3 are needed to understand the possible effects on RD hospital admissions.

6.5.3. THE ASSOCIATION OF AIR POLLUTION MIXTURE AND CARDIOVASCULAR DISEASE HOSPITAL ADMISSIONS

Ambient air pollution exposure has shown to increase in CVD mortality and hospital admissions.[44] Studies have also showed that exposure to low air pollutant concentration levels have a negative impact on CVD hospitalisation and mortality.[45-46] There have been strong associations between cardiac mortality and hospitalisation, to exposure of particulate matter and gaseous pollutants, such as SO_2 and nitrogen oxides.[47-50]

Although mixture 1 did not significantly increase CVD hospital admissions in this project, a Cape Town study found that individuals who were exposed to an air pollutant mixture of PM_{10}, NO_2, and SO_2 were at 1.2 times higher risk of being admitted for CVD.[30]

When $PM_{2.5}$ was in higher concentration levels, NO_2 varied from low to high concentrations and SO_2 was in the lowest concentration levels; these levels increased the risk of CVD hospital admission by 1.11. A Cape Town study investigating air pollution on CVD mortality found that long-term exposure to NO_2 and SO_2 increased the risk of CVD mortality by 3.4% and 2.6%, respectively.[51] However, the study only investigated the effects of single pollutants. Similar to the interactions with RD hospital admissions, mixtures with higher NO_2 concentration levels also showed to increase CVD hospital admissions. Increased PM_{10} and NO_2, and $PM_{2.5}$ and NO_2, have been associated with composite CVD, acute coronary events, heart failure, atrial fibrillation, as well as all-cause CVD mortality.[52-55] Other studies have shown that increased NO_2 also cause left ventricular diastolic dysfunction.[56]

NO_2 levels are difficult to quantify in isolation and it is not always the main pollutant focused on in research.[57] PM_{10} and SO_2 are usually used as NO_2 indicators of health effects.[58] The exposure to high levels of NO_2 have been found to affect the endocrine system in females and may be responsible for fertility and cardiovascular problems.[59-60] Short-term exposure to $PM_{2.5}$ and NO_2 can influence the onset of disease and

increase hospital admissions for arrhythmia.[61] $PM_{2.5}$, SO_2, NO_2, and CO have been associated with an increased risk of hospitalisation and death caused by congestive heart failure.[62]

There has been a close temporal association between exposure to gaseous and particulate air pollutants and hospitalisation due to stroke.[63-64] Studies have found this relationship for short-term exposure to SO_2, NO_2, and CO, as well as O_3.[65-67] Subpopulation groups including children, women and the elderly have shown higher susceptibility of CVD hospital admission from air pollution exposure. Children may be more likely to suffer CVD due to air pollution, because of the underdevelopment of their cardiovascular system compared to adults.[68] Short-term exposure to ambient air pollution has also been associated with increased cardiovascular-related hospital admissions among women and the elderly.[69-70] Furthermore, elderly individuals are most susceptible to have pre-existing chronic conditions.[71] Some African studies have shown such associations but there is still much to be done to explore these associations.[72]

This project provides epidemiological evidence that exposure to different air pollution mixtures can lead to increased risks of RD and CVD hospital admission. Although the VTAPA has been designated a national priority area and has specific air monitoring management plans,[73-74] more needs to be done in these areas. For instance, communities in low-income settlement areas have inconsistent refuse collection by their local municipalities and depend on burning their waste, which jeopardises their health and safety.[4] Therefore, basic services like refuse collection can improve their air quality. Revising and enforcing stricter emission policies of industry can improve the quality of air in this area. Investment in alternative energy sources in South Africa can also lead to improved air quality, which will have a direct impact on health.[75-76] Lastly, consistent maintenance and upgraded air quality monitoring systems can increase more epidemiological research.

6.6. STRENGTHS AND LIMITATIONS

This study has a number of strengths that can be mentioned. Firstly, this study took into consideration the different air pollution mixtures and their possible joint effects on RD and CVD hospital admissions. This helps increase knowledge on the joint effects of multi-pollutant exposures on RD and CVD hospital admissions. It also investigates

seven air pollutant mixtures, as opposed to one pollutant. Only two other studies have investigated PM_{10}, NO_2, and SO_2 mixtures in Cape Town, South Africa.[30,77]

Secondly, the study used classification and regression tree statistical modelling on South African data. The modified CART analysis by Gass et al, looked at both multi-pollutant effects as well as day type and factor correlations.[78] This is advantageous in that it can break down, as far as possible, more specific day types that could be harmful to increasing RD and CVD hospital admissions in South Africa.

Lastly, this study used imputation to complete the missing air pollution data. Missing data are a continual limitation in environmental studies, as discussed in detail in Chapter 5. This study used a completed air pollution dataset.

The study did have some limitations, one of which is that, despite using a completed air pollution dataset, the imputed values are only estimates and could affect joint effect outcomes. Another limitation to the study is that it used private hospital admission data, which has been a limitation of similar studies.[17,19,79-80] This is because private hospital admission data are available in electronic format, but not for public hospital admissions. However, only 16% of the South African population use private healthcare.[81] Thus, the findings cannot be ascribed to the general population.

Another limitation is the possible issue of measurement error. This is from the assumption that the measured air pollution and meteorological data were the same across the few air monitoring sites in the VTAPA. Lastly, the study did not investigate different subpopulation groups, i.e. sex and different age groups. Additionally, it only investigated relationships at a two-day cumulative lag. These limitations provide baselines for further investigations that study different population groups and time lags.

6.7. CONCLUSION

SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} mixtures showed to be associated with RD and CVD hospital admissions. The mixtures showed that the higher concentrations of NO_2 in combination with varying concentrations of SO_2, O_3, $PM_{2.5}$, and PM_{10} can lead to increased risk of both RD and CVD hospitalisation. The findings of the study in indicate there is still more to investigate concerning joint effects of different air pollutant mixtures. This also adds to epidemiological evidence that can help policy makers introduce stricter policies to improve the air quality of high priority areas in South Africa, such as VTAPA.

6.8. REFERENCES

1. Feig G, Ncipha X, Vertue B, Naidoo S, Mabaso D, Ngcukana N, et al. Analysis of a period of elevated ozone concentration reported over the Vaal Triangle on 2 June 2013. Clean Air Journal= Tydskrif vir Skoon Lug. 2014; 24(1):10-6.

2. Feig GT, Vertue B, Naidoo S, Ncgukana N, Mabaso D. Measurement of atmospheric black carbon in the Vaal Triangle and Highveld Priority Areas. Clean Air Journal. 2015; 25(1):46-.

3. Govender K, Sivakumar V. A decadal analysis of particulate matter (pm2. 5) and surface ozone (O3) over Vaal Priority Area, South Africa. Clean Air Journal. 2019; 29(2)

4. Muyemeki L, Burger R, Piketh SJ, Beukes JP, Van Zyl PG. Source apportionment of ambient PM10-25 and PM2.5 for the vaal triangle, South Africa. South African Journal of Science. 2021; 117(5-6):1-11.

5. Trinh TT, Trinh TT, Le TT, Nguyen TD, Tu BM. Temperature inversion and air pollution relationship, and its effects on human health in Hanoi city, Vietnam. Environmental Geochemistry and Health. 2019; 41(2):929-37.

6. Sager L. Estimating the effect of air pollution on road safety using atmospheric temperature inversions. Journal of Environmental Economics and Management. 2019; 98:102250. doi:https://doi.org/10.1016/j.jeem.2019.102250.

7. Hersey SP, Garland RM, Crosbie E, Shingler T, Sorooshian A, Piketh S, et al. An overview of regional and local characteristics of aerosols in South Africa using satellite, ground, and modeling data. Atmospheric Chemistry and Physics. 2015; 8:4259-78

8. Chen K-S, Ho YT, Lai CH, Tsai YA, Chen S-J. Trends in concentration of ground-level ozone and meteorological conditions during high ozone episodes in the Kao-Ping Airshed, Taiwan. Journal of the Air & Waste Management Association. 2004; 54(1):36-48.

9. Dai H, Zhu J, Liao H, Li J, Liang M, Yang Y, et al. Co-occurrence of ozone and PM2.5 pollution in the Yangtze River Delta over 2013-2019: Spatiotemporal distribution and meteorological conditions. Atmospheric Research. 2021; 249 doi:10.1016/j.atmosres.2020.105363.

10. Aung W, Noguchi M, Pan-Nu Yi E, Thant Z, Uchiyama S, Win-Shwe T, et al. Preliminary assessment of outdoor and indoor air quality in Yangon City, Myanmar. Atmospheric pollution research. 2019; 10(3):722-30. doi:10.1016/j.apr.2018.11.011.

11. Fayiga AO, Ipinmoroti MO, Chirenje T. Environmental pollution in Africa. Environment, Development and Sustainability. 2018; 20(1):41-73. doi:10.1007/s10668-016-9894-4.

12. World Health Organization. WHO air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide - global update 2005 - summary of risk assessment. Geneva; 2006.

13. Garland RM, Wernecke B, Feig G, Langerman K. The new WHO global air quality guidelines: What do they mean for South Africa? Clean Air Journal. 2021; 31(2):1-2.

14. Stafoggia M, Breitner S, Hampel R, Basagaña X. Statistical approaches to address multi-pollutant mixtures and multiple exposures: The state of the science. Current Environmental Health Reports. 2017; 4(4):481-90. doi:10.1007/s40572-017-0162-z.
15. Valavanidis A, Loridas S, Vlahogianni T, Fiotakis K. Influence of ozone on traffic-related particulate matter on the generation of hydroxyl radicals through a heterogeneous synergistic effect. Journal of Hazardous Materials. 2009; 162(2):886-92. doi:10.1016/j.jhazmat.2008.05.124.

16. Seinfeld JH, Pandis SN. Atmospheric chemistry and physics : From air pollution to climate change. Third edition. ed. Hoboken, New Jersey: John Wiley & Sons; 2016.

17. Lokotola CL, Wright CY, Wichmann J. Temperature as a modifier of the effects of air pollution on cardiovascular disease hospital admissions in Cape Town, South Africa. Environmental Science and Pollution Research. 2020; 27:16677-85. doi:10.1007/s11356-020-07938-7.

18. Thabethe NDL, Voyi K, Wichmann J. Association between ambient air pollution and cause-specific mortality in Cape Town, Durban, and Johannesburg, South Africa: Any susceptible groups? Environmental Science and Pollution Research. 2021; 28:42868-76 doi:10.1007/s11356-021-13778-w.

19. Mwase N, Olutola B, Wichmann J. Temperature modifies the association between air pollution and respiratory disease hospital admissions in an industrial area of South Africa: The Vaal Triangle Air Pollution Priority Area. Clean Air Journal. 2022; 32(2). doi:10.17159.caj.2022.32.2.14588.

20. Lee H, Myung W, Cheong HK, Yi SM, Hong YC, Cho SI, et al. Ambient air pollution exposure and risk of migraine: Synergistic effect with high temperature. Environment International. 2018; 121(Pt 1):383-91. doi:10.1016/j.envint.2018.09.022.

21. Coker E, Kizito S. A narrative review on the human health effects of ambient air pollution in Sub-Saharan Africa: An urgent need for health effects studies. International Journal Of Environmental Research and Public Health. 2018; 15(3):427.

22. Mustapha BA, Blangiardo M, Briggs DJ, Hansell AL. Traffic air pollution and other risk factors for respiratory illness in schoolchildren in the Niger-Delta region of Nigeria. Environmental Health Perspectives. 2011; 119(10):1478-82.

23. Ana G, Odeshi T, Sridhar M, Ige M. Outdoor respirable particulate matter and the lung function status of residents of selected communities in Ibadan, Nigeria. Perspectives in Public Health. 2014; 134(3):169-75.

24. Capraz O, Deniz A, Dogan N. Effects of air pollution on respiratory hospital admissions in Istanbul, Turkey, 2013 to 2015. Chemosphere. 2017; 181:544-50. doi:10.1016/j.chemosphere.2017.04.105.

25. Chen K, Glonek G, Hansen A, Williams S, Tuke J, Salter A, et al. The effects of air pollution on asthma hospital admissions in Adelaide, South Australia, 2003-2013: Time-series and case-crossover analyses. Clinical & Experimental Allergy. 2016; 46(11):1416-30. doi:10.1111/cea.12795.

26. Nhung NTT, Schindler C, Dien TM, Probst-Hensch N, Perez L, Künzli N. Acute effects of ambient air pollution on lower respiratory infections in Hanoi children: An eight-year time series study. Environment International. 2018; 110:139-48. doi:10.1016/j.envint.2017.10.024.

27. Deng Q, Lu C, Norbäck D, Bornehag CG, Zhang Y, Liu W, et al. Early life exposure to ambient air pollution and childhood asthma in China. Environmental Research. 2015; 143(Pt A):83-92. doi:10.1016/j.envres.2015.09.032.

28. Ding L, Zhu D, Peng D, Zhao Y. Air pollution and asthma attacks in children: A case-crossover analysis in the city of Chongqing, China. Environmental Pollution. 2017; 220(Part A):348-53. doi:10.1016/j.envpol.2016.09.070.

29. DeVries R, Kriebel D, Sama S. Outdoor air pollution and COPD-related emergency Department visits, hospital admissions, and mortality: A meta-analysis. International Journal of Chronic Obstructive Pulmonary Disease. 2017; 14(1):113-21. doi:10.1080/15412555.2016.1216956.

30. Nunu MB. The association of joint effects of air pollution on cardiovascular and respiratory disease hospital admissions in Cape Town: 2011- 2016: University of Pretoria; 2018.

31. Carugno M, Consonni D, Randi G, Catelan D, Grisotto L, Bertazzi PA, et al. Air pollution exposure, cause-specific deaths and hospitalizations in a highly polluted Italian region. Environmental Research. 2016; 147:415-24. doi:10.1016/j.envres.2016.03.003.

32. Pannullo F, Lee D, Neal L, Dalvi M, Agnew P, O'Connor FM, et al. Quantifying the impact of current and future concentrations of air pollutants on respiratory disease risk in England. Environmental Health. 2017; 16(1):29. doi:10.1186/s12940-017-0237-1.

33. Albers PN, Mathee A, Voyi KVV, Wright CY. Household fuel use and child respiratory ill health in two towns in Mpumalanga, South Africa. SAMJ: South African Medical Journal. 2015; 105(7):573-7. doi:10.7196/SAMJNEW.7934.

34. Guarnieri M, Balmes JR. Outdoor air pollution and asthma. Lancet. 2014; 383(9928):1581-92. doi:10.1016/s0140-6736(14)60617-6.

35. Liu Y, Pan J, Zhang H, Shi C, Li G, Peng Z, et al. Short-term exposure to ambient air pollution and asthma mortality. American Journal of Respiratory and Critical Care Medicine. 2019; 200(1):24-32. doi:10.1164/rccm.201810-1823OC.

36. Zhou M, He G, Liu Y, Yin P, Li Y, Kan H, et al. The associations between ambient air pollution and adult respiratory mortality in 32 major Chinese cities, 2006-2010. Environmental Research. 2015; 137:278-86. doi:10.1016/j.envres.2014.12.016.

37. Zoran M, Dida MR, Savastru R, Savastru D, Dida A, Ionescu O. Ground level ozone (O3) associated with radon (222rn) and particulate matter (PM) concentrations in Bucharest metropolitan area and adverse health effects. Journal of Radioanalytical and Nuclear Chemistry. 2014; 300(2):729-46. doi:10.1007/s10967-014-3041-1.

38. Szyszkowicz M, Kousha T, Castner J, Dales R. Air pollution and emergency department visits for respiratory diseases: A multi-city case crossover study. Environmental Research. 2018; 163:263-9. doi:10.1016/j.envres.2018.01.043.

39. Boezen HM, Vonk JM, van der Zee SC, Gerritsen J, Hoek G, Brunekreef B, et al. Susceptibility to air pollution in elderly males and females. European Respiratory Journal. 2005; 25(6):1018-24.

40. Clougherty JE. A growing role for gender analysis in air pollution epidemiology. Ciencia & Saude Coletiva. 2011; 16:2221-38.

41. Cabello N, Mishra V, Sinha U, DiAngelo SL, Chroneos ZC, Ekpa NA, et al. Sex differences in the expression of lung inflammatory mediators in response to ozone. American Journal of Physiology-Lung Cellular and Molecular Physiology. 2015; 309(10):L1150-L63.

42. Nhung NTT, Schindler C, Dien TM, Probst-Hensch N, Künzli N. Association of ambient air pollution with lengths of hospital stay for Hanoi children with acute lower-respiratory infection, 2007-2016. Environmental Pollution. 2019; 247:752-62. doi:10.1016/j.envpol.2019.01.115.

43. Zheng XY, Orellano P, Lin HL, Jiang M, Guan WJ. Short-term exposure to ozone, nitrogen dioxide, and sulphur dioxide and emergency department visits and hospital admissions due to asthma: A systematic review and meta-analysis. Environment International. 2021; 150:106435. doi:10.1016/j.envint.2021.106435.

44. World Health Organization Europe. Review of evidence on health aspects of air pollution – REVIHAAP project: Final technical report. Copenhagen; 2013.

45. Cesaroni G, Forastiere F, Stafoggia M, Andersen ZJ, Badaloni C, Beelen R, et al. Long term exposure to ambient air pollution and incidence of acute coronary events: Prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE project. BMJ. 2014; 348:f7412. doi:10.1136/bmj.f7412.

46. Stafoggia M, Cesaroni G, Peters A, Andersen ZJ, Badaloni C, Beelen R, et al. Long-term exposure to ambient air pollution and incidence of cerebrovascular events: Results from 11 European cohorts within the ESCAPE project. Environmental Health Perspectives. 2014; 122(9):919-25. doi:10.1289/ehp.1307301.

47. Abdolahnejad A, Jafari N, Mohammadi A, Miri M, Hajizadeh Y, Nikoonahad A. Cardiovascular, respiratory, and total mortality ascribed to PM10 and PM2.5 exposure in Isfahan, Iran. Journal of Education and Health Promotion. 2017; 6:109. doi:10.4103/jehp.jehp_166_16.

48. Kaufman JD, Adar SD, Barr RG, Budoff M, Burke GL, Curl CL, et al. Association between air pollution and coronary artery calcification within six metropolitan areas in the USA (the multi-ethnic study of atherosclerosis and air pollution): A longitudinal cohort study. Lancet. 2016; 388(10045):696-704. doi:10.1016/s0140-6736(16)00378-0.

49. Robertson S, Miller MR. Ambient air pollution and thrombosis. Particle and Fibre Toxicology. 2018; 15(1):1-16. doi:10.1186/s12989-017-0237-x.

50. Zhang Z, Guo C, Lau AKH, Chan TC, Chuang YC, Lin C, et al. Long-term exposure to fine particulate matter, blood pressure, and incident hypertension in Taiwanese adults. Environmental Health Perspectives. 2018; 126(1):017008. doi:10.1289/ehp2466.

51. Wichmann J, Voyi K. Ambient air pollution exposure and respiratory, cardiovascular and cerebrovascular mortality in Cape Town, South Africa: 2001-2006. International Journal of Environmental Research and Public Health. 2012; 9(11):3978-4016. doi:10.3390/ijerph9113978.

52. Cai Y, Hodgson S, Blangiardo M, Gulliver J, Morley D, Fecht D, et al. Road traffic noise, air pollution and incident cardiovascular disease: A joint analysis of the HUNT, EPIC-Oxford and UK Biobank cohorts. Environment International. 2018; 114:191-201. doi:10.1016/j.envint.2018.02.048.

53. Gandini M, Scarinzi C, Bande S, Berti G, Carnà P, Ciancarella L, et al. Long term effect of air pollution on incident hospital admissions: Results from the Italian longitudinal study within life med hiss project. Environment International. 2018; 121(Part 2):1087-97. doi:10.1016/j.envint.2018.10.020.

54. Kim H, Kim J, Kim S, Kang S-H, Kim H-J, Kim H, et al. Cardiovascular effects of long-term exposure to air pollution: A population-based study with 900 845 person-years of follow-up. Journal of the American Heart Association. 2017; 6(11) doi:10.1161/JAHA.117.007170.

55. Wolf K, Stafoggia M, Cesaroni G, Andersen ZJ, Beelen R, Galassi C, et al. Long-term exposure to particulate matter constituents and the incidence of coronary events in 11 European cohorts. Epidemiology. 2015; 26(4):565-74. doi:10.1097/EDE.0000000000000300.

56. Zheng C, Tang H, Wang X, Chen Z, Zhang L, Kang Y, et al. Left ventricular diastolic dysfunction and cardiovascular disease in different ambient air pollution conditions: A prospective cohort study. Science of the Total Environment. 2022; 831 doi:10.1016/j.scitotenv.2022.154872.

57. Saucy A, Röösli M, Künzli N, Tsai M-Y, Sieber C, Olaniyan T, et al. Land use regression modelling of outdoor NO2 and PM2.5 concentrations in three low income areas in the Western Cape Province, South Africa. International Journal of Environmental Research and Public Health. 2018; 15(7):1452. https://doi.org/10.3390/ijerph15071452.

58. Khaniabadi YO, Goudarzi G, Daryanoosh SM, Borgini A, Tittarelli A, De Marco A. Exposure to PM10, NO2, and O3 and impacts on human health. Environmental Science and Pollution Research. 2017; 24(3):2781-9. https://doi.org/10.1007/s11356-016-8038-6.

59. Conforti A, Mascia M, Cioffi G, De Angelis C, Coppola G, De Rosa P, et al. Air pollution and female fertility: A systematic review of literature. Reproductive Biology and Endocrinology. 2018; 16(1):117. https://doi.org/10.1186/s12958-018-0433-z.

60. Ghio AJ, Kim C, Devlin RB. Concentrated ambient air particles induce mild pulmonary inflammation in healthy human volunteers. American Journal of Respiratory and Critical Care Medicine. 2000; 162(3):981-8.

61. Mordukhovich I, Coull B, Kloog I, Koutrakis P, Vokonas P, Schwartz J. Exposure to sub-chronic and long-term particulate air pollution and heart rate variability in an elderly cohort: The normative aging study. Environmental Health 2015; 14:87.

62. Shah AS, Langrish JP, Nair H, McAllister DA, Hunter AL, Donaldson K, Newby DE, Mills NL. Global association of air pollution and heart failure: A systematic review and meta-analysis. Lancet. 2013; 382:1039-48.

63. Shah AS, Lee KK, McAllister DA, Hunter A, Nair H, Whiteley W, et al. Short term exposure to air pollution and stroke: Systematic review and meta-analysis. BMJ (Clinical research ed.). 2015; 350:h1295. doi:10.1136/bmj.h1295.

64. Vidale S, Arnaboldi M, Bosio V, Corrado G, Guidotti M, Sterzi R, et al. Short-term air pollution exposure and cardiovascular events: A 10-year study in the urban area of Como, Italy. International Journal of Cardiology. 2017; 248:389-93. doi:10.1016/j.ijcard.2017.06.037

65. Henrotin J, Zeller M, Lorgis L, Cottin Y, Giroud M, Bejot Y. Evidence of the role of short-term exposure to ozone on ischaemic cerebral and cardiac events: The Dijon Vascular Project(DIVA). Heart. 2010; 96:1990-6.

66. Ljungman PL, Mittleman MA. Ambient air pollution and stroke. Stroke. 2014; 45:3734-41.

67. Shin H, Burnett R, Cohen A, Hubbell BJ. Outdoor fine particles and nonfatal strokes: Systematic review and meta-analysis. Epidemiology. 2014; 25:835-42.

68. Gouveia N, Junger WL, Romieu I, Cifuentes LA, de Leon AP, Vera J, et al. Effects of air pollution on infant and children respiratory mortality in four large Latin-American cities. Environmental Pollution. 2018; 232:385-91.

69. Dastoorpoor M, Sekhavatpour Z, Masoumi K, Mohammadi MJ, Aghababaeian H, Khanjani N, et al. Air pollution and hospital admissions for cardiovascular diseases in Ahvaz, Iran. Science of the Total Environment. 2019; 652:1318-30. doi:10.1016/j.scitotenv.2018.10.285.

70. Yitshak-Sade M, Nethery R, Abu Awad Y, Mealli F, Dominici F, Kloog I, et al. Lowering air pollution levels in massachusetts may prevent cardiovascular hospital admissions. Journal of the American College of Cardiology. 2020; 75(20):2642-4. doi:10.1016/j.jacc.2020.03.056.

71. Simoni M, Baldacci S, Maio S, Cerrai S, Sarno G, Viegi G. Adverse effects of outdoor pollution in the elderly. Journal of Thoracic Disease. 2015; 7(1):34.

72. Health Effects Institute. The state of air quality and health impacts in Africa: A report from the state of global air initiative. 2022. Available from: https://www.stateofglobalair.org/sites/default/files/documents/2022-10/soga-africa-report.pdf.

73. Deparment of Environmental Affairs And Tourism. Executive summary of the Vaal Triangle Airshed Proirity Area air quality managenet plan. 2008.

74. Department of Environment Forestry and Fisheries. Draft second generation air quality management plan for Vaal Triangle Airshed Priority Area. Pretoria,: Government Gazette. 2020.

75. Gielen D, Boshell F, Saygin D, Bazilian MD, Wagner N, Gorini R. The role of renewable energy in the global energy transformation. Energy Strategy Reviews. 2019; 24:38-50.

76. Panwar NL, Kaushik SC, Kothari S. Role of renewable energy sources in environmental protection: A review. Renewable and Sustainable Energy Reviews. 2011; 15(3):1513-24.

77. Alfeus Anna. Air pollution source apportionment and joint effects on mortality in Cape Town, South Africa: University of Pretoria; 2021.

78. Gass K, Klein M, Sarnat SE, Winquist A, Darrow LA, Flanders WD, et al. Associations between ambient air pollutant mixtures and pediatric asthma emergency department visits in three cities: A classification and regression tree approach. Environmental Health. 2015; 14(1):58.

79. Lokotola C, Wichmann J, Wright C. Effect modification of temperature on air pollution associated with hospital admission for respiratory diseases in Cape Town, South Africa. Environmental Epidemiology. 2019; 3:249.

80. Olutola B, Wichmann J. Does apparent temperature modify the effects of air pollution on respiratory disease hospital admissions in an industrial area of South Africa? Clean Air Journal. 2021; 31(2):1-11. doi:10.17159/caj/2021/31/2.11366.

81. Govender K, Girdwood S, Letswalo D, Long L, Meyer-Rath G, Miot J. Primary healthcare seeking behaviour of low-income patients across the public and private health sectors in South Africa. BMC Public Health. 2021; 21(1):1-10.

CHAPTER 7: UNSUPERVISED MACHINE LEARNING TO INVESTIGATE THE JOINT EFFECTS OF SO_2, NO_2, O_3, $PM_{2.5}$ AND PM_{10} ON RESPIRATORY AND CARDIOVASCULAR HOSPITAL ADMISSIONS.

This chapter discusses the use of some unsupervised Machine Learning methods to determine the joint effect of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} on respiratory and cardiovascular hospital admissions in Vereeniging and Vanderbijlpark, in the Vaal Triangle Airshed Priority Area, from January 2011 to February 2020.

7.1. RESULTS

The same air pollution and hospital admission data were used as in Chapter 6.

7.1.1. DETERMINING OPTIMAL NUMBER OF CLUSTERS

Prior to clustering the air pollutant data, the data were scaled using standardised scaling. This is done in order for the model to better learn the data, creating uniformity among the different variables. Data prior to scaling and data after scaling can be seen in Figures 7.1 and 7.2, respectively. Thereafter, the optimal cluster size was determined using the elbow method and a silhouette method, illustrated in Figures 7.3 and 7.4, respectively.

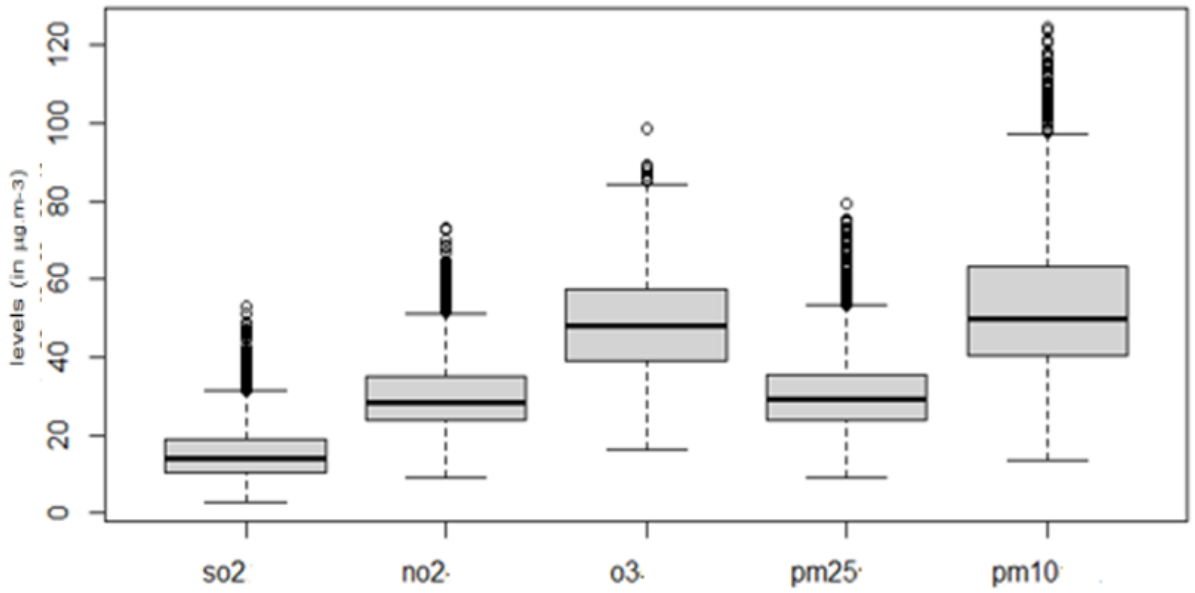


Figure 7.1: Air pollution data SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020, prior to using standardised scaling.

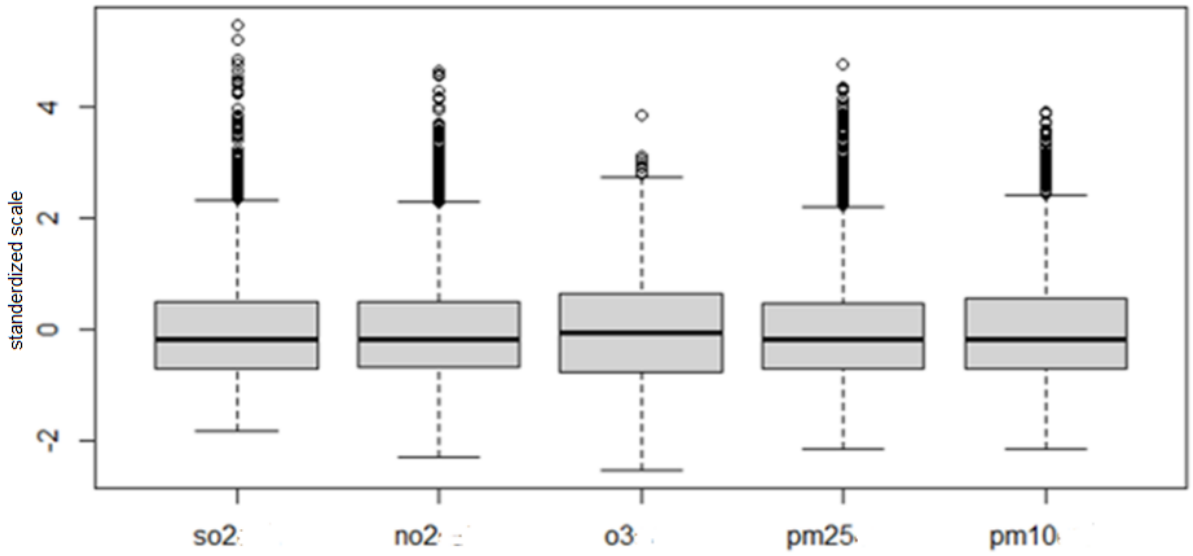


Figure 7.2: Air pollution data SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020, after standardised scaling.

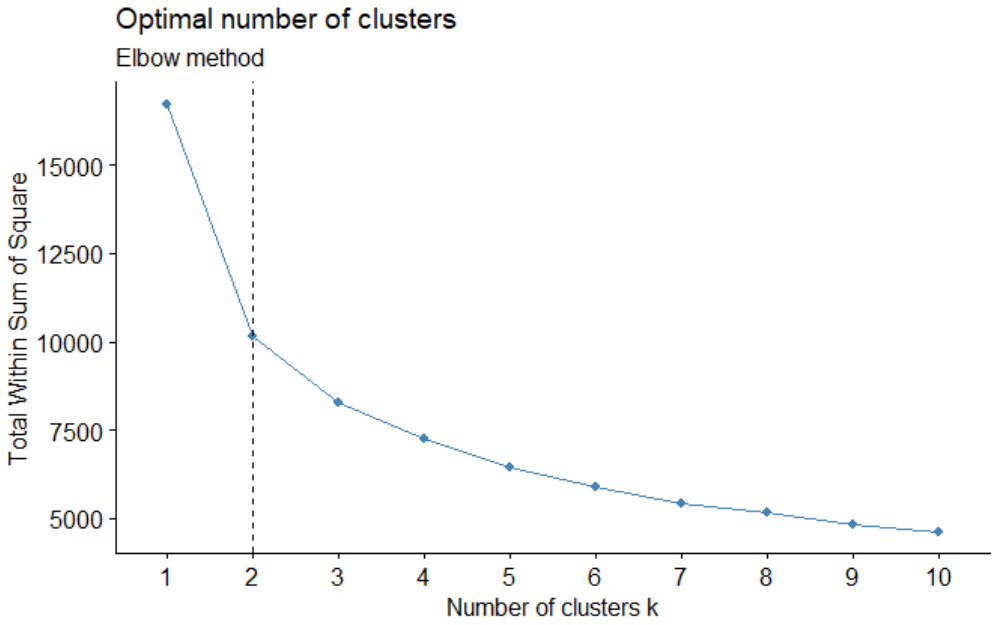


Figure 7.3: Optimal number of clusters according to 'elbow method' for SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

Figure 7.3 shows the optimal number of clusters for the dataset is two. This is where the 'elbow' bend is the most prominent. The silhouette method (Figure 7.4) also indicates two as the optimal number of clusters to run for the data in the clustering models.

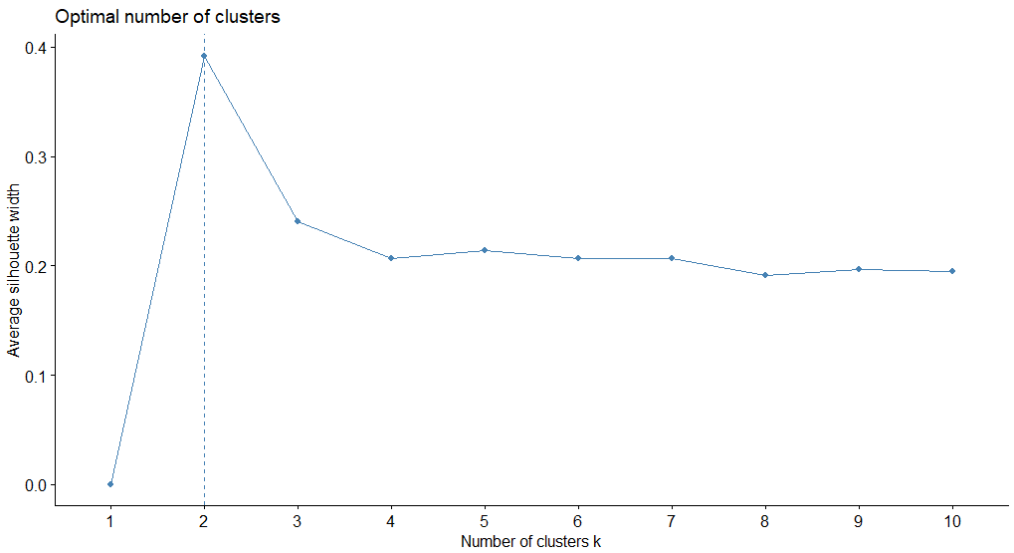


Figure 7.4: Optimal number of clusters according to 'silhouette' method for SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

Figure 7.5 shows a histogram of the best suggested number of clusters to use for the dataset. Two was the optimal number of clusters, but three was also taken as a cluster number for demonstration purposes. Lastly, as a third method to determine optimal number of clusters for clustering, an algorithm similar to principal component analysis shows that the data can only be classified into two factors, thus two is the optimal number of clusters (k) to run in the clustering models (Figure 7.6).

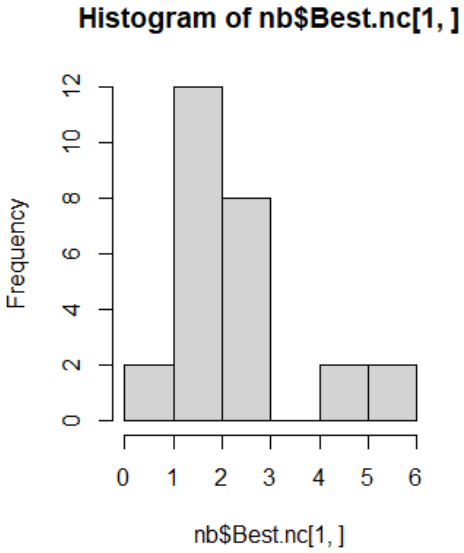


Figure 7.5: Histogram of best number of clusters for SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

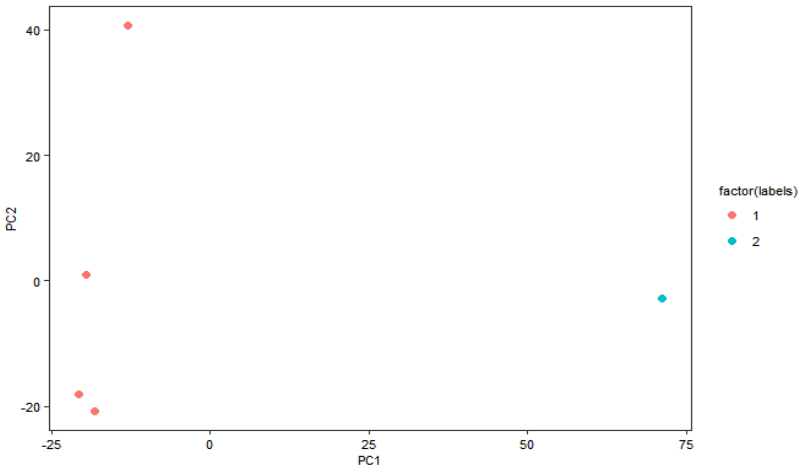


Figure 7.6: Using the Spectrum package to find of best number of clusters for SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020.

7.1.2. K-MEANS CLUSTERING

7.2.2.1. 2 CLUSTER K-MEANS MODEL

Figure 7.7 shows the two clusters formed from the air pollution data. Table 7.1 shows the descriptive statistics of RD hospital admissions and ambient air pollutant data (SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}) per cluster. Cluster one (n=867) had the lowest number of observations, but the highest mean concentration averages of SO_2, NO_2, $PM_{2.5}$, and PM_{10}. Within cluster one the highest average of RD hospital admissions was observed at ~20, which was higher than the initial average of RD hospital admissions (~16). However, concerning the CVD hospital admissions, the difference in the average hospital admissions was only slightly higher in cluster one than cluster two. Cluster two (n=2479) had the highest number of observations, but the lowest average of RD hospital admissions. In this cluster, O_3 was at its highest concentration average, while SO_2, NO_2, $PM_{2.5}$, and PM_{10} were relatively lower than in cluster one. Wilcoxon rank-sum tests showed that there were significant differences in SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentration levels between the two clusters (p-value < 0.01).

Figure 7.7: 2 k-means cluster model for SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020. Dim- Dimension reduction representing a certain amount of variation contained in the original dataset.

Table 7.1: Summary statistics of RD and CVD hospital admissions and daily air pollutants in VTPA for k-means with 2 clusters.

Variable	Mean	Min	P25	Median	P75	Max
Cluster 1						
N= 867						
RD hosp adm	19.82	4	15	19	24	55
CVD hosp adm	7.01	0	5	7	9	29
SO_2 ($\mu g.m^{-3}$)	21.70	4.15	16.55	20.43	25.52	63.57
NO_2 ($\mu g.m^{-3}$)	41.21	20.44	35.19	39.75	46.48	80.81
O_3 ($\mu g.m^{-3}$)	40.39	16.48	32.28	38.63	46.80	80.22
$PM_{2.5}$ ($\mu g.m^{-3}$)	43.17	20.26	36.22	42.00	48.27	79.41
PM_{10} ($\mu g.m^{-3}$)	75.10	31.51	64.27	73.45	124.63	85.35
Cluster 2						
N=2479						
RD hosp adm	15.18	1	11	14	19	53
CVD hosp adm	6.63	0	4	6	9	21
SO_2 ($\mu g.m^{-3}$)	13.27	2.99	9.56	12.44	15.96	42.70
NO_2 ($\mu g.m^{-3}$)	26.39	8.98	22.54	26.26	29.85	47.25
O_3 ($\mu g.m^{-3}$)	52.07	19.62	44.07	51.08	60.08	98.62
$PM_{2.5}$ ($\mu g.m^{-3}$)	26.53	8.93	22.31	26.60	30.59	47.96
PM_{10} ($\mu g.m^{-3}$)	45.36	13.57	37.22	45.41	52.32	89.71

Abbreviations: SO_2: sulphur dioxide; NO_2: nitrogen dioxide; O_3: Ozone; $PM_{2.5}$: particulate matter with an aerodynamic diameter of less than 2.5 μm; PM_{10}: particulate matter with an aerodynamic diameter of less than 10 μm; Tapp: apparent temperature; P25: 25^{th} percentile; P75: 75^{th} percentile, RD-respiratory disease, CVD-cardiovascular disease, hosp adm-hospital admissions.

The joint effects of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} on RD hospital admissions, obtained in the adjusted regression models, showed cluster mixture one did not significantly increase RD hospital admissions, with an RR of 1.04 (95% CI: 1.00, 1.07). No cluster mixtures were found to significantly increase CVD hospital admissions (p-value > 0.05).

7.2.1.3. 3 CLUSTER K-MEANS MODEL

The 3-cluster k-means model for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} is shown in Figure 7.8.Table 7.2 shows the descriptive statistics of RD hospital admissions and the ambient air pollutants (SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}). Cluster one (n=1519) had neither the highest nor lowest number of observations, with mean concentration averages of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}. Cluster two (n=517), however, had the

lowest number of observations, but showed the highest average of RD and CVD hospital admissions. The lowest average concentration level of O_3 was observed while SO_2, NO_2, $PM_{2.5}$, and PM_{10} were at their highest average concentration levels. Lastly, cluster three (n=1310) had the largest number of observations, as well as the lowest average of RD and CVD hospital admissions. The average O_3 concentration levels were the highest while SO_2, NO_2, $PM_{2.5}$, and PM_{10} were relatively lower in comparison to clusters one and two. Kruskal Wallis tests showed that there were significant differences in SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentration levels among the three clusters (p-value < 0.01).

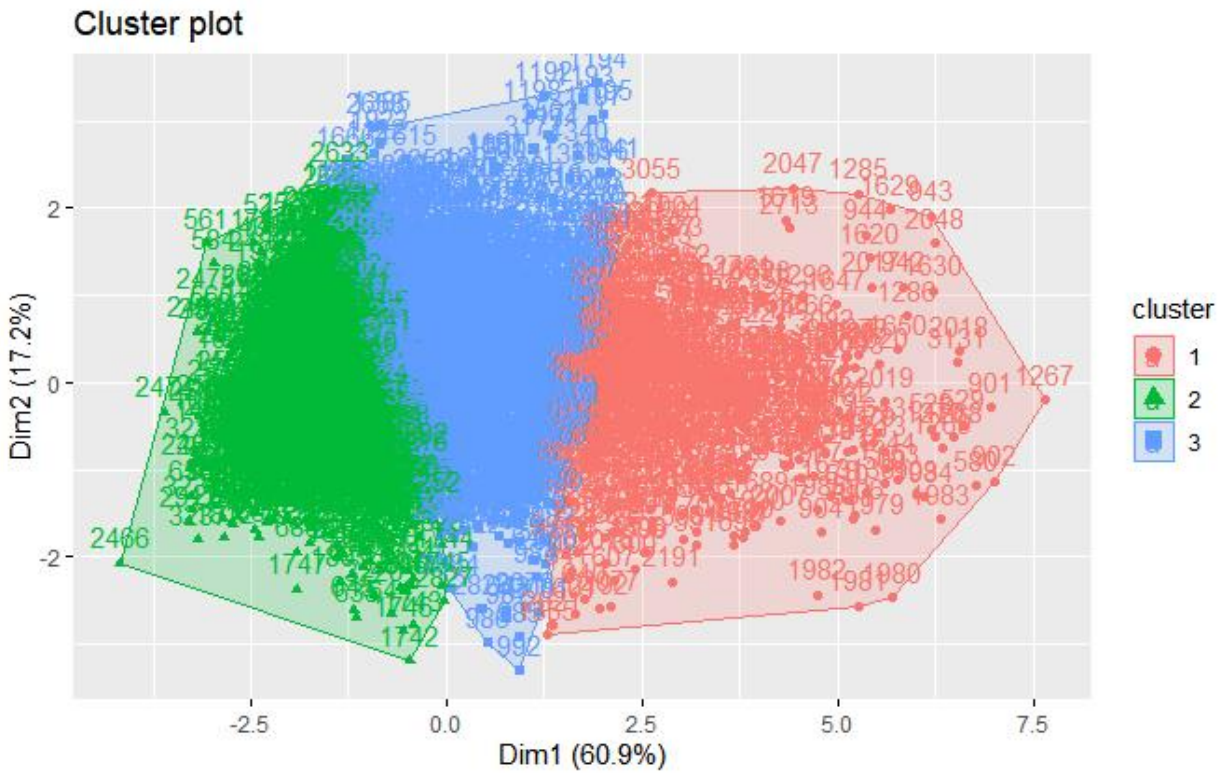


Figure 7.8: 3 k-means cluster for SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020. Dim- Dimension reduction representing a certain amount of variation contained in the original dataset.

Table 7.2: Summary statistics of RD and CVD hospital admissions and daily air pollutants for k-means with 3 clusters.

Variable	Mean	Min	P25	Median	P75	Max
N= 1519						
RD hosp adm	17.33	2	12	16	21	55
CVD hosp adm	6.77	0	4	6	9	21
SO_2 ($\mu g.m^{-3}$)	16.59	3.22	12.55	15.48	19.33	46.00
NO_2 ($\mu g.m^{-3}$)	30.54	13.88	26.83	30.01	34.07	50.47
O_3 ($\mu g.m^{-3}$)	44.68	19.34	37.33	44.83	50.43	81.96
$PM_{2.5}$ ($\mu g.m^{-3}$)	31.25	13.22	27.58	30.90	34.72	50.70
PM_{10} ($\mu g.m^{-3}$)	54.02	22.24	46.48	52.80	61.06	101.25
Cluster 2						
N=517						
RD hosp adm	20.06	5	15	19	24	53
CVD hosp adm	7.03	0	5	7	9	29
SO_2 ($\mu g.m^{-3}$)	22.93	10.22	17.87	21.55	26.64	53.06
NO_2 ($\mu g.m^{-3}$)	45.39	24.97	39.45	44.08	50.53	73.14
O_3 ($\mu g.m^{-3}$)	39.08	16.48	31.79	37.51	44.75	80.22
$PM_{2.5}$ ($\mu g.m^{-3}$)	47.94	29.99	41.35	46.53	52.97	79.41
PM_{10} ($\mu g.m^{-3}$)	83.00	49.23	72.64	81.88	91.00	124.63
Cluster 3						
N=1310						
RD hosp adm	13.82	1	10	13	17	48
CVD hosp adm	6.56	0	4	6	9	19
SO_2 ($\mu g.m^{-3}$)	11.19	2.99	8.41	10.41	13.12	32.64
NO_2 ($\mu g.m^{-3}$)	23.90	8.98	20.82	23.80	26.94	40.57
O_3 ($\mu g.m^{-3}$)	58.04	26.90	50.61	57.58	65.51	98.62
$PM_{2.5}$ ($\mu g.m^{-3}$)	23.62	8.93	20.03	23.35	26.88	43.41
PM_{10} ($\mu g.m^{-3}$)	40.16	13.57	33.08	39.85	46.85	81.78

Abbreviations: SO_2: sulphur dioxide; NO_2: nitrogen dioxide; O_3: Ozone; $PM_{2.5}$: particulate matter with an aerodynamic diameter of less than 2.5 μm; PM_{10}: particulate matter with an aerodynamic diameter of less than 10 μm; Tapp: apparent temperature; P25: 25th percentile; P75: 75th percentile, RD-respiratory disease, CVD-cardiovascular disease, hosp adm-hospital admissions.

Despite the significant difference among the air pollutants' concentrations levels, there were no significant cluster mixtures that increased RD or CVD hospital admissions (p-values > 0.05).

7.1.3. SPECTRAL CLUSTERING

7.1.2.1. TWO CLUSTERS USING LAPLACIAN MATRIX

Figure 7.9 shows the cluster distribution using spectral clustering with the Laplacian matrix. Table 7.3 shows the descriptive statistics of air pollutants and hospital admission data. In a Wilcoxon rank-sum test the air pollutants of the two clusters show to be significantly different for SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, (p-value < 0.01), respectively. However, the clusters do not show much difference in how the mean concentrations are in contrast with each other. Additionally, none of the cluster mixtures show to significantly increase risk of RD hospital admissions (p-value = 0.40) or CVD hospital admissions (p-value = 0.49).

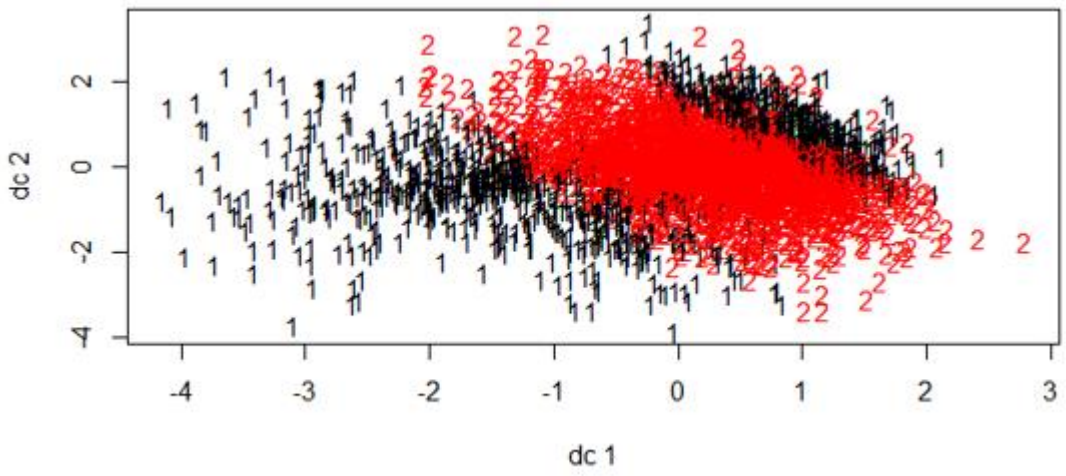


Figure 7.9: 2 cluster distribution for SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020, using spectral clustering the Laplacian matrix. dc-data centres

Table 7.3: Summary statistics of daily air pollutants in VTPA for spectral clustering with 2 clusters using Laplacian matrix.

Variable	Mean	Min	P25	Median	P75	Max
Cluster 1						
N= 1099	17.41	2.00	12.00	17.00	22.00	53.00
SO_2 ($\mu g.m^{-3}$)	15.48	3.29	10.64	14.35	18.89	48.89
NO_2 ($\mu g.m^{-3}$)	30.83	8.98	24.70	29.12	35.24	73.14
O_3 ($\mu g.m^{-3}$)	48.46	17.01	38.88	47.46	56.50	98.62
$PM_{2.5}$ ($\mu g.m^{-3}$)	30.62	10.27	23.56	29.15	35.32	72.18
PM_{10} ($\mu g.m^{-3}$)	52.24	15.96	39.09	49.52	61.20	118.45
Cluster 2						
N=2247	15.88	1.00	11.00	15.00	20.00	55.00
SO_2 ($\mu g.m^{-3}$)	15.45	2.99	10.40	14.01	18.94	53.06
NO_2 ($\mu g.m^{-3}$)	29.94	11.12	23.58	28.18	34.77	72.63
O_3 ($\mu g.m^{-3}$)	49.33	16.48	39.34	48.67	57.96	89.41
$PM_{2.5}$ ($\mu g.m^{-3}$)	30.95	8.93	23.80	29.13	35.69	79.41
PM_{10} ($\mu g.m^{-3}$)	53.47	13.57	40.71	49.98	63.87	124.63

Abbreviations: SO_2: sulphur dioxide; NO_2: nitrogen dioxide; O_3: Ozone; $PM_{2.5}$: particulate matter with an aerodynamic diameter of less than 2.5 μm; PM_{10}: particulate matter with an aerodynamic diameter of less than 10 μm; Tapp: apparent temperature; P25: 25th percentile; P75: 75th percentile, RD-respiratory disease, CVD-cardiovascular disease, hosp adm-hospital admissions.

7.1.2.2. THREE CLUSTERS USING LAPLACIAN MATRIX

Figure 7.10 shows the cluster distribution using spectral clustering with the Laplacian matrix. Table 7.4 shows the descriptive statistics of air pollutants and hospital admission data. The cluster distribution of observations in cluster three was very low (n=11). Interestingly, SO_2, NO_2, $PM_{2.5}$, and PM_{10} mean concentration levels are highest in this cluster three and lowest in cluster two, which has the highest number of observations (Table 7.4). A Kruskal Wallis test showed that SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} in each cluster were significantly different (p-value < 0.01). None of the cluster mixtures significantly increased the risk of RD or CVD hospital admission.

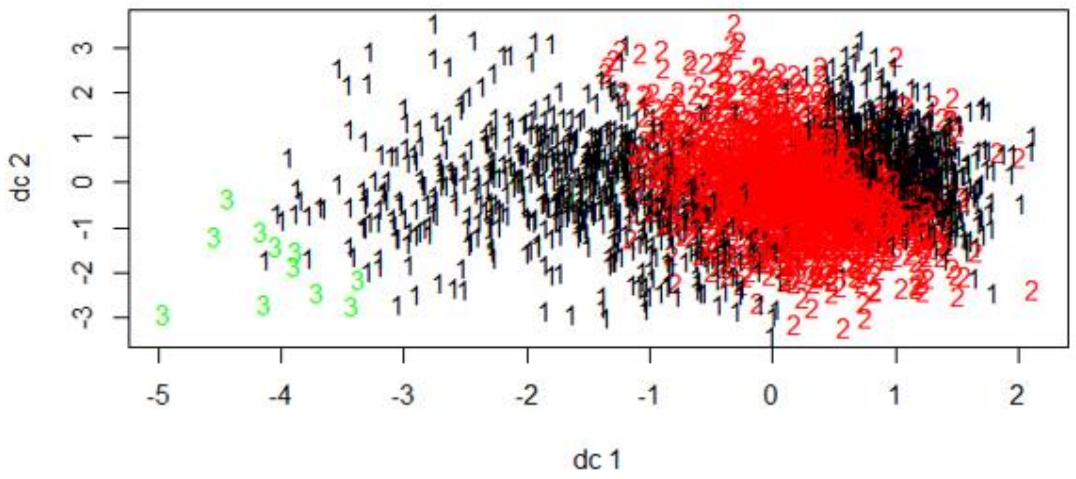


Figure 7.10: 3 cluster distribution for SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020, using spectral clustering the Laplacian matrix. dc-data centres.

Table 7.4: Summary statistics of daily air pollutants for spectral clustering with 3 clusters.

Variable	Mean	Min	P25	Median	P75	Max
Cluster 1						
N= 1343	17.01	2.00	12.00	16.00	21.00	53.00
SO_2 ($\mu g.m^{-3}$)	16.56	2.99	9.98	15.19	21.35	51.15
NO_2 ($\mu g.m^{-3}$)	32.72	8.98	22.65	31.09	41.22	73.14
O_3 ($\mu g.m^{-3}$)	46.38	16.48	35.91	44.75	55.85	89.41
$PM_{2.5}$ ($\mu g.m^{-3}$)	33.17	8.93	21.26	30.72	44.08	74.87
PM_{10} ($\mu g.m^{-3}$)	57.15	13.57	35.52	51.68	78.15	121.60
Cluster 2						
N=1992	15.94	1.00	11.00	15.00	20.00	55.00
SO_2 ($\mu g.m^{-3}$)	14.60	3.22	10.68	13.68	17.50	41.32
NO_2 ($\mu g.m^{-3}$)	28.41	9.25	24.56	28.10	31.86	52.57
O_3 ($\mu g.m^{-3}$)	50.88	19.34	42.75	49.80	58.74	98.62
$PM_{2.5}$ ($\mu g.m^{-3}$)	29.06	12.63	24.97	28.89	32.58	50.70
PM_{10} ($\mu g.m^{-3}$)	49.98	18.84	42.56	49.52	56.25	91.80
Cluster 3						
N=11	21.00	9.00	15.00	21.00	24.00	44.00
SO_2 ($\mu g.m^{-3}$)	36.77	26.42	31.05	34.88	41.58	53.06
NO_2 ($\mu g.m^{-3}$)	56.05	47.67	52.47	55.45	58.10	67.09
O_3 ($\mu g.m^{-3}$)	41.34	30.55	35.10	38.36	46.74	55.62
$PM_{2.5}$ ($\mu g.m^{-3}$)	69.08	58.55	64.67	70.08	73.31	79.41
PM_{10} ($\mu g.m^{-3}$)	114.59	106.34	109.66	113.47	118.36	124.63

Abbreviations: SO_2: sulphur dioxide; NO_2: nitrogen dioxide; O_3: Ozone; $PM_{2.5}$: particulate matter with an aerodynamic diameter of less than 2.5 μm; PM_{10}: particulate matter with an aerodynamic diameter of less than 10 μm; Tapp: apparent temperature; P25: 25th percentile; P75: 75th percentile, RD-respiratory disease, CVD-cardiovascular disease, hosp adm-hospital admissions.

7.1.2.3. TWO CLUSTERS USING NORMALISED LAPLACIAN MATRIX

Figure 7.11 shows the cluster distribution using spectral clustering with the normalised Laplacian matrix. Table 7.5 shows the descriptive statistics of the air pollutants and hospital admission data per cluster. Cluster one had the lowest number of observations, but the highest mean concentrations levels of SO_2, NO_2, $PM_{2.5}$, and PM_{10}, with the lowest mean concentration level of O_3. Cluster two had the inverse means

concentration levels for the air pollutants, with the highest O_3 mean concentration levels. The Wilcoxon rank-sum test showed that the SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} mean concentrations were significantly different in each cluster (p-value < 0.01).

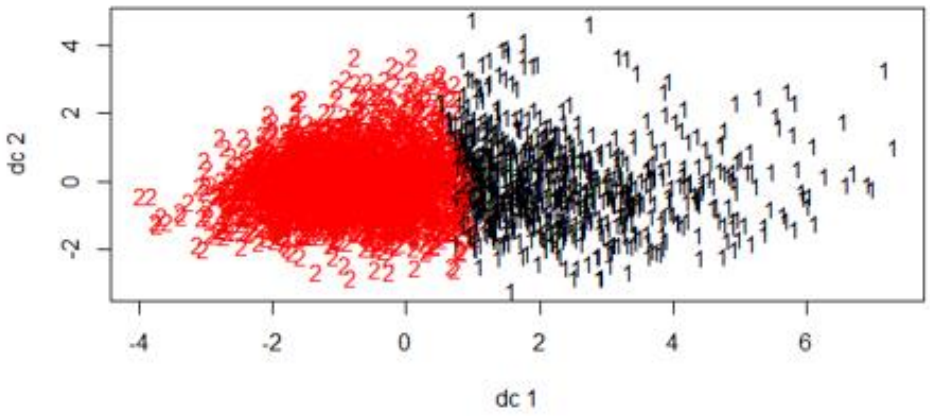


Figure 7.11: 2 cluster distribution for SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020, using spectral clustering the normalised Laplacian matrix. dc-data centres

Table 7.5: Summary statistics of daily air pollutants in VTPA for spectral clustering with 2 clusters using normalised Laplacian matrix

Variable	Mean	Min	P25	Median	P75	Max
Cluster 1						
N= 858	19.35	4.00	14.00	18.00	24.00	53.00
SO_2 ($\mu g.m^{-3}$)	21.60	7.67	16.44	20.36	25.52	53.06
NO_2 ($\mu g.m^{-3}$)	40.79	18.13	34.49	39.59	46.25	73.14
O_3 ($\mu g.m^{-3}$)	46.09	16.48	34.83	43.41	55.70	89.41
$PM_{2.5}$ ($\mu g.m^{-3}$)	42.88	17.42	35.54	41.44	48.36	79.41
PM_{10} ($\mu g.m^{-3}$)	74.97	33.85	64.41	73.87	85.49	124.63
Cluster 2						
N=2488	15.36	1.00	11.00	14.00	19.00	55.00
SO_2 ($\mu g.m^{-3}$)	13.34	2.99	9.58	12.47	16.07	40.40
NO_2 ($\mu g.m^{-3}$)	26.59	8.98	22.57	26.44	30.12	52.57
O_3 ($\mu g.m^{-3}$)	50.07	19.34	41.45	49.29	57.98	98.62
$PM_{2.5}$ ($\mu g.m^{-3}$)	26.69	8.93	22.32	26.64	30.70	47.99
PM_{10} ($\mu g.m^{-3}$)	45.52	13.57	37.24	45.58	52.75	89.71

Abbreviations: SO_2: sulphur dioxide; NO_2: nitrogen dioxide; O_3: Ozone; $PM_{2.5}$: particulate matter with an aerodynamic diameter of less than 2.5 μm; PM_{10}: particulate matter with an aerodynamic diameter of less than 10 μm; Tapp: apparent temperature; P25: 25th percentile; P75: 75th percentile, RD-respiratory disease, CVD-cardiovascular disease, hosp adm-hospital admissions.

Although the air pollutant concentrations did differ per cluster, the cluster mixtures significantly increased the risk of RD hospital admission. However, cluster mixture one barely showed significance (p-value = 0.0503). No cluster mixtures showed to significantly increase CVD hospital admissions.

7.1.2.4. THREE CLUSTERS USING NORMALISED LAPLACIAN MATRIX

Figure 7.12 shows the 3-cluster distribution using spectral clustering with the normalised Laplacian matrix. Table 7.6 shows the descriptive statistics of air pollutants and hospital admission data. The concentration levels in the different clusters showed to significantly differ among SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} (p-value < 0.01). Cluster three had the highest SO_2, NO_2, $PM_{2.5}$, and PM_{10} levels, but this cluster mixture did not significantly increase the risk of RD hospital admissions. Cluster one did have higher SO_2, NO_2, $PM_{2.5}$, and PM_{10} mean concentrations levels and did show to significantly increase the risk of RD hospital admissions by an RR of 1.04 (95% CI, 1.01-1.08). Cluster mixture two had the highest O_3 mean concentration levels and the lowest SO_2, NO_2, $PM_{2.5}$, and PM_{10} mean concentrations levels, and showed a significant RR of 0.96 (95% CI 0.94, 0.99). This suggests that for cluster mixture two, there was a lower risk of RD hospital admission compared to cluster mixtures one and three. However, none of the cluster mixtures showed to significantly increase CVD hospital admissions.

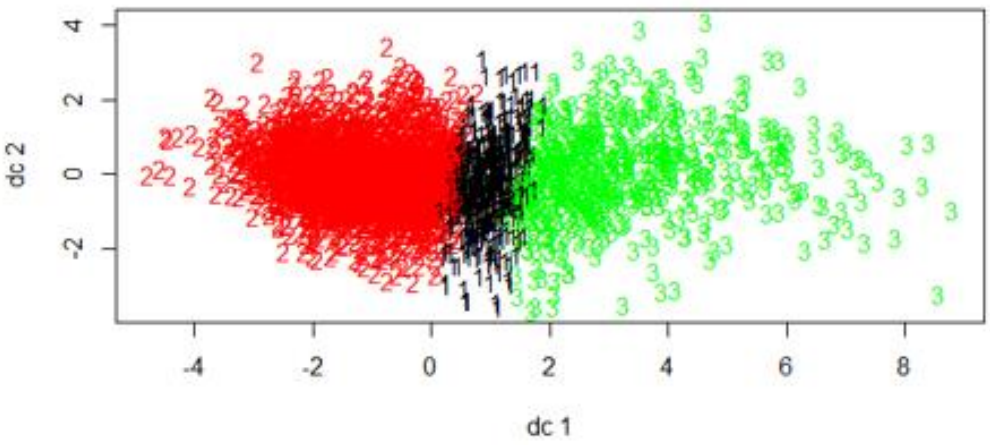


Figure 7.12: 3 cluster distribution for SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020, using spectral clustering the normalised Laplacian matrix. Dc- data centres

Table 7.6: Summary statistics of daily air pollutants for spectral clustering with 3 clusters.

Variable	Mean	Min	P25	Median	P75	Max
Cluster 1						
N= 508	18.05	1.00	12.00	17.00	22.00	55.00
SO_2 ($\mu g.m^{-3}$)	18.11	4.15	13.71	17.07	21.12	42.70
NO_2 ($\mu g.m^{-3}$)	33.11	18.13	24.19	28.25	32.47	53.48
O_3 ($\mu g.m^{-3}$)	49.67	19.84	39.34	48.55	58.24	86.53
$PM_{2.5}$ ($\mu g.m^{-3}$)	34.53	17.42	30.91	34.48	38.27	50.70
PM_{10} ($\mu g.m^{-3}$)	60.30	31.75	51.74	60.15	68.19	91.80
Cluster 2						
N=2232	15.12	1.00	11.00	14.00	19.00	53.00
SO_2 ($\mu g.m^{-3}$)	12.87	2.99	9.34	12.07	15.40	39.69
NO_2 ($\mu g.m^{-3}$)	25.99	8.98	22.25	25.90	29.30	47.59
O_3 ($\mu g.m^{-3}$)	50.08	19.34	41.65	49.37	57.84	98.62
$PM_{2.5}$ ($\mu g.m^{-3}$)	25.92	8.93	21.79	25.83	29.84	47.99
PM_{10} ($\mu g.m^{-3}$)	44.09	13.57	36.29	44.25	51.23	81.14
Cluster 3						
N=606	19.66	4.00	15.00	19.00	24.00	49.00
SO_2 ($\mu g.m^{-3}$)	22.75	7.74	17.55	21.48	26.48	53.06
NO_2 ($\mu g.m^{-3}$)	43.48	23.32	37.71	42.27	48.78	73.14
O_3 ($\mu g.m^{-3}$)	44.72	16.48	33.83	41.56	54.67	89.41
$PM_{2.5}$ ($\mu g.m^{-3}$)	45.87	22.82	38.72	45.16	51.61	79.41
PM_{10} ($\mu g.m^{-3}$)	80.07	40.42	69.57	80.08	89.60	124.63

Abbreviations: SO_2: sulphur dioxide; NO_2: nitrogen dioxide; O_3: Ozone; $PM_{2.5}$: particulate matter with an aerodynamic diameter of less than 2.5 μm; PM_{10}: particulate matter with an aerodynamic diameter of less than 10 μm; Tapp: apparent temperature; P25: 25^{th} percentile; P75: 75^{th} percentile, RD-respiratory disease, CVD-cardiovascular disease, hosp adm-hospital admissions.

7.1.4. DENSITY BASED SPATIAL CLUSTERING WITH APPLICATION OF NOISE (DBSCAN) CLUSTERING

Similar to k-means and spectral clustering, there are methods to determine optimal parameters, such as the radius of a circle for a core point (eps).[1] Figure 7.13 shows the different k-nearest neighbours (KNN) distances, to determine the eps for the air

pollution data that has not been scaled: (a) at 3 (KNN), (b) at 4 (KNN), and (c) at 5 (KNN). The optimal radius was 0.7.

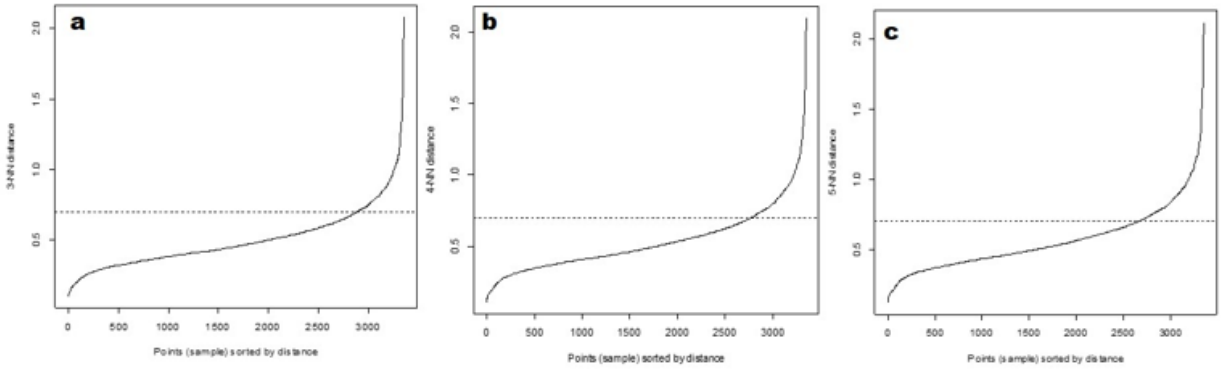


Figure 7.13: KNN graphs to determine optimal radius for dbscan clustering SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020. This data was scaled.

Different reachability minimum numbers of points (MinPts) were used to run the clusters. Figure 7.14 shows the different cluster plots at: (a) 3 MinPts, (b) 4 MinPts, and (c) 5 MinPts. Table 7.7 shows how the DBSCAN clustering algorithm distributed the number of points into clusters. As the minimum number of points increase, the amount of noise also increases. 'Noise' being the observations not assigned a cluster. Furthermore, each cluster shows that the first cluster has the highest number of observations and other clusters seem to have a negligible number of observations.

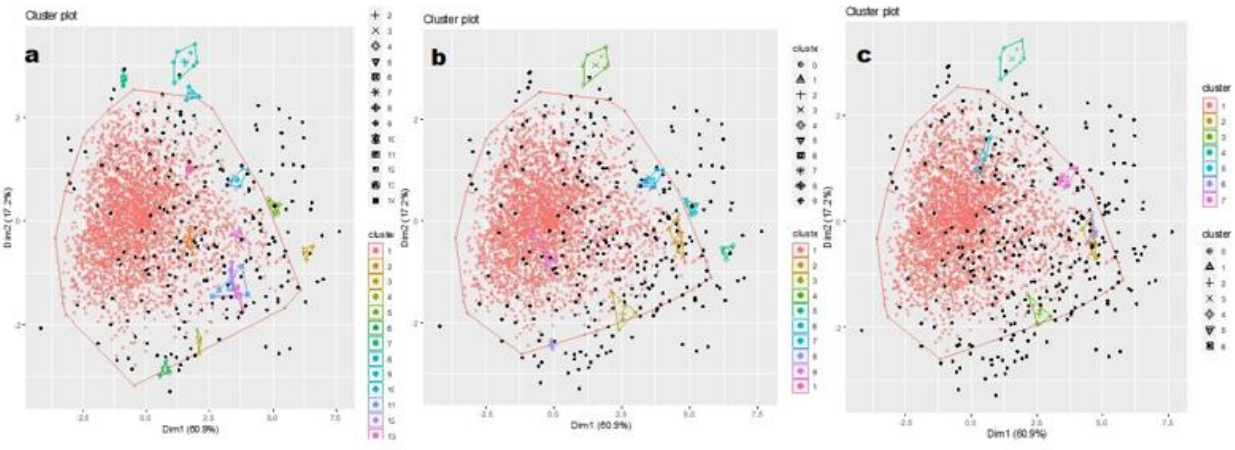


Figure 7.14: Cluster plots for dbscan clustering at a) 3 minimum number of points, b) 4 minimum number of points and c) 5 minimum number of points, for SO_2, NO_2, O_3, $PM_{2.5}$ and PM_{10} in VTAPA, South Africa during 2 January 2011 to 29 February 2020. Dim- Dimension reduction representing a certain amount of variation contained in the original dataset.

Table 7.7: Distribution of observation in each cluster made via dbscan clustering at different minimum number of points.

3 MinPts	noise	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15
n	206	3082	3	4	3	6	3	3	8	3	7	5	3	4	3	3
4 MinPts	noise	c1	c2	c3	c4	c5	c6	c7	c8	c9	c10					
n	264	3037	6	6	8	4	6	7	2	3	3					
5 MinPts	noise	c1	c2	c3	c4	c5	c6	c7								
n	313	2997	6	5	8	5	5	7								

MinPts- reachability minimum number of points, n=number of observations, c-cluster

7.2. DISCUSSION

The aim of this part of the project was to use unsupervised Machine Learning (ML) clustering to determine the joint effects of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} mixtures on RD and CVD hospital admissions.

Supervised ML is often used to investigate health outcomes based on the different variables.[2] In this study, the hospital admissions data was already available. Thus, methods such as random forest and decision trees would be inappropriate and better suited if hospital admissions were being predicted from the air pollutant data. In this case, the grouping of the five air pollutants' data were unknown (unlabelled) and the grouping of this data are needed in order to investigate its potential impact on the RD and CVD hospital admissions. Furthermore, designating a testing and training dataset was not required prior to running the clustering algorithms, since this is more commonly done for supervised ML techniques.[3-6]

The data for all runs was scaled in order to prevent outliers, minimums, and maximums from having too much influence in the clustering models. Setting the number of clusters in a k-means algorithm is subjective to the researcher,[7] but there are functions that can be applied to provide an optimal number of clusters.[8-10] By determining the optimal number of parameters, such as clusters and radius measures, a point of reference is formed for the researcher. The number of clusters and radiuses were increased for comparison purposes.

K-means clustering has been the more popularly used ML clustering method in studies similar to this project.[11-15] These studies used both supervised and unsupervised ML k-means clustering, mainly to determine air pollution mixtures. Although most of the studies found significant increased risk of negative health outcomes from exposure to the pollution mixtures,[11,13-14] significant risks for RD and CVD hospitalisation were not

observed from the project results. The results did, however, show distinct differences in SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} concentrations levels among the formed clusters. The cluster models allowed for five air pollutants to be added into one model and created the pollution mixtures. Other studies have investigated the effects of individual pollutants such as SO_2, NO_2, CO, FeNO, $PM_{2.5}$, and PM_{10} on RD and CVD hospital admissions.[16-18] While other studies have included more than one pollutant in the regression models, the number of pollutants were often limited to two or three pollutants for the association with hospital admissions.[19-21] The CART analyses done in Chapter 6 only permitted three pollutants per model and resulted in the use of seven different mixtures.

Spectral clustering has been found to be a better preforming clustering method compared to k-means clustering.[22-23] The results show that there was a clear difference in results when applying the different matrices. The normalised Laplacian matrix showed more reasonable results. The normalised Laplacian matrix outperforms the Laplacian matrix, as it considers the heterogeneous distribution of node-degree when performing spectral analysis.[24] Spectral clustering is said to depend on the affinity matrix used to improve the clustering.[25] This could explain the difference in observation distribution when a three-cluster model was run using the Laplacian matrix. The two-cluster model mixtures obtained under the Laplacian and normalised Laplacian matrices, did not significantly increase the risk of RD and CVD hospital admission.

The three-cluster spectral model, using the normalised Laplacian matrix, showed to significantly increase RD hospital admission when a cluster mixture had SO_2, NO_2, $PM_{2.5}$, and PM_{10} at higher concentration levels and O_3 at lower concentration levels. However, when O_3 had higher concentration levels, patients were at lower risk of RD hospital admission compared to the other cluster mixtures. The reduced risk in the presence of higher O_3 concentrations would suggest the pollutant is not as harmful as the other pollutants. This suggestion is contrary to other studies that have shown O_3 to have adverse effects on the respiratory system.[26-27] Studies have also shown that O_3 in the presence of PM is associated with increased risk for conditions such as stroke.[28-29] This is contrary to the findings in this study which found that none of the cluster mixtures formed increased the risk of CVD hospital admmissions. Majority of the health studies that have used ML, have used supervised ML methods,[15,30-34] and few used unsupervised ML methods.

Interestingly, in the CART analyses (in Chapter 6), there was an indication that some mixtures with O_3 in lower quartiles had a reduced risk of RD hospital admissions. The increased risk of RD hospitalisation seems to be driven by a combination of higher $PM_{2.5}$ and NO_2, or PM_{10} and NO_2 concentrations. A study that used k-means clustering on air pollutants to investigate the association with mixtures and type 2 diabetes mellitus incidence,[14] showed an increased incidence for cluster mixtures containing higher concentrations of SO_2, NO_2, CO, $PM_{2.5}$, and PM_{10}, as compared to clusters with higher O_3 concentration levels. Results from the aforementioned study and this research project suggests that O_3 exposure may not have as detrimental an effect on health as SO_2, NO_2, $PM_{2.5}$, and PM_{10}. Perhaps, the inverse relationship between O_3 and the other air pollutants could be a key factor in the results.[35-36]

Another well-known clustering method is the DBSCAN. This method can cluster nested data at any shape, where k-means clusters data at fixed shapes.[1] The DBSCAN results were not used to run further associative investigations on RD and CVD hospital admissions. Although clusters were formed, it placed some observations as noise and reduced the dataset for comparison with the RD and CVD hospital admissions data. Furthermore, the clusters formed were not usable for further associative investigations, because of the evident severe irregularity in observation distribution.

Clustering-based approaches, such as k-means clustering, have proven useful in the definition of multipollutant exposure profiles in estimating the associations between air pollution and diseases, such as breast cancer and diabeties.[11-13,37] The main advantages to clustering-based approaches are computational speed, large dataset capacity, and its ability to find pure sub-clusters.[38-41] However, k-means does not identify outliers, is quite restricted to data that has a centroid, and cluster dimensions are limited to spherical formations.[41] Spectral clustering, on the other hand, is more versatile and does not hold the same restrictions in cluster formation as k-means.[23] The findings do not clearly show whether spectral clustering is superior to k-means. However, using the 3-cluster spectral clustering model yielded reasonable results regarding the cluster mixtures' effects on RD hospital admissions.

The Riches et al,[14] study further investigated $PM_{2.5}$ and its different elemental species. In their analysis they keep extreme values and use a k-medoids clustering algorithm to mitigate for k-means' sensitivity to outliers.[14] This addition to the study can be included in further studies. Few studies have used unsupervised ML to investigate air pollution

mixtures' effects on health outcomes. Nonetheless, public health data are ever-growing in both scale and complexity and ML shows promise for addressing how to handle such large data sets.[42]

7.3. CONCLUSION

Unsupervised ML could have a place in determining joint effects of air pollutant mixtures on hospital admission and other health outcomes. The clustering methods used in this study were quick to run and analyse. Using the clustering methods was also less time consuming in comparison to the CART analyses. Unsupervised ML clustering allowed for more than three air pollutants in the mixture as compared to CART analyses. The process also showed promise for analysing multiple air pollutants in a more probable mixture, despite the different interactions. However, it is evident that more studies are needed before considering unsupervised ML a reliable and definite tool to study joint effects of air pollution on different health outcomes. It is highly recommended that more studies be done using unsupervised ML clustering methods such as k-means and spectral clustering. The same process can be run in areas of the country with air quality and health outcome profiles different from the VTAPA.

7.4. REFERENCES

1. Chakraborty S, Nagwani NK, Dey L. Performance comparison of incremental k-means and incremental dbscan algorithms. arXiv preprint arXiv:1406.4751. 2014.

2. Kebalepile MM, Dzikiti LN, Voyi K. Supervised kohonen self-organizing maps of acute asthma from air pollution exposure. International Journal of Environmental Health Research. 2021; 18(21):11071.

3. Bhavsar H, Ganatra A. A comparative study of training algorithms for supervised machine learning. International Journal of Soft Computing and Engineering 2012; 2(4):2231-307.

4. Jabbar H, Khan RZ. Methods to avoid over-fitting and under-fitting in supervised machine learning (comparative study). Computer Science, Communication and Instrumentation Devices. 2015; 70:163-72.

5. Mahesh B. Machine learning algorithms-a review. International Journal of Science and Research. 2020; 9:381-6.

6. Nasteski V. An overview of the supervised machine learning methods. Horizons. b. 2017; 4:51-62.

7. Fränti P, Sieranoja S. How much can k-means be improved by using better initialization and repeats? Pattern Recognition. 2019; 93:95-112.

8. Dangeti P. Statistics for machine learning: Packt Publishing Ltd; 2017.

9. Saputra DM, Saputra D, Oswari LD, editors. Effect of distance metrics in determining k-value in k-means clustering using elbow and silhouette method. Sriwijaya International Conference on Information Technology and Its Applications (SICONIAN 2019); 2020: Atlantis Press.

10. Yuan C, Yang H. Research on k-value selection method of k-means clustering algorithm. J. 2019; 2(2):226-35.

11. Keller J, editor. Predictive clustering methods to identify differential toxicity of air pollutant mixtures. ISEE Conference Abstracts; 2022.

12. Mistry S, Riches NO, Gouripeddi R, Facelli JC. Environmental exposures in machine learning and data mining approaches to diabetes etiology: A scoping review. Artificial Intelligence In Medicine. 2023; 135. doi:10.1016/j.artmed.2022.102461.

13. Riches NO, Gouripeddi R, Facelli JC. 73432 assessment of multi-pollutant ambient air composition on type 2 diabetes mellitus using machine learning. Journal of Clinical and Translational Science. 2021; 5(s1):46.

14. Riches NO, Gouripeddi R, Payan-Medina A, Facelli JC. K-means cluster analysis of cooperative effects of CO, NO2, O3, PM2.5, PM10, and SO2 on incidence of type 2 diabetes mellitus in the US. Environmental Research. 2022; 212(Part B). doi:10.1016/j.envres.2022.113259.

15. Zhang D, Du L, Wang W, Zhu Q, Bi J, Scovronick N, et al. A machine learning model to estimate ambient PM2.5 concentrations in industrialized highveld region of South Africa. Remote Sensing of Environment. 2021; 266. doi:10.1016/j.rse.2021.112713.

16. Abdolahnejad A, Jafari N, Mohammadi A, Miri M, Hajizadeh Y, Nikoonahad A. Cardiovascular, respiratory, and total mortality ascribed to PM10 and PM2.5 exposure in Isfahan, Iran. Journal of Education and Health Promotion. 2017; 6:109. doi:10.4103/jehp.jehp_166_16.

17. Capraz O, Deniz A, Dogan N. Effects of air pollution on respiratory hospital admissions in Istanbul, Turkey, 2013 to 2015. Chemosphere. 2017; 181:544-50. doi:10.1016/j.chemosphere.2017.04.105.

18. Hart JE, Puett RC, Rexrode KM, Albert CM, Laden F. Effect modification of long-term air pollution exposures and the risk of incident cardiovascular disease in US women. Journal of the American Heart Association. 2015; 4(12).

19. Mwase N, Olutola B, Wichmann J. Temperature modifies the association between air pollution and respiratory disease hospital admissions in an industrial area of South Africa: The Vaal Triangle Air Pollution Priority Area. Clean Air Journal. 2022; 32(2). doi:10.17159.caj.2022.32.2.14588.

20. Olutola B, Wichmann J. Does apparent temperature modify the effects of air pollution on respiratory disease hospital admissions in an industrial area of South Africa? Clean Air Journal. 2021; 31(2):1-11. doi:10.17159/caj/2021/31/2.11366.

21. Yao Y, Chen X, Chen W, Wang Q, Fan Y, Han Y, et al. Susceptibility of individuals with chronic obstructive pulmonary disease to respiratory inflammation associated with short-term exposure to ambient air pollution: A panel study in Beijing. Science of the Total Environment. 2021; 766: 142639. doi:10.1016/j.scitotenv.2020.142639.

22. Langone R, Van Barel M, Suykens Johan AK. Efficient evolutionary spectral clustering. Pattern Recognition Letters. 2016; 84:78-84. doi:10.1016/j.patrec.2016.08.012.

23. von Luxburg U, Belkin M, Bousquet O. Consistency of spectral clustering. The Annals of Statistics. 2008; 36(2):555-86.

24. Shen H-W, Cheng X-Q. Spectral methods for the detection of network community structure: A comparative analysis. Journal of Statistical Mechanics: Theory and Experiment. 2010; 2010(10):P10020.

25. Chang H, Yeung D-Y. Robust path-based spectral clustering. Pattern Recognition. 2008; 41(1):191-203. doi:10.1016/j.patcog.2007.04.010.

26. Nhung NTT, Schindler C, Dien TM, Probst-Hensch N, Künzli N. Association of ambient air pollution with lengths of hospital stay for Hanoi children with acute lower-respiratory infection, 2007-2016. Environmental Pollution. 2019; 247:752-62. doi:10.1016/j.envpol.2019.01.115.

27. Zheng X-Y, Orellano P, Lin H-L, Jiang M, Guan W-J. Short-term exposure to ozone, nitrogen dioxide, and sulphur dioxide and emergency department visits and hospital admissions due to asthma: A systematic review and meta-analysis. Environment International. 2021; 150:106435. doi:10.1016/j.envint.2021.106435.

28. Ljungman PL, Mittleman MA. Ambient air pollution and stroke. Stroke. 2014; 45:3734-41.

29. Shin H, Burnett R, Cohen A, Hubbell BJ. Outdoor fine particles and nonfatal strokes: Systematic review and meta-analysis. Epidemiology. 2014; 25:835-42.

30. Aguiar FS, Torres RC, Pinto JV, Kritski AL, Seixas JM, Mello FC. Development of two artificial neural network models to support the diagnosis of pulmonary tuberculosis in hospitalized patients in Rio de Janeiro, Brazil. Medical & Biological Engineering & Computing. 2016; 54(11):1751-9. doi:10.1007/s11517-016-1465-1.

31. Andrade BB, Reis-Filho A, Barros AM, Souza-Neto SM, Nogueira LL, Fukutani KF, et al. Towards a precise test for malaria diagnosis in the Brazilian Amazon: Comparison among field microscopy, a rapid diagnostic test, nested PCR, and a computational expert system based on artificial neural networks. Malaria Jounal. 2010; 9:117. doi:10.1186/1475-2875-9-117.

32. Dlamini Z, Francies FZ, Hull R, Marima R. Artificial intelligence (AI) and big data in cancer and precision oncology. Computational and Structural Biotechnology Journal. 2020; 18:2300-11. doi:10.1016/j.csbj.2020.08.019.

33. Moyo S, Doan TN, Yun JA, Tshuma N. Application of machine learning models in predicting length of stay among healthcare workers in underserved communities in South Africa. Human resources for health. 2018; 16(1):68. doi:10.1186/s12960-018-0329-1.

34. van Heerden A, Young S. Use of social media big data as a novel HIV surveillance tool in South Africa. PLoS One. 2020; 15(10):e0239304. doi:10.1371/journal.pone.0239304.

35. Chen KS, Ho YT, Lai CH, Tsai YA, Chen SJ. Trends in concentration of ground-level ozone and meteorological conditions during high ozone episodes in the Kao-Ping airshed, Taiwan. Journal of the Air & Waste Management Association. 2004; 54(1):36-48.

36. Dai H, Zhu J, Liao H, Li J, Liang M, Yang Y, et al. Co-occurrence of ozone and PM2.5 pollution in the Yangtze River Delta over 2013-2019: Spatiotemporal distribution and meteorological conditions. Atmospheric Research. 2021; 249 doi:10.1016/j.atmosres.2020.105363.

37. White AJ, Keller JP, Zhao S, Carroll R, Kaufman JD, Sandler DP. Air pollution, clustering of particulate matter components, and breast cancer in the sister study: A US-wide cohort. Environmental Health Perspectives. 2019; 127(10):107002.

38. Aggarwal CC. Data clustering: algorithms and applications.Ser. Chapman & hall/crc data mining and knowledge discovery series. CRC Press; 2013.

39. Chen M, Wang P, Chen Q, Wu J, Chen X. A clustering algorithm for sample data based on environmental pollution characteristics. Atmospheric Environment. 2015; 107:194-203. doi:10.1016/j.atmosenv.2015.02.042.

40. Jain AK, Murty MN, Flynn PJ. Data clustering a review. ACM Computing Surveys. 1999; 31(3):264-323. doi:10.1145/331499.331504.

41. Sonagara D, Badheka S. Comparison of basic clustering algorithms. International Journal of Computer Science and Mobile Applications. 2014; 3(10):58-61.

42. Goldsmith J, Sun Y, Fried LP, Wing J, Miller GW, Berhane K. The emergence and future of public health data science. Public Health Reviews. 2021; 42:1604023. doi:10.3389/phrs.2021.1604023.

CHAPTER 8: POSITIVE MATRIX FACTORISATION SOURCE APPORTIONMENT OF $PM_{2.5}$

This chapter summarises the modelling process and results by PMF on the $PM_{2.5}$ samples collected in Pretoria, using trace metals as markers. The samples were taken over a four-year period, from 18 April 2017 to 12 February 2021. The main sources and contributions to the total $PM_{2.5}$ included industry/base metal (8.7%), road traffic (11.3%), secondary sulphur (12.1%), mining (43.2%), biomass/coal burning (14.2%), resuspended dust (8.5%), and general exhaust (2.0%) emissions.

8.1. DESCRIPTIVE STATISTICS

A total of 428 samples were collected over the forty-six-month period, from 18 April 2017 to 12 February 2021, used in the PMF analysis. The $PM_{2.5}$ levels ranged from 0.3 $\mu g/m^3$ to 138.9 $\mu g/m^3$, with a mean of 21.8 ± 17.9. $\mu g/m^3$ (Table 8.1). The mean $PM_{2.5}$ level for the study period was lower than the daily National Air Quality Standard (NAAQS) of 40 $\mu g/m^3$, and only 65 out of 428 days exceeded the daily limit. The mean $PM_{2.5}$ level exceeded the daily WHO air quality guideline (15 $\mu g/m^3$),[1] with 238 out of 428 days being above this guideline.

During the study period, winter had the highest $PM_{2.5}$ mean concentration levels (37.9 ± 20.22 $\mu g/m^3$), followed by autumn (20.7 ± 17.9. $\mu g/m^3$), spring (17.3 ± 9.8. $\mu g/m^3$), and summer (10.64 ± 6.7. $\mu g/m^3$). There were 98, 57, 57, and 26 days that exceeded the WHO daily guideline in winter, autumn, spring, and summer, respectively. Then, 49, 13, 3, and 0 days that exceeded the NAAQS daily standard in winter, autumn, spring, and summer, respectively.

Table 8.1: Summary statistics of $PM_{2.5}$ measured at the School of Health System and Public Health, University of Pretoria between 18 April 2017 and 12 February 2021.

Variable	Mean	SD	Min	Max
$PM_{2.5}$	21.8	17.9	0.3	138.9
BC	2.8	2.5	<LoD	11.4
UV-PM	2.4	1.7	<LoD	7.6
S	1072.94	1639.92	0.39	11840.58
Cl	65.03	185.25	<LoD	1800.73
K	254.06	346.52	1.33	1917.24
Ca	145.06	156.84	<LoD	996.88
Ti	31.93	28.20	0.07	211.51
Fe	265.90	213.35	0.72	1453.88
Ni	47.14	79.36	0.05	266.26
Cu	14.13	31.43	<LoD	487.01
Zn	51.45	87.72	0.11	809.19
Br	16.60	24.98	0.04	253.67
U	2.94	10.86	<LoD	222.20
Si	579.63	654.52	4.14	5531.63

Units: $PM_{2.5}$ ($\mu g/m^3$), soot ($m^{-1} \times 10^{-5}$), BC ($\mu g/m^3$), UV-PM ($\mu g/m^3$), trace elements (ng/m^3), <LoD-under the limit of detection.

8.2. MODEL PARAMETERS SETTING

The data from April 2017 to March 2020 from two previous PhD studies was used, [2-4] thereafter, an additional 10 months was added for analysis. The additional samples were collected during the COVID-19 pandemic National Lockdown. There were five lockdown levels: lockdown level five was implemented between 27 March to 30 April 2020, followed by level four (1-31 May 2020), level three (1 June to 17 August 2020), level two (18 August to 20 September 2020), and level one (21 September to 28 December 2020). South Africa was on an adjusted lockdown level three from 29 December 2020 until the end of the PMF study period (12 February 2021).

The relationship between concentration and corresponding uncertainty data was determined by the signal-to-noise ratio (S/N).[5] $PM_{2.5}$ was set as the total variable and the S/N is set as 'weak' by default. Fifteen species were used in the base runs, but species identified in previous studies used As and Pb as markers in seasonal runs.[2-4]

Table 8.2 shows fifteen species identified by XRF, all of which have 'strong' S/N outcomes.

Table 8.2: Input Data Statistics for positive matrix factorisation modelling.

	Species	Category	S/N
1	**$PM_{2.5}$**	**Weak**	**5.75**
2	BC	Strong	4.39
3	UVPM	Strong	4.32
4	S	Strong	7.51
5	Cl	Strong	6.22
6	K	Strong	8.20
7	Ca	Strong	7.01
8	Ti	Strong	6.23
9	Fe	Strong	7.90
10	Ni	Strong	4.77
11	Cu	Strong	5.07
12	Zn	Strong	7.00
13	Br	Strong	6.66
14	U	Strong	2.60
15	Si	Strong	7.76

Three model runs were completed using different factors; these were mainly determined by literature for previous studies within the area, i.e. five-, six-, and seven-factor runs.[2,4,6-7] The base runs were done at random seeds, and 100 runs were done for each dataset. The error estimates were also set to 100 runs.

8.3. SOURCES OF $PM_{2.5}$ IDENTIFIED BY PMF

The study location heavily influences the process of source identification.[8-10] The specific area of Pretoria, South Africa, where the sampling took place was urban/industrial.

8.3.1. GOODNESS-OF-FIT (Q STATISTIC)

The Q (Robust) value indicates the stability of the run and the lower the value the greater the stability of the run.[9] Meaning, the Q (Robust) values are goodness-of-fit parameters that are highlighted if a model converges. Within the base run, the residuals are also taken into consideration and can be compared at each run.[9] Table 8.2 shows the Q (robust) values at each run. The original $PM_{2.5}$ values and the modelled values were compared, and the correlation between the two is indicated by the R^2 value.

Lastly, the model runs underwent error estimations. After considering all these factors, the seven-factor model run was considered for further analysis.

Table 8.3: The 5, 6 and 7-factor Q values and the run which was decided upon for the positive matrix factorisation model.

2017-2021	Run #	Q(Robust)	Q(True)	Converged	Q(true)/Qexp	Modelled (R^2)
5-factor	38	74938.5	603457	Yes	159.77	0.47
6-factor	3	60545.4	533060	Yes	159.88	0.59
7-factor	**52**	**43627.7**	**195763**	**Yes**	**67.71**	**0.60**

8.3.3. MODEL RESULTS

Table 8.4 shows the possible sources identified when the PMF model outputs were set to for five, six, and seven factors. Figures 8.1 to 8.3 show the mean distributions (concentrations in µg/m^3) of $PM_{2.5}$.

Table 8.4: A summary of the main trace elements in each factor configuration.

5	1	Secondary sulphur	S, U, Si,
	2	Biomass	Cl, K, Br
	3	Vehicular exhaust	Zn, Br, Cl
	4	Resuspended dust	Ni, Ca, Ti, Fe
	5	Mining/Industry	$PM_{2.5}$, BC, UVPM, Fe, Cu, Br, Ca, Ni
6	1	Base metal / Refined oil	U, Ni, Si, Ti, Fe
	2	Secondary sulphur	S, Si
	3	Mining/ Industry	$PM_{2.5}$, BC, UVPM, Fe, Ti, K, Ca
	4	Resuspended dust	Ca, Cu, Ni, Ti
	5	Biomass/ coal burning	Cl, K, Br
	6	General exhaust	Zn, Br, Cl
7	1	Industry/ Base metal	Ni, Cu, Si
	2	Road traffic	$PM_{2.5}$, BC, UVPM, Zn, Br, Cu, Fe
	3	Secondary sulphur	S, U, Si
	4	Mining	$PM_{2.5}$, BC, UVPM, Ti, Fe, Br, U, Cu
	5	Biomass/Coal Burning	K, Br
	6	Resuspended dust	Ca, Ti
	7	General exhaust	Cl, Br, U

After considering all these factors, the 7-factor model run was considered for further analysis.

The seven-factor run produced the relatively better model. Figure 8.4 shows the time-series of the sampled $PM_{2.5}$ concentration levels against the PMF modelled concentrations. The full study concentration mean was 21.8 µg/m^3 (range 0.3 µg/m^3 - 138.9 µg/m^3). The $PM_{2.5}$ concentration levels against the concentration levels modelled by PMF which showed a concentration mean of concentration mean was 17.4 µg/m^3 (0.4 µg/m^3 - 96.3 µg/m^3).

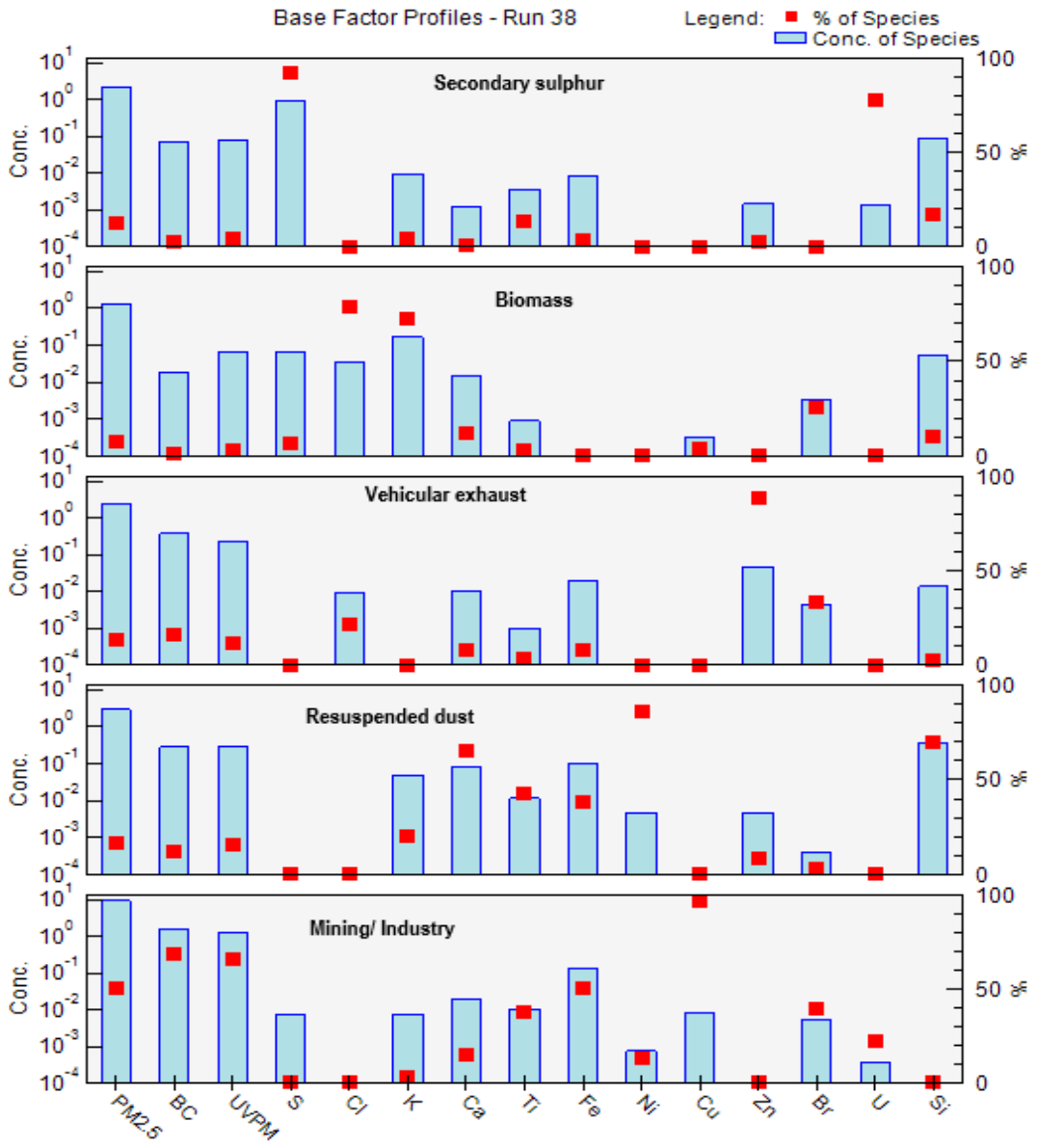


Figure 8.1: The 5-factor Positive matrix factorisation solution for sources and mean contributions (concentrations are in µg/m^3) of $PM_{2.5}$ measured at the School of Health System and Public Health, University of Pretoria between 18 April 2017 and 12 February 2021.

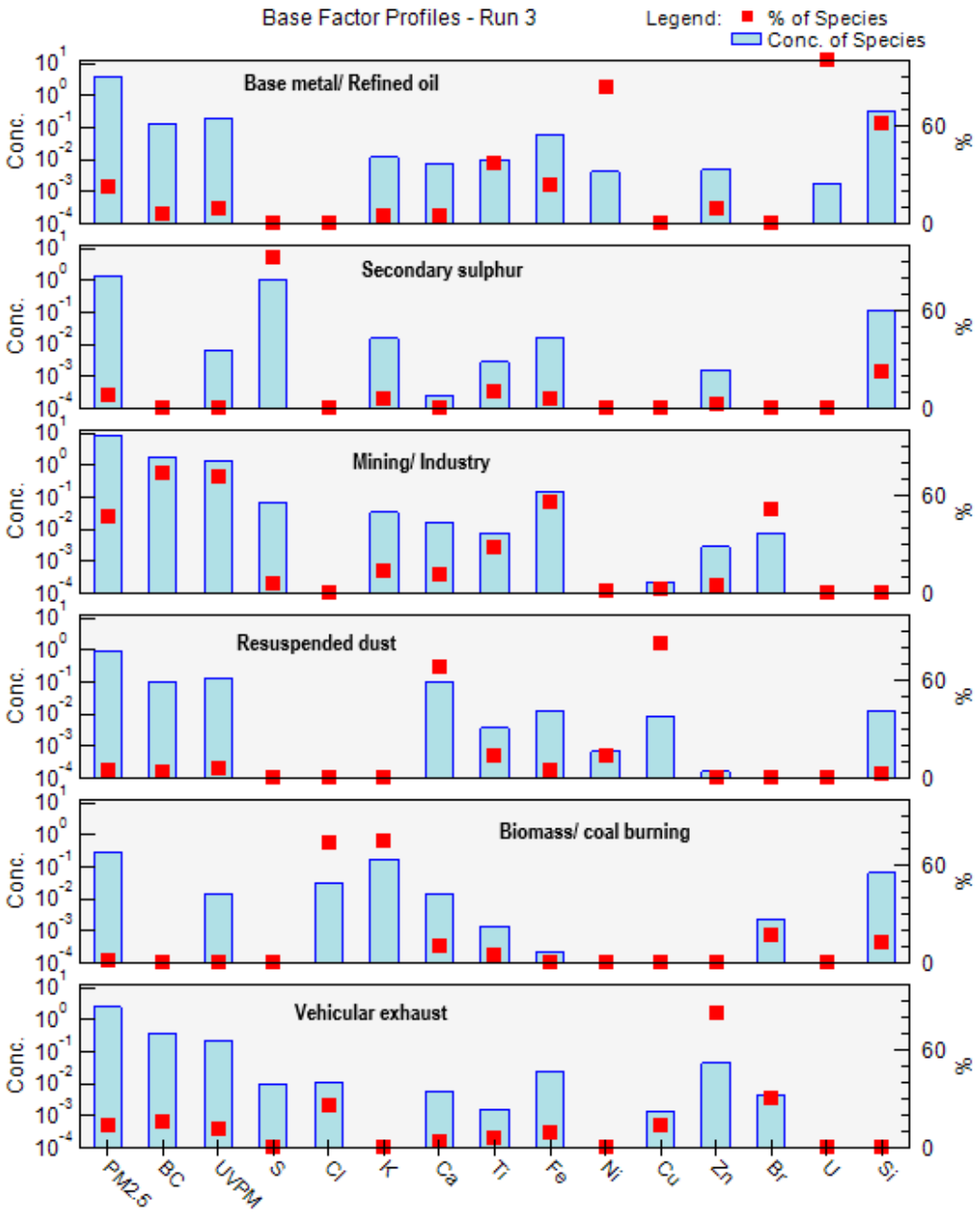


Figure 8.2: The 6-factor Positive matrix factorisation solution for sources and mean contributions (concentrations are in µg/m^3) of $PM_{2.5}$ measured at the School of Health System and Public Health, University of Pretoria between 18 April 2017 and 12 February 2021.

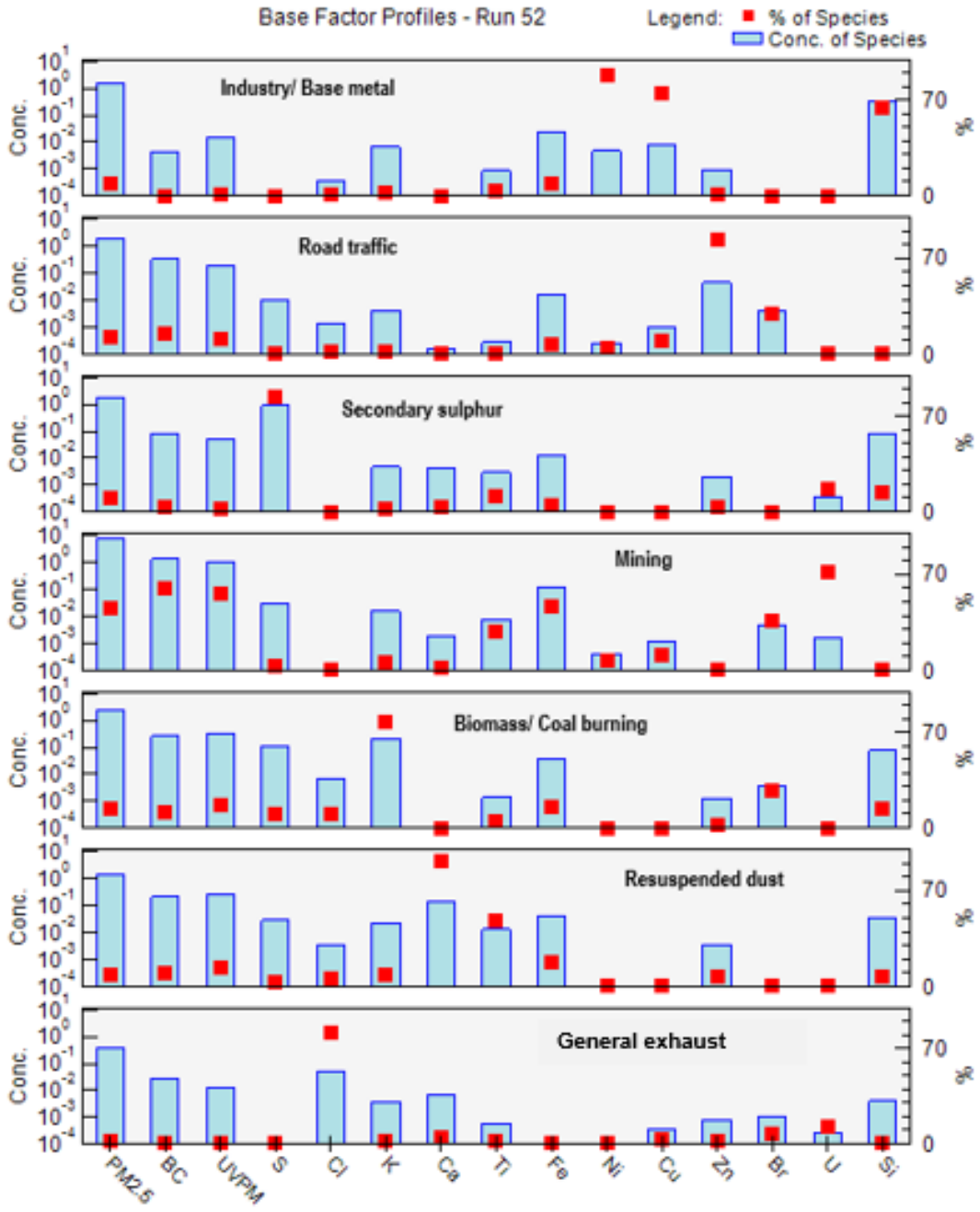


Figure 8.3: The 7-factor Positive matrix factorisation solution for sources and mean contributions (concentrations are in µg/m^3) of $PM_{2.5}$ measured at the School of Health System and Public Health, University of Pretoria between 18 April 2017 and 12 February 2021.

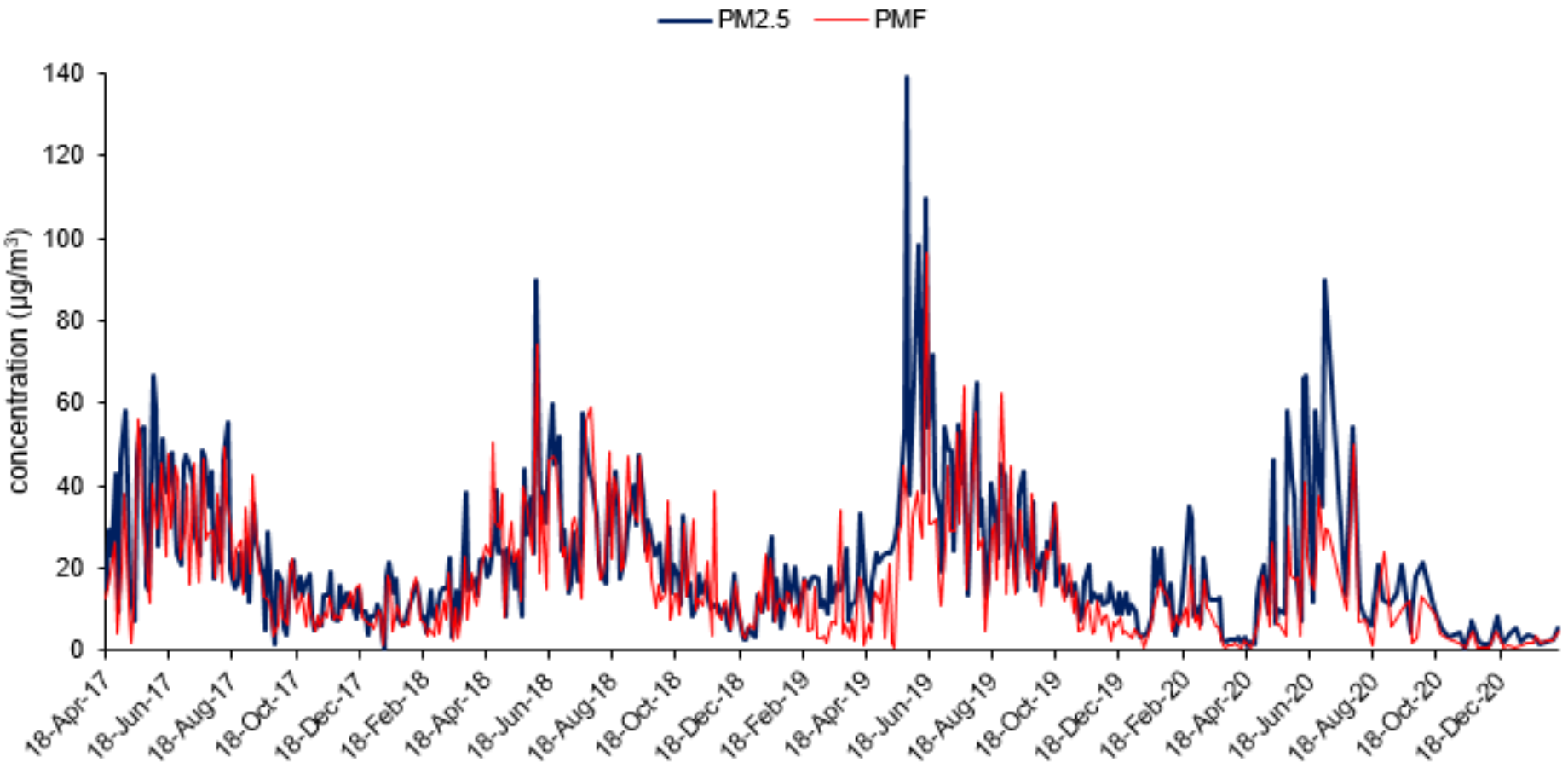


Figure 8.4: Time series of original $PM_{2.5}$ concentrations against concentration levels modelled by PMF from the 7-factor model.

Figure 8.5 and figure 8.6 shows the source contributions and percentages of $PM_{2.5}$. Mining was the largest contribution source of $PM_{2.5}$ at 7.8 µg/m^3 (43.2%). General exhaust was the contributed the least $PM_{2.5}$ at 0.4 µg/m^3 (2%). Figure 8.7 shows the contributions of $PM_{2.5}$ using the seven-factor model. The graphs show the concentration levels as they occurred through the year, each year, demarcated by the orange line.

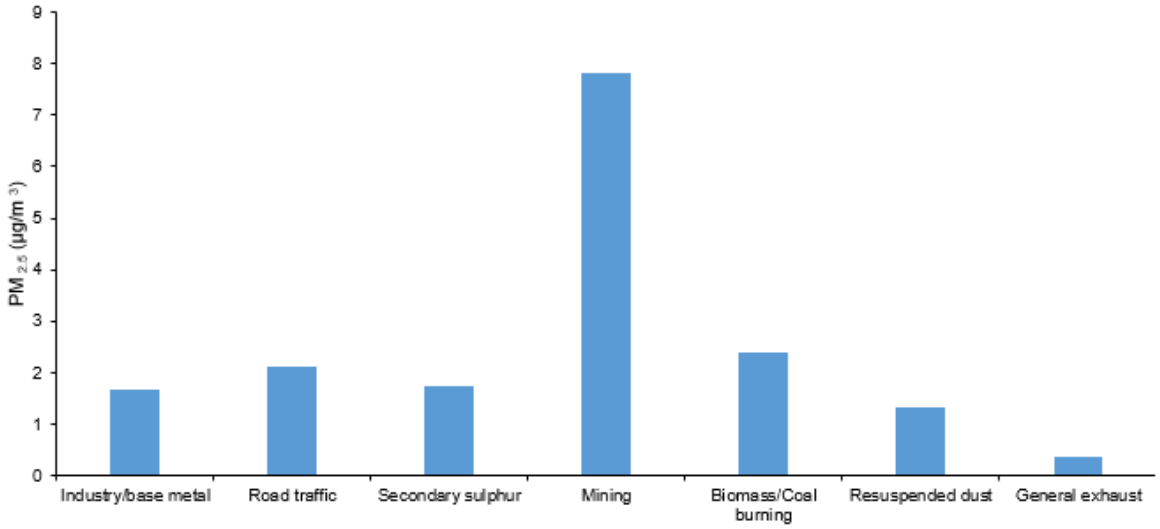


Figure 8.5: Mean source contributions from the 7-factor positive matrix analysis.

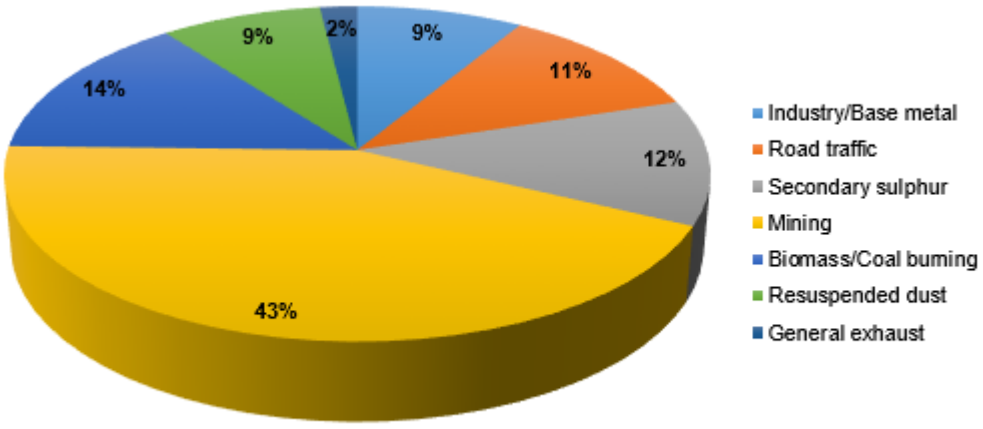


Figure 8.6: Percentage contributions from the 7-factor positive matrix analysis.

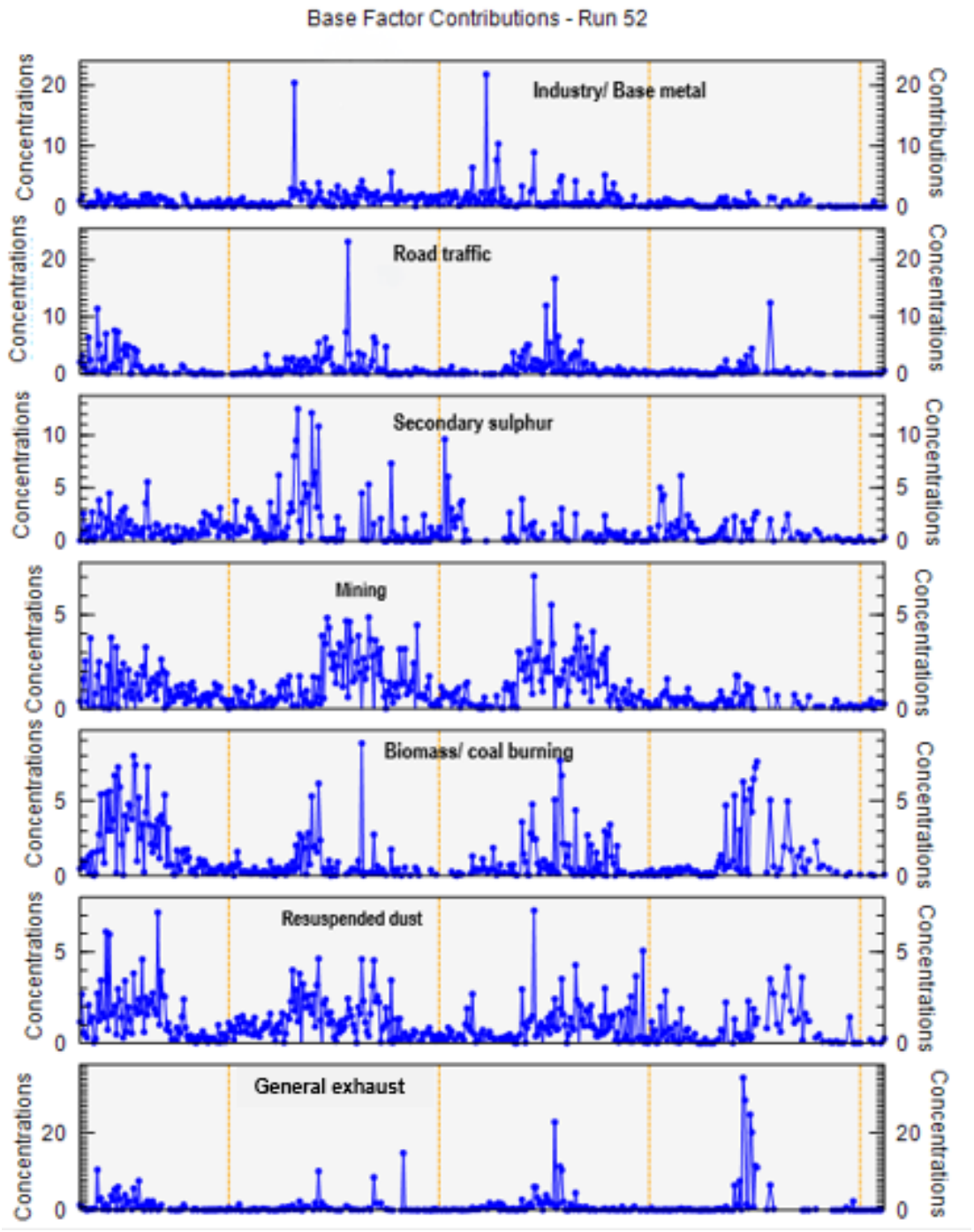


Figure 8.7: Time series factor contributions (in µg/m^3) of $PM_{2.5}$ from the 7-factor PMF analysis.

8.3.2. POSSIBLE FACTORS

FACTOR 1. INDUSTRY/ BASE METAL

Ni, Cu, Si

Low amounts of $PM_{2.5}$ combination was attributed to Industry/base metal pollution, which attributed to 8.7% of the overall $PM_{2.5}$ (1.67 µg/m^3). The factor contained high amounts of Ni (88%), Cu (75%), and Si (65%) over the four-year study period. The high Ni concentrations could indicate that it is a secondary product of manufacturing,[11] as well as an implication of base metal (pyrometallurgy) transportation activities.[2] These are more clearly seen when there is a Ni and U combination, or Ni and V combination. The combination of Ni and Vanadium (V), particularly, are considered markers for oil combustion emitted from shipping or industrial plants.[12] Si also had high contributions, which could be attributed to tail-pipe particulates.[13]

The combination of Ni and Fe (9%) can also be an indication of pyrometallurgical-related factors and base metal processing.[2] Ferrochromium smelters in the Bushveld Igneous Complex in the Gauteng Province could be the reason for the trace amounts of Fe produced.[2]

FACTOR 2. ROAD TRAFFIC

$PM_{2.5}$, BC, UVPM, Zn, Br, Cu, Fe

Road traffic contributed 11.3% of the overall $PM_{2.5}$ concentration (2.1 µg/m^3) over the four-year study period. The combination of Zn, Br, Cu, Fe, and low traces of Cl are attributed to traffic-related sources.[14] This factor contained high contributions of Zn (84%), Cu (9%), and Fe (6%), and a fair contribution of Br (29%). The presence of BC (14%) would suggest combustion was part of the process coupled with the metals, could likely be from road traffic emmisions.[15] Cu and Zn are elemental products associated with tail-pipe emissions.[16] An additional presence of Sb would suggest traffic emission as Cu, Zn, and Sb are traditionally traffic-related elements that are a product of deceleration. Preliminary results used by the Council for Scientific and Industrial Research (CSIR) showed that the strict lockdown measures in South Africa reduced NO_2, SO_2, and O_3 emissions, which could be a result of reduced road use.[17]

FACTOR 3. SECONDARY SULPHUR

S, U, Si

Secondary sulphur contributed 12.1% of the overall $PM_{2.5}$ (1.75 µg/m^3) concentration over the four-year study period. This factor contained high contributions of S (83%), U (16%), and Si (14%). The presence of sulphur is present in multiple forms, including pyrite, metal sulphide smelter gases, coal, and crude oil.[18] There has been considerable pyrite mineralisation in Pretoria, with additional sulphur mining, which causes leeching into U.[18] Other possible sources of S, could include photochemical action and coal burning.[6,19] Additionally, heavy mineral mining from countries like Zambia contributes to high sulphate concentrations over Southern Africa.[2,20-22]

FACTOR 4. MINING

$PM_{2.5}$, BC, UVPM, Ti, Fe, Br, U, Cu

Mining showed to be the highest contributor of the overall $PM_{2.5}$, accounting for 43.2% (7.82 µg/m^3) over the four-year study period. This factor contained high contributions of $PM_{2.5}$ (44.92%), BC (59.42%), UVPM (55.68%), Ti (28.72%), Fe (46.06%), Br (35.83%), U (71.28%), and Cu (10.81%). Figure 8.7 shows the locations of mining areas in South Africa. According to the Department of Mineral Resources, there are over 150 mine sites and related industries located across the country.[23] Approximately five to ten of these sites are located within a 200 km radius of Pretoria.[24-25] Minerals such as gold, diamond, and coal are mined in these areas; common species produced from mining these minerals include Cu, Fe, S and Ca.[24-25]

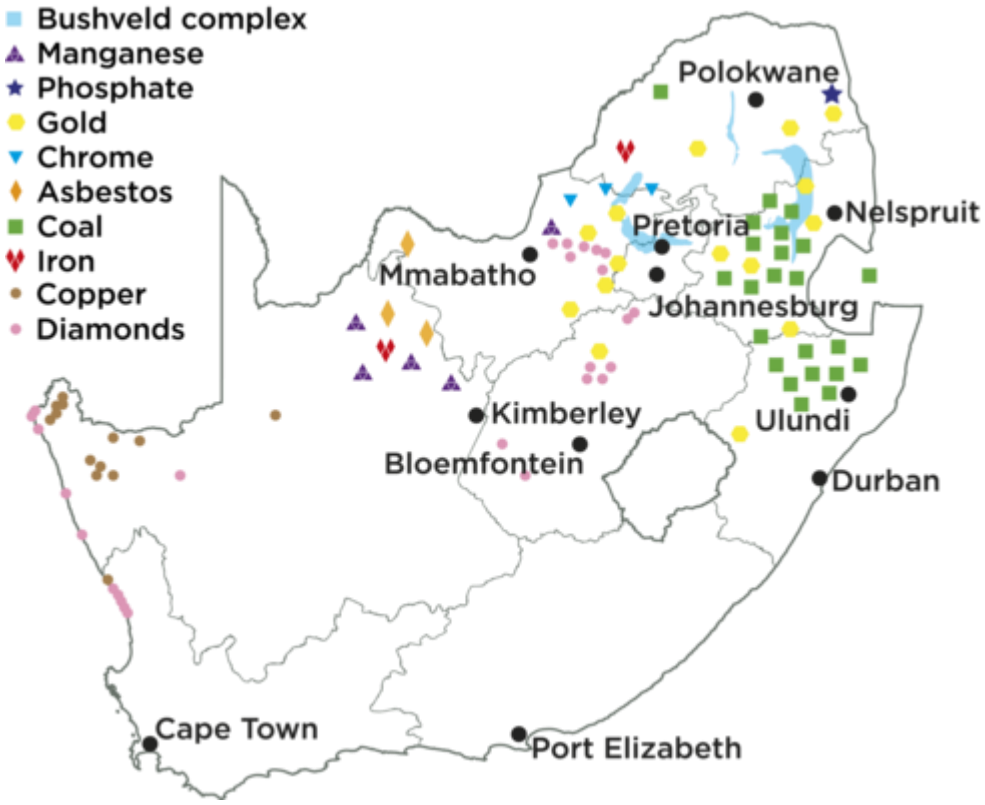


Figure 8.8. Map of Mines in South Africa in proximity to nine main cities in South Africa.[26]

Coal mining is a high producer of hydrocarbons that, in turn, produce levels of BC and UVPM, because of the incomplete combustion.[27] Industrial dust is often highly enriched with Cr, Cu, Pb, and Mn, while coal mining dust often has high V concentrations.[28] Although V was a 'bad' species in this model over the four-year period, the previous 2017-2018 study in the same study area included V as part of the model run.[2] This could suggest that the industrial activity that emitted V has stopped or halted manufacturing after 2018.

FACTOR 5. BIOMASS/ COAL BURNING

K, Br

The contribution of biomass and coal burning to the overall $PM_{2.5}$ was 14.2% (2.4 µg/m3) over the four-year study period. This factor contained high concentrations of K (77.9%), Br (27%), and Fe (15.4%). K is a useful species marker to indicate biomass burning.[2,4,29,30] Lowenthal et al, found that K concentration levels increase in colder months due to the increase in burning activity, as well as changes in meteorological conditions. [31]

FACTOR 6. RESUSPENDED DUST

Ca, Ti, Fe

The resuspended dust factor contributed to 8.5% of the total assessed $PM_{2.5}$ in the PMF model. This factor contained high contributions of Ca (91.2%), Ti (47.7%), and Fe (17.2%), with lower traces of Si (6.5%) and Zn (6.2%). Polluted road-side dust sources ranged from mine ash dumps, waste burning sites, unpaved roads, seasonal grass fires, non-exhaust traffic sources, and industrial sites.[32-33] A large percentage (73%) of the annual contribution of Ca could be attributed to cement kilns[34-35] and natural coarse long-range dust sources.[36] The larger proportion of Ti (44%) can be attributed to both crustal soil sources and mining.[34,37] The Si contribution could be attributed to soil dust,[13] road dust,[38] and tail-pipe particulates.[13]

FACTOR 7. GENERAL EXHAUST

Cl, Br, U

General exhaust fumes contributed to 2.0% of the total assessed $PM_{2.5}$ in the PMF model. This factor contained high contributions of Cl (81.5%), Br (7.4%), and U (12.3%). There were also lower traces of Ca (4.4%), Cu (3.5%), and Zn (1.4%). The combination of Cu and Zn can be attributed to vehicle exhaust and industrial emissions.[4] This combination has been found to come from exhaust fumes specifically stemming from industrial activities.[39-40] However this combination of Cu and Zn can also be associated with lubricating oil in car fuels.[4,13] Trace amounts of Cu can also be a product of the deterioration of tires, clutch, and brake lining in vehicles.[2,7]

8.4. WEEKLY, MONTHLY AND SEASONAL TRENDS OF $PM_{2.5}$ SOURCES IDENTIFIED FROM THE PMF MODEL

Table 8.5 shows the weekly $PM_{2.5}$ levels over the four-year study period. Showing that there was no significant difference in $PM_{2.5}$ concentration levels between the weekends and weekdays across the different identified sources. Figure 8.9 shows the mining to have the highest concentration average through each day of the week.

Table 8.5: The weekly mean $PM_{2.5}$ levels ($\mu g/m^3$) from the 7-factor positive matrix analysis.

	Industry	Fossil fuels	Secondary sulphur	Mining	Coal burning	Resuspended dust	General exhaust
Sunday (N=61)	2.1	1.8	1.7	7.2	2.2	1.1	0.4
Monday (N=59)	1.4	1.9	1.6	7.9	2.7	1.4	0.4
Tuesday (N=62)	1.4	1.7	1.5	7.7	2.1	1.0	0.2
Wednesday(N=61)	**2.2**	2.0	**2.1**	6.7	**2.7**	1.4	**0.6**
Thursday(N=61)	1.8	2.3	1.8	**9.3**	2.5	**1.6**	0.4
Friday (N=61)	1.4	**2.8**	1.8	6.4	2.6	1.4	0.3
Saturday (N=63)	1.5	2.2	1.9	9.4	2.0	1.3	0.3
Kwallis (p-value)	0.96	0.40	0.98	0.84	0.97	0.70	0.85

Bold-Highest average, *-p-value significant

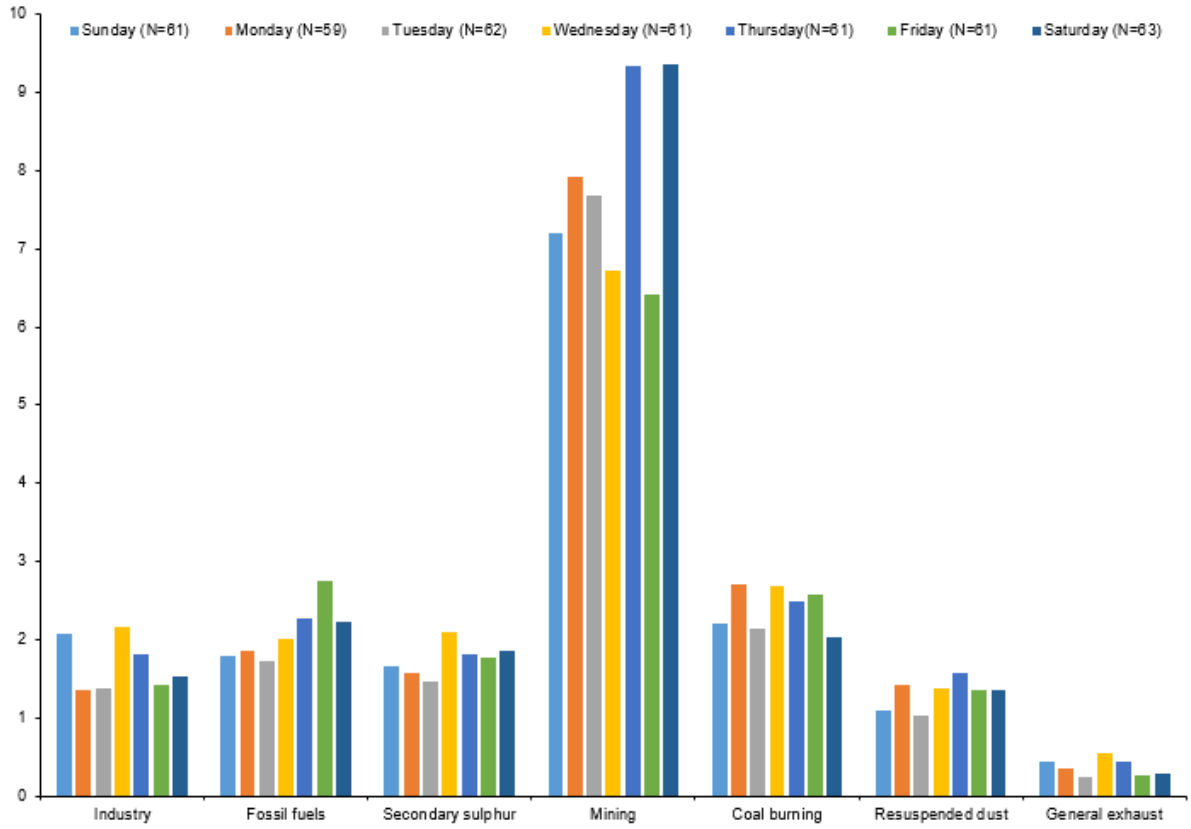


Figure 8.9: The weekly $PM_{2.5}$ concentration levels ($\mu g/m^3$) from the 7-factor positive matrix analysis.

Table 8.6 shows the monthly averages of the different identified sources from the seven-factor PMF model. There was no significant difference in monthly $PM_{2.5}$ concentration levels found for the industrial factor, but significant differences were seen in the other identified sources. There was a significant difference in monthly $PM_{2.5}$ levels for fossil fuels, which showed the lowest average in December (0.2 $\mu g/m^3$) and

the highest average in July (7.7 µg/m^3). Secondary sulphur levels were at the lowest average in September (0.2 µg/m^3) and the highest average in May (3.7 µg/m^3). Mining levels were at the lowest average in January (3.0 µg/m^3) and the highest average in June (14.1 µg/m^3). For coal burning, the lowest average was in December (0.3 µg/m^3) and the highest average in June (5.9 µg/m^3). Lastly, for resuspended dust, the lowest average was in November (0.5 µg/m^3) and the highest average in August (2.6 µg/m^3). Monthly differences were also seen in general exhaust emissions with the lowest average being in November (0.02 µg/m^3) and the highest average in June (1.7 µg/m^3).

Table 8.6: The monthly mean $PM_{2.5}$ levels (µg/m^3) from the 7-factor positive matrix analysis.

	Industry	Fossil fuels	Secondary sulphur	Mining	Coal burning	Resuspended dust	General exhaust
January (N=38)	1.2	0.3	2.6	3.0	0.5	0.7	0.04
February (N=31)	1.3	0.4	2.4	3.9	0.5	0.8	0.04
March (N=29)	2.4	0.6	1.5	2.5	0.8	0.6	0.1
April (N=35)	**2.9**	1.4	1.8	3.6	0.9	0.9	0.2
May (N=42)	1.5	3.5	**3.7**	7.1	3.4	1.5	0.3
June (N=39)	1.9	5.1	1.9	**14.1**	**5.9**	2.1	**1.7**
July (N=35)	1.5	**7.7**	1.3	13.5	5.8	1.8	1.1
August (N=36)	2.0	2.2	1.5	12.9	4.7	**2.6**	0.3
September (N=36)	1.4	1.9	0.8	12.8	2.4	1.9	0.2
October (N=36)	1.9	0.8	1.2	8.6	2.1	1.1	0.2
November (N=35)	1.0	0.3	0.9	6.2	0.7	0.5	0.02
December (N=36)	0.9	0.2	1.2	3.8	0.3	0.7	0.1
Kwallis (p-value)	0.53	0.0001*	0.0003*	0.0001*	0.0001*	0.0001*	0.001*

Bold-Highest average, *-p-value significant

Generally there were higher concentrations of $PM_{2.5}$ from most of the identified sources in colder months, i.e. May to August, and lower concentrations in warmer months, i.e. September to January. Figure 8.10 shows the distribution for each month, where it is clearer to see Mining is continuously high across all months. However, the $PM_{2.5}$ concentration levels seem to be higher from June to September.

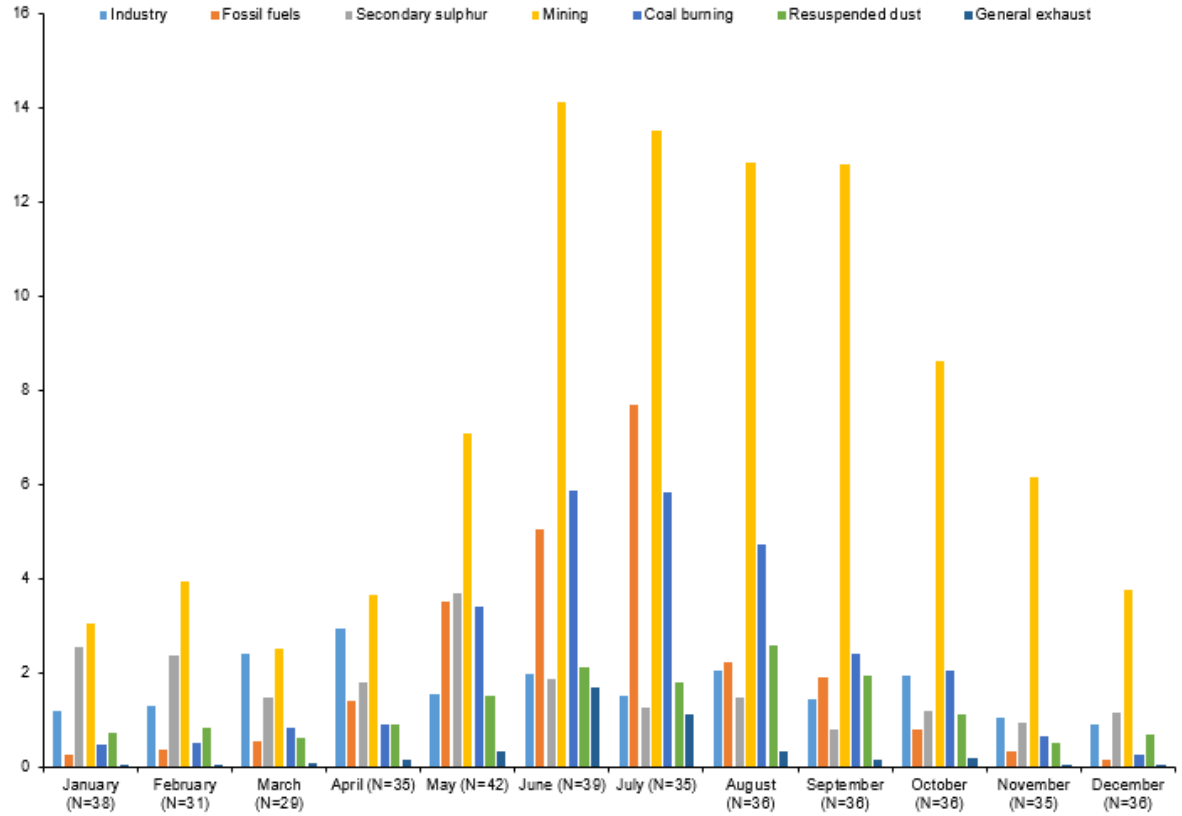


Figure 8.10: The monthly $PM_{2.5}$ concentration levels (µg/m^3) from the 7-factor positive matrix analysis.

Table 8.7 shows the seasonal $PM_{2.5}$ concentration levels for the four-year study period. Industry did not show seasonal differences in $PM_{2.5}$ levels, which corresponds with the monthly findings. This could suggest that there are particular industries that continuously run throughout the year. However, the highest average $PM_{2.5}$ concentration was seen in April, i.e. autumn. Fossil fuels (4.9µg/m^3), mining (13.5 µg/m^3), coal burning (5.5 µg/m^3), resuspended dust (2.2 µg/m^3), and general exhaust (1.1 µg/m^3) all showed significantly higher concentration levels in winter. This could be because of the temperature inversion that occurs in colder months, which does not release air pollutants easily out from the atmosphere.[41] Additionally, activities such as mining and coal burning significantly increase in winter when there is a higher demand for heating within homes and industries in South Africa.[4] The findings are consistent with previous studies conducted within Pretoria.[2,7] The findings of this project are also consistent with areas in the Vaal and Highveld that experience similar seasonal patterns as Pretoria, when higher $PM_{2.5}$ levels are experienced in late winter and early spring.[42-44] During this period activities such as domestic burning is at its highest.[43]

However, in Cape Town, a coastal city, and Thohoyandou, a largely rural town, $PM_{2.5}$ levels were in higher spring, not winter.[7,45]

Table 8.7: The seasonal mean $PM_{2.5}$ levels (µg/m^3) from the 7-factor positive matrix analysis.

	Industry	Fossil fuels	Secondary sulphur	Mining	Coal burning	Resuspended dust	General exhaust
Autumn (N = 106)	**2.2**	2.0	**2.5**	4.7	1.9	1.1	0.2
Winter (N = 110)	1.9	**4.9**	1.6	**13.5**	**5.5**	**2.2**	**1.1**
Spring (N= 107)	1.5	1.0	1.0	9.2	1.7	1.2	0.1
Summer (N = 105)	1.1	0.3	2.0	3.6	0.4	0.8	0.04
Kwallis (p-value)	0.09	0.0001*	0.002*	0.0001*	0.0001*	0.0001*	0.0001*

Bold-Highest average, *-p-value significant

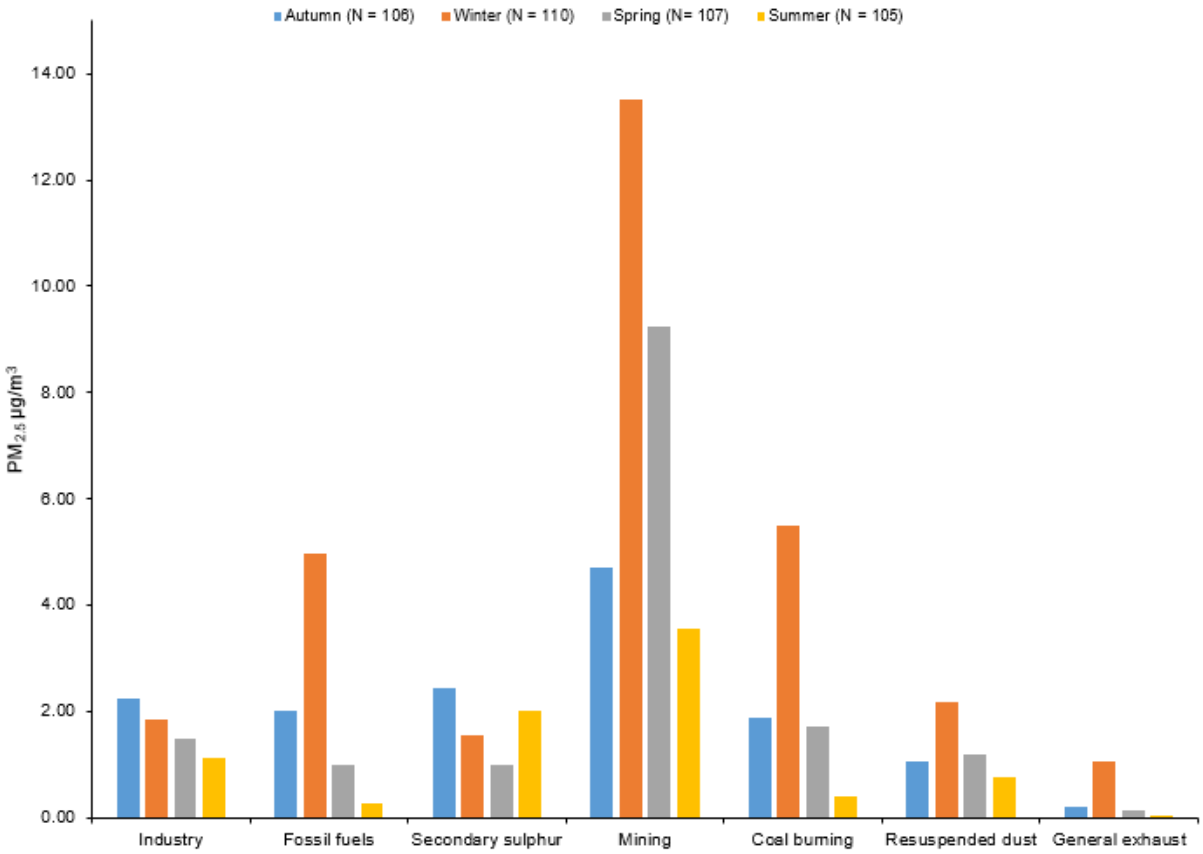


Figure 8.11: The seasonal $PM_{2.5}$ concentration levels (µg/m^3) from the 7-factor positive matrix analysis.

8.5. DISCUSSION

Fine particulate matter ($PM_{2.5}$) has been identified as a major environmental health risk that is highly associated with aggravating or worsening disease.[46-51] A case-crossover study conducted in Pretoria, South Africa showed an increased risk in

respiratory hospitalisation from exposure to $PM_{2.5}$ and $PM_{2.5}$-bound trace elements.[52] Thus, source apportionment studies are important because they assist in identifying contributing sources of $PM_{2.5}$ and influence policy to reduce emissions.[53-54] The seven-factor model suggested that the main sources and contributions to the total $PM_{2.5}$ were industry/base metal (8.7%), road traffic (11.3%), secondary sulphur (12.1%), mining (43.2%), biomass/coal burning (14.2%), resuspended dust (8.5%), and general exhaust (2.0%). These results are similar to those obtained in two previous studies of the same area for 2017 to 2018,[2] and 2017 to 2020.[4] However, in a study by McDuffie et al, coal combustion was estimated to contribute 20.5% and 26.1% of total $PM_{2.5}$ pollution in South Africa and Johannesburg, respectively. [55] Unlike the previous studies, HYSPLIT trajectory was not done, since it was not the focus of this study. This part of the study determined the number of pollution sources over a particular duration of time, using a popular method of source apportionment.[8-9,15,56-59]

The sampling site for $PM_{2.5}$ may have been in an urban background, however, the source apportionment through PMF showed a variety of potential sources. PMF had modelled an average $PM_{2.5}$ concentration of 17.4 µg/m^{3,} compared to the measured $PM_{2.5}$ sample average of 21.8 µg/m^3. The actual measured $PM_{2.5}$ average was close to an estimated $PM_{2.5}$ average for South Africa, which was 28.8 µg/m^3.[55] However, the modelled $PM_{2.5}$ was much lower than the estimated $PM_{2.5}$ in Johannesburg which was 40.2 µg/m^3.[55] Although the variations in the modelled average concentration levels $PM_{2.5}$ and the actual measured $PM_{2.5}$ do not largely vary they provide a good baseline of the average anticipated $PM_{2.5}$ level in areas of the country that are of similar background. There are limited studies that focus on the collection of $PM_{2.5}$ in Africa, and so, the comparisons may not be as representative of the different African countries.

Majority of the source apportionment studies in South Africa have been conducted in urban and rural backgrounds with few conducted in industry backgrounds.[60] $PM_{2.5}$ concentration levels in the administrative capital of Pretoria have recorded levels lower than the NAAQS standard (40 µg/m^3), but are generally higher than the WHO guideline.[2-4]. The findings in this project show higher mean $PM_{2.5}$ levels in an urban background compared to the mean $PM_{2.5}$ levels recorded in the coastal city of Cape Town (13.3 µg/m^3) [45] and rural non-industrial town of Thohoyandou (10.9 µg/m^3).[7].

However, in national priority areas such as the Vaal Triangle Airshed Priority Area (VTAPA), South Africa, higher mean $PM_{2.5}$ levels were recorded (30.4 µg/m^3).[61] The measurements from the latter were higher than recorded by the South African government at twenty-one air monitoring stations from five provinces in the country.[62] The difference in background shows that the increased sampling in different areas of the country can better inform specific air quality management plans for the different areas in South Africa. In Africa ambient $PM_{2.5}$ levels are only available in 22 of 54 countries, and all exceed the WHO daily guideline.[63] Thus more sampling across the continent can also provide more specific mean concentration levels of $PM_{2.5,}$ better equipping the air quality management in the different countries.

The project briefly outlines the BC levels obtained in the XRF results. $PM_{2.5}$ and BC have a strong correlation; for example higher $PM_{2.5}$ levels in the VTAPA, recorded higher BC levels (3.2 µg/m^3)[61], compared to Thohoyandou (1.3 µg/m^3) [7] which recorded lower $PM_{2.5}$ levels. This project recorded a mean BC level of 2.8 µg/m^3, which is similar to an average BC of 2.7 µg/m^3 recorded by a previous study conducted in Pretoria.[4] A Kenyan study conducted in Nairobi recorded a mean BC level of 2.7 µg/m^3.[64] Benin and South Africa recorded the highest and lowest estimated mean BC level in Africa, at 16 µg/m^3 and 2.1 µg/m^3, respectively.[65] The latter show consistent mean BC levels found in this project that probably reflect samples taken from urban backgrounds.

The correlation between meteorological conditions like relative humidity, wind speed, and temperature have an influence on the air pollution of an area.[43-44] Studies have shown that meteorological conditions such as wind speed, wind direction, rainfall, temperature, and relative humidity can influence the trajectory of sources onto a sampling area.[2,4,7,45] In the one-year and three-year studies done in Pretoria prior to this project, a range of four to five cluster trajectories showed to influence the identified sources.[2-3] Long-range and local influences are indeed important to acknowledge during source apportionment studies. Pretoria is located in close proximity to mining and industrial areas at both long- and local-range.[4] Results from this project showed that industry and mining sectors contributed over 50% of the $PM_{2.5}$ during the study period.

In the northern landlocked areas of South Africa which includes Pretoria, these areas generally experience dry winters and wet summers, meaning the colder months record higher concentrations of air pollution.[66-69] The movement of air disperses air particles such as $PM_{2.5}$ and can reduce the particles in the atmosphere.[70] The shift in temperature, wind speed, relative humidity, and rainfall can increase and decrease the concentrations of particulate matter.[71-72] Additionally, stronger inversion can be created through low wind speeds, and low to no precipitation.[67,71,73-74] During warmer months, air pollutants are released easier into the troposphere.[68] This literature corresponds with the findings of this project, which shows that concentrations of $PM_{2.5}$ concentrations were higher in the colder months and lower in warmer months. Majority of South Africa's local emission comes from traffic emissions, power generation by coal, industrial emissions, wood burning, and paraffin use.[75-77] With some activities increasing during winter. The long-range influence could be contributed from the national designated priority areas, which have exceedingly high air pollution levels from high industrial production, domestic fuel burning, mining, and waste burning.[78-80]

8.6. CONCLUSION

Source apportionment using PMF analyses showed that mining and industry are the main contributing factors to $PM_{2.5}$ in Pretoria. There is a great need for more studies that sample $PM_{2.5}$ in the Africa. Source apportionment studies are vital in the evaluation of policy to protect communities from the detrimental health effects of $PM_{2.5}$. The running and interpretation of model results from the PMF software was relatively easy. Source identification may seem rather subjective, but the results were similar to previous studies conducted in the same area. The main limitation of the process was that the three model runs only showed 0.4-0.6 correlation with the original data.

8.7. REFERENCES

1. World Health Organization. WHO global air quality guidelines. Particulate matter (PM2.5 and PM10), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide. 2021 978-92-4-003422-8.

2. Adeyemi A, Molnar P, Boman J, Wichmann J. Source apportionment of fine atmospheric particles using positive matrix factorization in Pretoria, South Africa. Environmental Monitoring and Assessment. 2021; 193(11):716. doi:10.1007/s10661-021-09483-3.

3. Howlett-Downing C. The association between sources of air pollution and respiratory health in Pretoria, South Africa: University of Pretoria; 2022.

4. Howlett-Downing C, Boman J, Molnár P, Shirinde J, Wichmann J. PM2.5 chemical composition and geographical origin of air masses in Pretoria, South Africa. Water, Air, & Soil Pollution. 2022; 233(7) doi:10.1007/s11270-022-05746-y.

5. Venter AD, Vakkar V, Beukes JP, van Zyl PG, Laakso H, Mabaso D, et al. An air quality assessment in the industrialised western Bushveld Igneous Complex, South Africa. South African Journal of Science. 108(9/10) doi:10.4102/sajs.v108i9/10.1059.

6. Molnár P, Tang L, Sjöberg K, Wichmann J. Long-range transport clusters and positive matrix factorization source apportionment for investigating transboundary PM2.5 in Gothenburg, Sweden. Environmental Science: Processes & Impacts. 2017; 19(10):1270-7. doi:10.1039/c7em00122c.

7. Novela RJ, Gitari WM, Chikoore H, Molnar P, Mudzielwana R, Wichmann J. Chemical characterization of fine particulate matter, source apportionment and long-range transport clusters in Thohoyandou, South Africa. Clean Air Journal; 30(2):1-2. doi:10.17159/caj/2020/30/2.8735.

8. Jianfei C, Chunfang L, Lixia Z, Quanyuan W, Jianshu L, Rodríguez-Seijo As. Source apportionment of potentially toxic elements in soils using apcs/mlr, PMF and geostatistics in a typical industrial and mining city in eastern China. Plos one. 2020; 15(9):e0238513. doi:10.1371/journal.pone.0238513.

9. Norris G DR, Brown S, Bai S. EPA positive matrix factorization (PMF) 5.0 fundamentals and user guide, US environmental protection agency. Washington, DC; 2014.

10. Zíková Nzd, Wang Y, Yang F, Li X, Tian M, Hopke PK. On the source contribution to Beijing PM2.5 concentrations. Atmospheric Environment. 2016; 134:84-95. doi:10.1016/j.atmosenv.2016.03.047.

11. Srivastava D, Goel A, Agrawal M. Particle bound metals at major intersections in an urban location and source identification through use of metal markers. Proceedings of the National Academy of Sciences, India Section A: Physical Sciences. 2016; 86(2):209-20. doi:10.1007/s40010-016-0268-y.

12. Iijima A, Saito Y, Kozawa K, Furuta N, Sato K, Fujitani Y, et al. Clarification of the predominant emission sources of antimony in airborne particulate matter and estimation of their effects on the atmosphere in Japan. Environmental Chemistry. 2009; 6(2):122-32. doi:10.1071/EN08107.

13. Yu L, Wang G, Zhang R, Zhang L, Song Y, Wu B, et al. Characterization and source apportionment of PM2.5 in an urban environment in Beijing. Aerosol and Air Quality Research. 2013; 13(2):574-83. doi:10.4209/aaqr.2012.07.0192.

14. Huang F, Zhou J, Chen N, Li Y, Li K, Wu S. Chemical characteristics and source apportionment of PM2.5 in Wuhan, China. Journal of Atmospheric Chemistry. 2019; 76(3):245-62. doi:10.1007/s10874-019-09395-0.

15. Mooibroek D, Schaap M, Weijers EP, Hoogerbrugge R. Source apportionment and spatial variability of PM2.5 using measurements at five sites in the netherlands. Atmospheric Environment. 2011; 45(25):4180-91. doi:https://doi.org/10.1016/j.atmosenv.2011.05.017.

16. Leelőssy Á, Molnár F, Izsák F, Havasi Á, Lagzi I, Mészáros R. Dispersion modeling of air pollutants in the atmosphere: A review. Open Geosciences. 2014; 6(3):257-78.

17. Mandaha D. Satellites show decrease in air pollution in South Africa during national lockdown CSIR,; 2020. Available from: https://www.csir.co.za/air-pollution-decrease-south-africa-during-national-lockdown.

18. Department of Mineral Resources. Review of the sulphur industry in the republic of south Africa, 2012. 2013.

19. Kim E, Hopke PK, Edgerton ES. Source identification of atlanta aerosol by positive matrix factorization. Journal of the Air & Waste Management Association. 2003; 53(6):731-9.

20. Kríbek B, Majer V, Veselovský F, Nyambe I. Discrimination of lithogenic and anthropogenic sources of metals and sulphur in soils of the central-northern part of the Zambian Copperbelt Mining District: A topsoil vs. Subsurface soil concept. Journal of Geochemical Exploration. 2010; 104(3):69-86.

21. Vojtěch E, Vítková M, Mihaljevič M, Šebek Oe, Klementová M, Veselovský Fe, et al. Dust from Zambian smelters: Mineralogy and contaminant bioaccessibility. Environmental Geochemistry and Health. 2014; 36(5):919-33. doi:10.1007/s10653-014-9609-4.

22. Macdonald R. Theory and objectives of air dispersion modelling. Modelling Air Emissions For Compliance. 2003.

23. Department of Mineral Resources and Energy. Operating mines in Gauteng,. 2022. Available from: https://www.dmr.gov.za/mineral-policy-promotion/operating-mines/gauteng.

24. Cairncross B, Dixon R. Minerals of South Africa. The geological society of South Africa. 1995.

25. Cairncross B. Field guide to rocks & minerals of Southern Africa. 2004.

26. Utembe W, Faustman E, Matatiele P, Gulumiam M. Hazards identified and the need for health risk assessment in the South African mining industry. Human & Experimental Toxicology. 2015; 34:1212-21. doi:10.1177/0960327115600370.

27. Wang S, Liu G, Yuan Z, Liu Y, Lam PK. Spatial variability and source apportionment of aliphatic hydrocarbons in sediments from the typical coal mining area. Bulletin of Environmental Contamination and Toxicology. 2020; 105(2):230-6.

28. Das A, Kumar R, Patel SS, Saha MC, Guha D. Source apportionment of potentially toxic elements in street dust of a coal mining area in Chhattisgarh, India, using multivariate and lead isotopic ratio analysis. Environmental Monitoring and Assessment. 2020; 192(6):1-14.

29. Luo L, Zhang Y-Y, Xiao H-Y, Xiao H-W, Zheng N-J, Zhang Z-Y, et al. Spatial distributions and sources of inorganic chlorine in PM2.5 across China in winter. Atmosphere. 2019; 10(9):505. doi:10.3390/atmos10090505.

30. Chow JC, Watson JG, Fujita EM, Lu Z, Lawson DR, Ashbaugh LL. Temporal and spatial variations of PM2.5 and PM10 aerosol in the southern California air quality study. Atmospheric Environment. 1994; 28(12):2061-80.

31. Lowenthal D, Gertler A, Labib M. Particulate matter source apportionment in cairo: Recent measurements and comparison with previous studies. International Journal of Environmental Science and Technology. 2014; 11(3):657-70.

32. Thabethe NDL, Engelbrecht JC, Wright CY, Oosthuizen MA. Human health risks posed by exposure to PM10 for four life stages in a low socio-economic community in South Africa. The Pan African Medical Journal. 2014; 18:206. doi:10.11604/pamj.2014.18.206.3393.

33. Belis CA, Karagulian F, Larsen BR, Hopke PK. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. Atmospheric Environment. 2013; 69:94-108. doi:10.1016/j.atmosenv.2012.11.009.

34. Kim E, Hopke PK, Edgerton ES. Source identification of atlanta aerosol by positive matrix factorization. Journal of the Air & Waste Management Association. 2003; 53(6):731-9.

35. Qi L, Zhang Y, Ma Y, Chen M, Ge X, Ma Y, et al. Source identification of trace elements in the atmosphere during the second asian youth games in Nanjing, China: Influence of control measures on air quality. Atmospheric Pollution Research. 2016; 7(3):547-56. doi:10.1016/j.apr.2016.01.003.

36. Thurston GD, Ito K, Lall R. A source apportionment of US fine particulate matter air pollution. Atmospheric Environment. 2011; 45(24):3924-36.

37. Mancilla Y, Hernandez Paniagua I, Mendoza A. Spatial differences in ambient coarse and fine particles in the Monterrey metropolitan area, Mexico: Implications for source contribution. Journal of the Air & Waste Management Association. 2019; 69(5):548-64.

38. Thorpe A, Harrison RM. Sources and properties of non-exhaust particulate matter from road traffic: A review. Science of the Total Environment. 2008; 400(1-3):270-82.

39. Negi B, Sadasivan S, Mishra U. Aerosol composition and sources in urban areas in India. Atmospheric Environment 1967; 21(6):1259-66.

40. Schwarz J, Cusack M, Karban Ji, Chalupníčková E, Havránek Vr, Smolík Ji, et al. PM2.5 chemical composition at a rural background site in central Europe, including correlation and air mass back trajectory analysis. Atmospheric Research. 2016; 176-177:108-20. doi:10.1016/j.atmosres.2016.02.017.

41. Sager L. Estimating the effect of air pollution on road safety using atmospheric temperature inversions. Journal of Environmental Economics and Management. 2019; 98:102250. doi:https://doi.org/10.1016/j.jeem.2019.102250.

42. Govender K, Sivakumar V. A decadal analysis of particulate matter (pm2. 5) and surface ozone (O3) over Vaal Priority Area, South Africa. Clean Air Journal. 2019; 29(2).

43. Hersey SP, Garland RM, Crosbie E, Shingler T, Sorooshian A, Piketh S, et al. An overview of regional and local characteristics of aerosols in South Africa using satellite, ground, and modeling data. Atmospheric Chemistry and Physics. 2015; 8:4259-78.

44. Muyemeki L, Burger R, Piketh SJ, Beukes JP, Van Zyl PG. Source apportionment of ambient PM10-25 and PM2.5 for the Vaal Triangle, South Africa. South African Journal of Science. 2021; 117(5-6):1-11.

45. Williams J, Petrik L, Wichmann J. PM2.5 chemical composition and geographical origin of air masses in Cape Town, South Africa. Air Quality, Atmosphere & Health. 2020; 14(3):431-42. doi:10.1007/s11869-020-00947-y.

46. Abdolahnejad A, Jafari N, Mohammadi A, Miri M, Hajizadeh Y, Nikoonahad A. Cardiovascular, respiratory, and total mortality ascribed to PM10 and PM2.5 exposure in Isfahan, Iran. Journal of Education and Health Promotion. 2017; 6:109. doi:10.4103/jehp.jehp_166_16.

47. Barnes B, Mathee A, E T, Bruce N. Household energy, indoor air pollution and child respiratory health in South Africa. Journal of Energy in Southern Africa. 2017; 20(1):4-13. doi:10.17159/2413-3051/2009/v20i1a3296.

48. Capraz O, Deniz A, Dogan N. Effects of air pollution on respiratory hospital admissions in Istanbul, Turkey, 2013 to 2015. Chemosphere. 2017; 181:544-50. doi:10.1016/j.chemosphere.2017.04.105.

49. Li P, Wu J, Wang R, Liu H, Zhu T, Xue T. Source sectors underlying pm2. 5-related deaths among children under 5 years of age in 17 low-and middle-income countries. Environment International. 2023:107756.

50. Lu F, Xu D, Cheng Y, Dong S, Guo C, Jiang X, et al. Systematic review and meta-analysis of the adverse health effects of ambient PM2.5 and PM10 pollution in the Chinese population. Environmental Research. 2015; 136:196-204. doi:10.1016/j.envres.2014.06.029.

51. Madaniyazi L, Xerxes S. Outdoor air pollution and the onset and exacerbation of asthma. Chronic Diseases and Translational Medicine. 2021; 7(2):100-6. doi:10.1016/j.cdtm.2021.04.003.

52. Howlett-Downing C, Boman J, Molnár P, Shirinde J, Wichmann J. Case-crossover study for the association between increased hospital admissions for respiratory diseases and the increase in atmospheric PM2.5 and PM2.5-bound trace elements in Pretoria, South Africa. International Journal of Environmental Health Research. 2023:1-15.

53. Chen M, Wang P, Chen Q, Wu J, Chen X. A clustering algorithm for sample data based on environmental pollution characteristics. Atmospheric Environment. 2015; 107:194-203. doi:10.1016/j.atmosenv.2015.02.042.

54. Hopke PK. Review of receptor modeling methods for source apportionment. Journal of the Air & Waste Management Association. 2016; 66(3):237-59. doi:10.1080/10962247.2016.1140693.

55. McDuffie EE, Martin RV, Spadaro JV, Burnett R, Smith SJ, O'Rourke P, et al. Source sector and fuel contributions to ambient pm2. 5 and attributable mortality across multiple spatial scales. Nature Communications. 2021; 12(1):3594.

56. Ashrafi K, Fallah R, Hadei M, Yarahmadi M, Shahsavani A. Source apportionment of total suspended particles (TSP) by positive matrix factorization (PMF) and chemical mass balance (CMB) modeling in Ahvaz, Iran. Archives of Environmental Contamination and Toxicology. 2018; 75(2):278-94. doi:10.1007/s00244-017-0500-z.

57. Han F, Kota SH, Wang Y, Zhang H. Source apportionment of PM2.5 in Baton Rouge, Louisiana during 2009–2014. Science of The Total Environment. 2017; 586:115-26. doi:https://doi.org/10.1016/j.scitotenv.2017.01.189.

58. Kim S, Kim TY, Yi SM, Heo J. Source apportionment of PM2.5 using positive matrix factorization (PMF) at a rural site in Korea. Journal of Environmental Management. 2018; 214:325-34. doi:10.1016/j.jenvman.2018.03.027.

59. Molnár P, Sallsten G. Contribution to PM(2.5) from domestic wood burning in a small community in Sweden. Environmental Science: Processes & Impacts. 2013; 15(4):833-8. doi:10.1039/c3em30864b.

60. Mathuthu M, Dudu VP, Manjoro M. Source apportionment of air particulates in South Africa: A review. Atmospheric and Climate Sciences. 2018; 9(1):100-13.

61. Mwase N, Olutola B, Wichmann J. Temperature modifies the association between air pollution and respiratory disease hospital admissions in an industrial area of South Africa: The Vaal Triangle Air Pollution Priority Area. Clean Air Journal. 2022; 32(2). doi:10.17159.caj.2022.32.2.14588.

62. Altieri KE, Keen SL. Public health benefits of reducing exposure to ambient fine particulate matter in South Africa. Science of the Total Environment. 2019; 684:610-20.

63. Agbo KE, Walgraeve C, Eze JI, Ugwoke PE, Ukoha PO, Van Langenhove H. A review on ambient and indoor air pollution status in Africa. Atmospheric Pollution Research. 2021; 12(2):243-60.

64. Gaita SM, Boman J, Pettersson JBC, Gatari MJ, Janhall S. Source apportionment and seasonal variation of PM2.5 in a sub-sahara african city: Nairobi, kenya. Atmospheric Chemistry and Physics Discussions. 2014; 14(7):9565-601. doi:10.5194/acpd-14-9565-2014.

65. Bachwenkizi J, Liu C, Meng X, Zhang L, Wang W, van Donkelaar A, et al. Fine particulate matter constituents and infant mortality in Africa: A multicountry study. Environment International. 2021; 156:106739.

66. Boffetta P, La Vecchia C, Moolgavkar S. Chronic effects of air pollution are probably overestimated. Risk Analysis. 2015; 35(5):766-9. doi:10.1111/risa.12320.

67. Hsu W, Hwang S, Kinney PL, Lin S. Seasonal and temperature modifications of the association between fine particulate air pollution and cardiovascular hospitalization in New York State. Science of the Total Environment. 2017; 578:626-36.

68. Jacob DJ, Winner DA. Effect of climate change on air quality. Atmospheric Environment. 2009; 43(1):51-63.

69. Zhang R, Jing J, Tao J, Hsu S-C, Wang G, Cao J, et al. Chemical characterization and source apportionment of PM2.5 in Beijing: Seasonal perspective. Atmospheric Chemistry and Physics. 2013; 13(14):7053-74.

70. Chen Y, Schleicher N, Fricker M, Cen K, Liu X, Kaminski U, et al. Long-term variation of black carbon and PM2.5 in Beijing, China with respect to meteorological conditions and governmental measures. Environmental Pollution. 2016; 212:269-78. doi:10.1016/j.envpol.2016.01.008.

71. Wang J, Ogawa S. Effects of meteorological conditions on PM2.5 concentrations in Nagasaki, Japan. International Journal of Environmental Research and Public Health. 2015; 12(8):9089-101. doi:10.3390/ijerph120809089.

72. Hou X, Fei D, Kang H, Zhang Y, Gao J. Seasonal statistical analysis of the impact of meteorological factors on fine particle pollution in China in 2013–2017. Natural Hazards. 2018; 93:677–98. doi:10.1007/s11069-018-3315-y.

73. Khare P, Baruah BP. Elemental characterization and source identification of PM2.5 using multivariate analysis at the suburban site of North-East India. Atmos Res. 2010; 98(1):148-62. doi:10.1016/j.atmosres.2010.07.001.

74. Wang J, Wang Y, Liu H, Yang Y, Zhang X, Li Y, et al. Diagnostic identification of the impact of meteorological conditions on PM2.5 concentrations in Beijing. Atmospheric Environment. 2013; 81:158-65. doi:10.1016/j.atmosenv.2013.08.033.

75. Guarnieri M, Balmes JR. Outdoor air pollution and asthma. Lancet. 2014; 383(9928):1581-92. doi:10.1016/s0140-6736(14)60617-6.

76. Anderson JO, Thundiyil JG, Stolbach A. Clearing the air: A review of the effects of particulate matter air pollution on human health. Journal of Medical Toxicology. 2012; 8(2):166-75. doi:10.1007/s13181-011-0203-1.

77. Hime NJ, Marks GB, Cowie CT. A comparison of the health effects of ambient particulate matter air pollution from five emission sources. International Journal of Environmental Research and Public Health. 2018; 15(6):1-24. doi:10.3390/ijerph15061206.

78. Department of Environment Forestry and Fisheries. Draft second generation air quality management plan for Vaal Triangle Airshed Priority Area. Pretoria,: Government Gazette. 2020.

79. Department of Environmental Affairs. Highveld Priority Area air quality management plan. Pretoria. 2012.

80. Department of Environmental Affairs. The Waterberg-Bojanala Priority Area air quality management plan: Baseline characterisation. 2014.

CHAPTER 9: UNSUPERVISED MACHINE LEARNING METHODS APPLIED IN PM2.5 SOURCE APPORTIONMENT

This chapter summarises the unsupervised ML process and results of k-means, spectral clustering and principal component analysis on the $PM_{2.5}$ samples collected in Pretoria. The samples were taken over a four-year period, from 18 April 2017 to 12 February 2021. The main sources identified for the total $PM_{2.5}$ through the seven-cluster spectral clustering model included; coal burning (42.9%), industry (22.0%), resuspended dust (10.4%), base metal (6.7%),road traffic (6.8%), vehicular exhaust (5.8%), and secondary sulphur (5.5%).

9.1. DATA

As mentioned in Chapter 8, there were 428 $PM_{2.5}$ filter samples collected from 17April 2017 to 12 February 2021. Twelve trace elements were detected: S, Cl, K, Ca, Ti, Fe, Ni, Cu, Zn, Br, U, and Si. The BC and UV-PM content of the $PM_{2.5}$ samples were also determined. The data was scaled using standardised scaling prior to applying the clustering method.

9.2. RESULTS

9.2.1. DETERMINING OF OPTIMAL CLUSTERS

The silhouette method (Figure 9a) and gap statistic (Figure 9b), were used to determine the optimal number of clusters to be used in the different clustering methods. The methods show that the optimal number of clusters for the data ranges from two to three clusters.

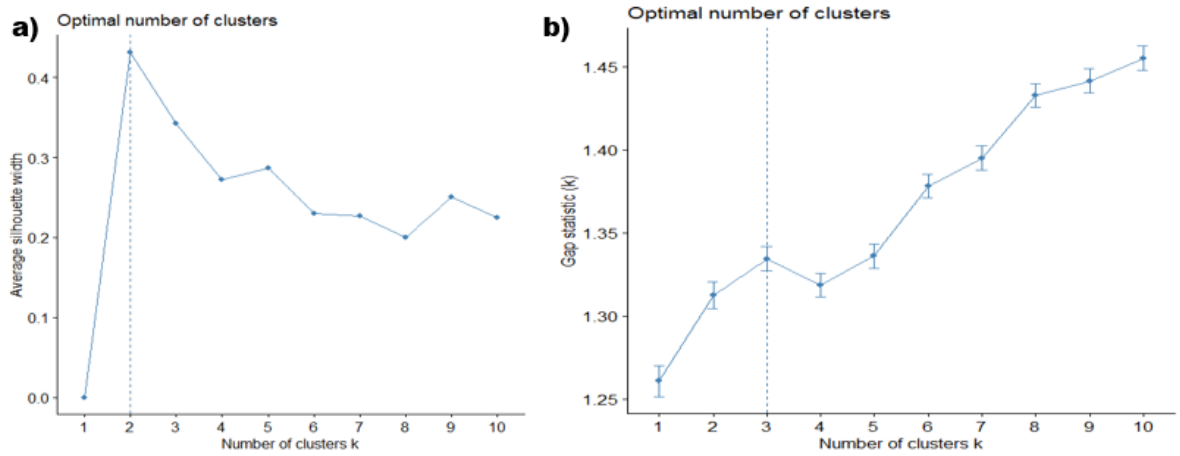


Figure 9.1: Methods a) silhouette and b) gap statistic, to determine the optimal number of clusters on the $PM_{2.5}$ and trace element data.

Although two to three clusters were shown to be ideal, clusters five to seven were included because of the number of identified sources found in Chapter 8.

9.2.2. TWO-CLUSTER K-MEANS

Figure 9.2 shows the clustering of the 428 observations on $PM_{2.5}$ BC, UVPM, S, Cl, K, Ca, Ti, Fe, Ni, Cu, Zn, Br, U, and Si, run through a two-cluster k-means model. Cluster one recorded 113 observations with a mean concentration average for $PM_{2.5}$, ~43.6 µg/m^3 and cluster two recorded 315 observations with a mean concentration average of ~14.0 µg/m^3.

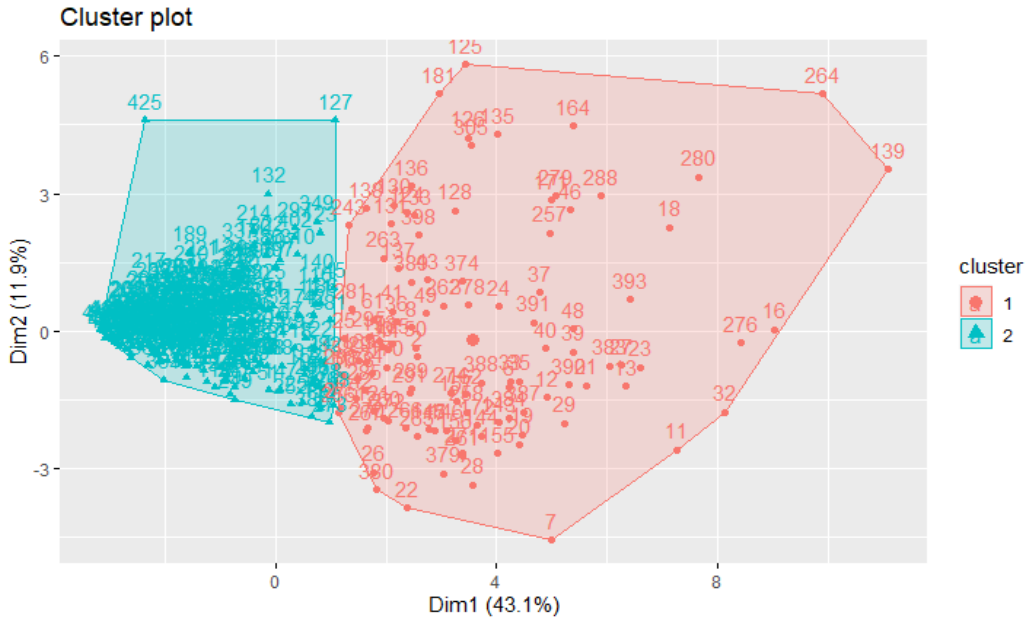


Figure 9.2: Two-cluster k-means model for $PM_{2.5}$ and trace element data. Dim-Dimension reduction representing a certain amount of variation contained in the original dataset.

Figure 9.3 indicates the percentage distribution of each species per cluster. The proportion of $PM_{2.5}$ in clusters one and two were 52.8% and 47.2%, respectively. The $PM_{2.5}$ concentration levels were significantly different between the clusters (p-value < 0.001).

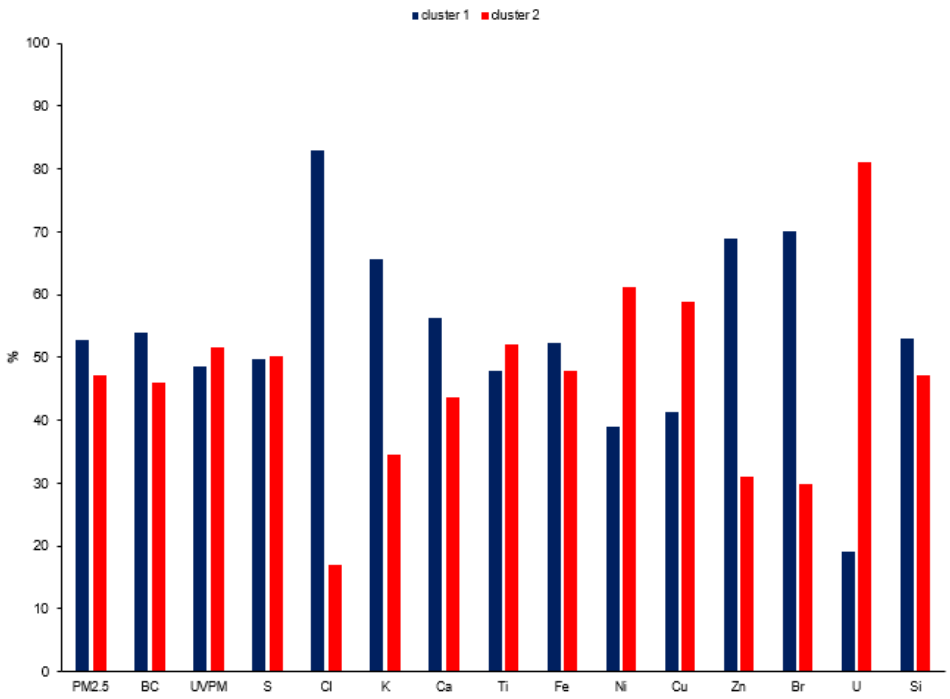


Figure 9.3: Bar graph of the percentage distribution for the two-cluster k-means model for $PM_{2.5}$ and trace element data.

9.2.3. FIVE-CLUSTER K-MEANS

Figure 9.4 shows the clusters formed using a five-cluster k-means model. Interestingly the model created a cluster that only had one observation.

Figure 9.4: Five-cluster k-means model for $PM_{2.5}$ and trace element data. Dim-Dimension reduction representing a certain amount of variation contained in the original dataset.

Figure 9.5 shows the percentage distribution of each species per cluster. The cluster distribution for clusters one, two, three, four, and five were 35, 56, 1, 268, and 68, respectively. The mean concentration of $PM_{2.5}$ was 51.7 µg/m^3 for cluster one, 23.6 µg/m^3 for cluster two, 1.5 µg/m^3 for cluster three, 12.5 µg/m^{3}for cluster four, and 41.9 µg/m^3 for cluster five. The percentage distribution for $PM_{2.5}$ for clusters one, two, three, four, and five were 19.4%, 14.2%, 0.02%, 35.9%, and 30.5%, respectively. The $PM_{2.5}$ concentration levels were significantly different among the clusters (p-value < 0.001).

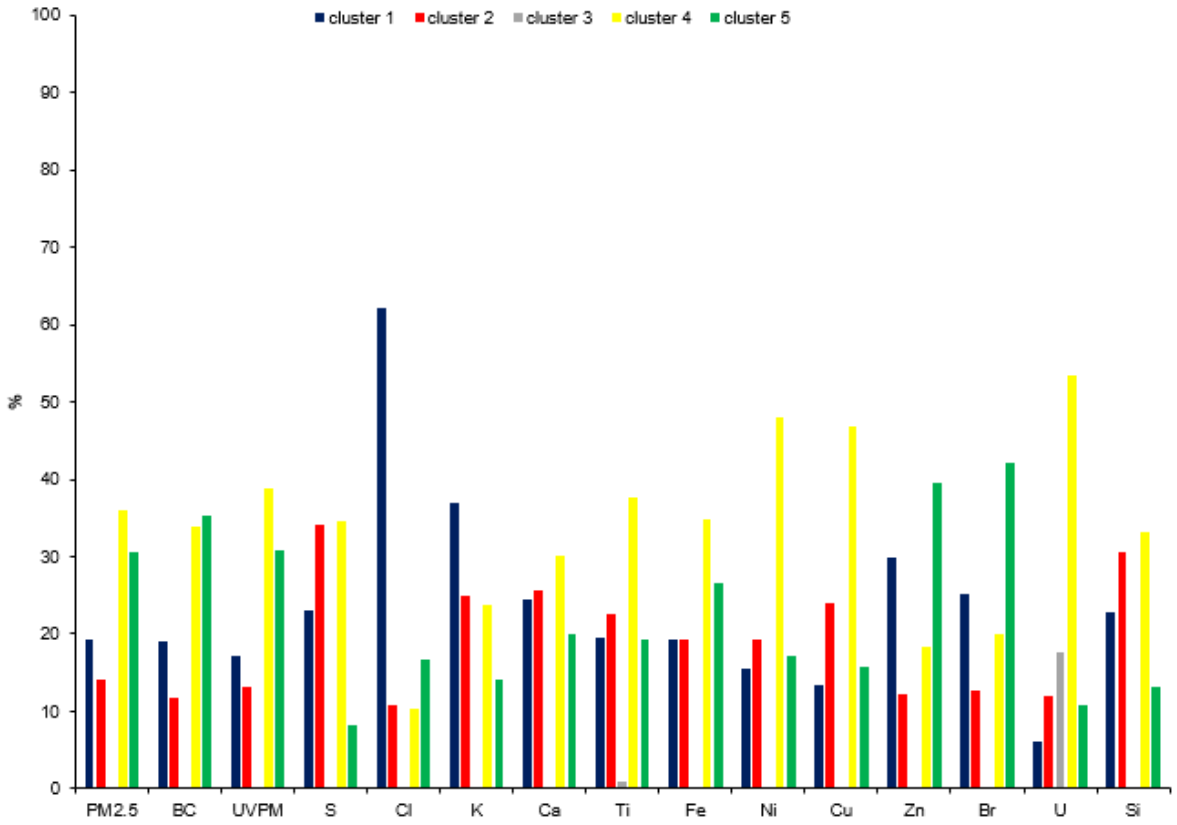


Figure 9.5: Bar graph of the percentage distribution for the five-cluster k-means model for $PM_{2.5}$ and trace element data.

9.2.4. SIX-CLUSTER K-MEANS

Figure 9.6 shows the cluster grouping when k was set at six-clusters and shows the same 'grouping' as observed when k was set at five.

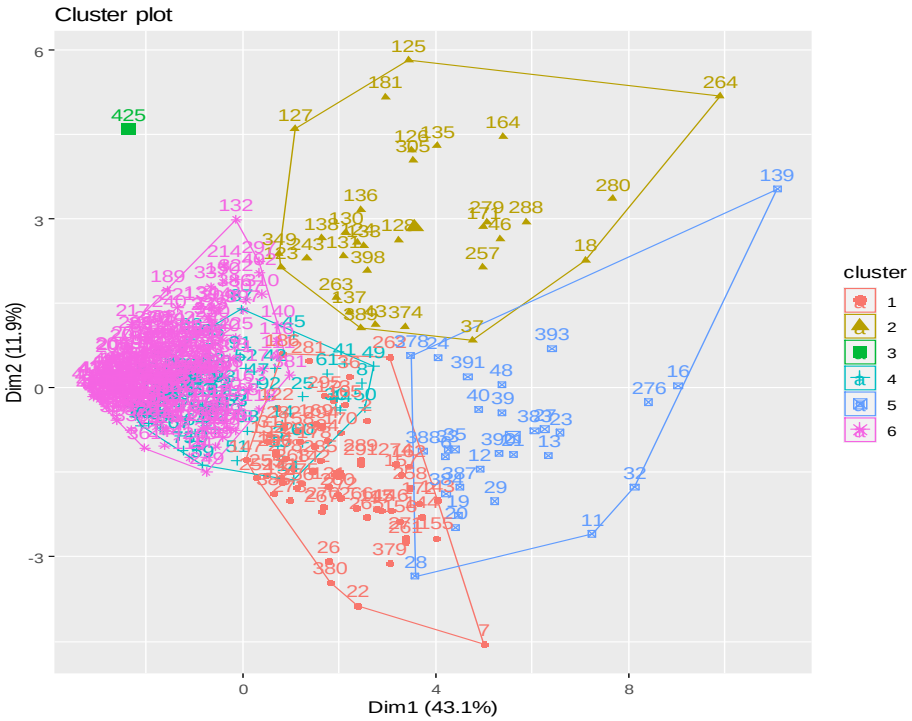


Figure 9.6: Six-cluster k-means cluster model for $PM_{2.5}$ and trace element data.
Dim- Dimension reduction representing a certain amount of variation contained in the original dataset.

The cluster distribution for clusters one, two, three, four, five, and six were 62, 32, 1, 64, 29, and 240, respectively. The mean concentration of $PM_{2.5}$ was 42.4 µg/m^3 for cluster one, 34.7 µg/m^3 for cluster two, 1.5 µg/m^3 for cluster three, 14.6 µg/m^3 for cluster four, 52.2 µg/m^3 for cluster five, and 13.1 µg/m^3 for cluster six. The percentage distribution for $PM_{2.5}$ for clusters one, two, three, four, five, and six were 28.2%, 11.9%, 0.02%, 10.0%, 16.2%, and 33.7%, respectively (Figure 9.7). The $PM_{2.5}$ concentration levels were significantly different among the clusters (p-value < 0.001).

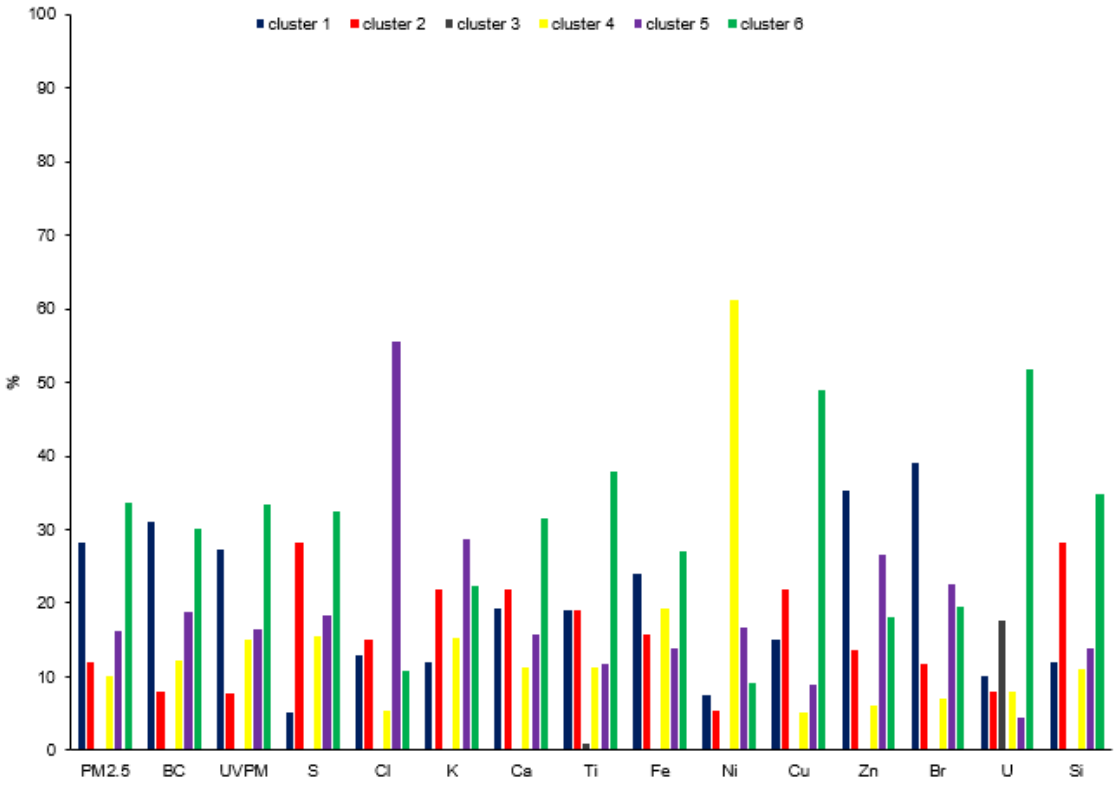


Figure 9.7: Bar graph of the percentage distribution for the six-cluster k-means model for $PM_{2.5}$ and trace element data.

Table 9.1 shows the centres of each cluster for the different cluster models. The two-cluster model shows distinct centre points for the clusters, e.g. for $PM_{2.5}$ in cluster one the centre was 1.22, and for cluster two the centre was -0.44. The larger cluster models do not show clear overall centre differences, e.g. in the five-cluster model the Cl centres are 2.31, -0.06, -0.35,-0.29, and 0.02. The centres are closer to each other and it is more prominent to see the overlapping of clusters in figures 9.4 and 9.6. Table 9.2 shows the possible sources identified when the k-mean clusters were set to two, five, and six.

Table 9.1: A summary of cluster centres for each cluster model, i.e. 2, 5 and 6-clusters, among the different species.

Model	PM2.5	BC	UVPM	S	Cl	K	Ca	Ti	Fe	Ni	Cu	Zn	Br	U
2 clusters														
1	1.22	1.19	1.16	0.58	0.75	1.09	1.05	0.92	1.22	0.28	0.25	0.95	1.10	-0.08
2	-0.44	-0.43	-0.41	-0.21	-0.27	-0.39	-0.38	-0.33	-0.44	-0.10	-0.09	-0.34	-0.40	0.03
5-clusters														
1	1.67	1.53	1.52	1.19	2.31	2.59	1.83	1.58	1.71	0.54	0.29	1.56	1.38	-0.07
2	0.01	-0.11	0.002	1.05	-0.06	0.67	0.89	0.83	0.59	0.29	0.37	-0.04	-0.01	-0.02
3	-1.14	-1.13	-1.32	-0.65	-0.35	-0.71	-0.91	2.96	-1.24	-0.58	-0.43	-0.58	-0.64	20.19
4	-0.52	-0.52	-0.53	-0.29	-0.29	-0.45	-0.48	-0.45	-0.55	-0.14	-0.11	-0.41	-0.45	-0.04
5	1.12	1.39	1.31	-0.32	0.02	-0.08	0.24	0.24	0.83	0.04	-0.003	0.87	1.1	-0.08
6-clusters														
1	1.15	1.3	1.23	-0.42	-0.04	-0.13	0.31	0.35	0.82	-0.29	0.04	0.84	1.12	-0.08
2	0.72	0.08	0.06	1.82	0.36	1.4	1.79	1.76	1.37	-0.16	0.873	0.49	0.38	0.02
3	-1.14	-1.13	-1.32	-0.65	-0.35	-0.71	-0.91	2.96	-1.24	-0.58	-0.43	-0.58	-0.64	20.12
4	-0.4	-0.21	0.01	0.03	-0.23	0.01	-0.22	-0.27	0.36	1.83	-0.3	-0.34	-0.35	-0.13
5	1.7	2.03	1.98	1.11	2.53	2.38	1.23	0.84	1.3	0.88	0.15	1.72	1.55	-0.1
6	-0.49	-0.03	0.56	-0.27	-0.28	-0.44	-0.4	-0.37	-0.65	-0.5	-0.06	-0.4	-0.43	-0.02

Table 9.2: A summary of the main trace elements in each cluster model using k-means clustering.

2	1	Biomass/Coal burning	Cl, K, Zn, Br
	2	Base metals	U, Ni, Cu
5	1	Vehicle exhaust	Cl, K, Zn
	2	Secondary sulphur	S, Si, K, Cu
	3	Industry	U
	4	Base metal/ Pyrometallurgy	UV-PM, U, Ni, Cu
	5	Biomass/ coal burning	BC, UV-PM, Zn, Br
6	1	Vehicle exhaust	BC, Zn, Br
	2	Industry	S, Si, K, Ca, Cu
	3		Ti, U
	4	Base metal/Pyrometallurgy	Ni, Fe, K, Si
	5	Biomass/Coal Burning	Cl, K, Zn, Br
	6	Manufacturing	U, Cu, Ti, Si

9.3. SPECTRAL CLUSTERING

Due to the findings in Chapter 7 only the normal Laplacian matrix was used for spectral clustering in this chapter.

9.3.1. FIVE-CLUSTER SPECTRAL CLUSTERING

Figure 9.8 shows the clusters formed using spectral clustering at a five-cluster model. Clusters one, two, three, four, and five had $PM_{2.5}$ mean concentration averages of 36.5 µg/m^3, 50.8 µg/m^3, 10.1 µg/m^3, 22.8 µg/m^3, and 16.7 µg/m^3, respectively. Figure 9.9 shows the proportion distribution of the species among the clusters. The percentage distribution for $PM_{2.5}$ for clusters one, two, three, four, and five were 35.9%, 20.7%, 18.9%, 8.8%, and 15.7%, respectively. Kruskal Wallis tests shows a significant difference in $PM_{2.5}$ among the different clusters (p-value < 0.001).

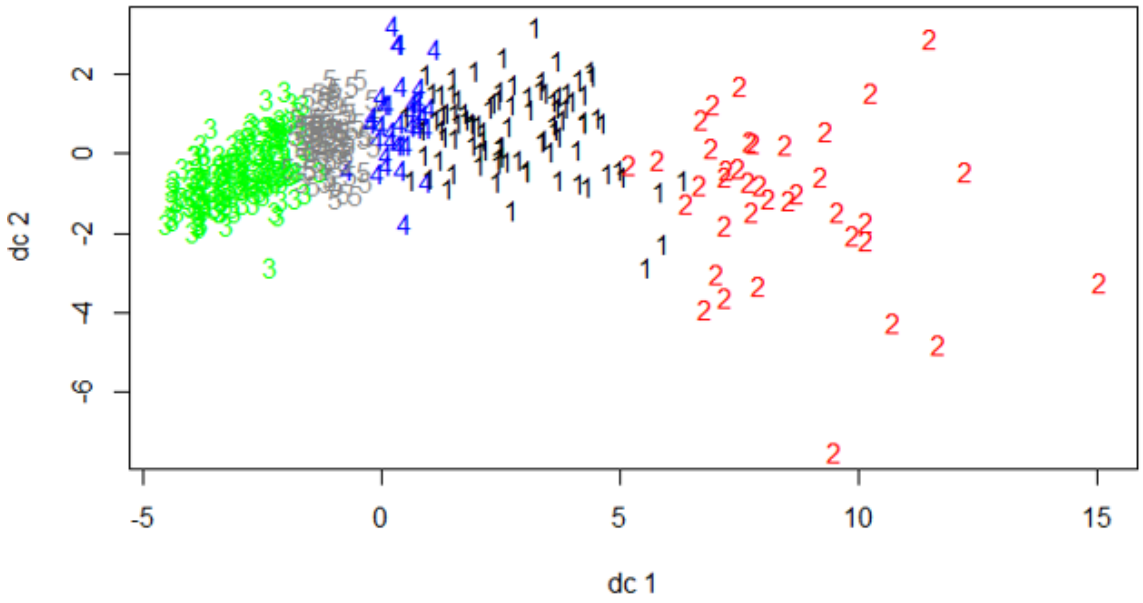


Figure 9.8: Five-cluster spectral cluster model for $PM_{2.5}$ and trace element data.
dc-data centres.

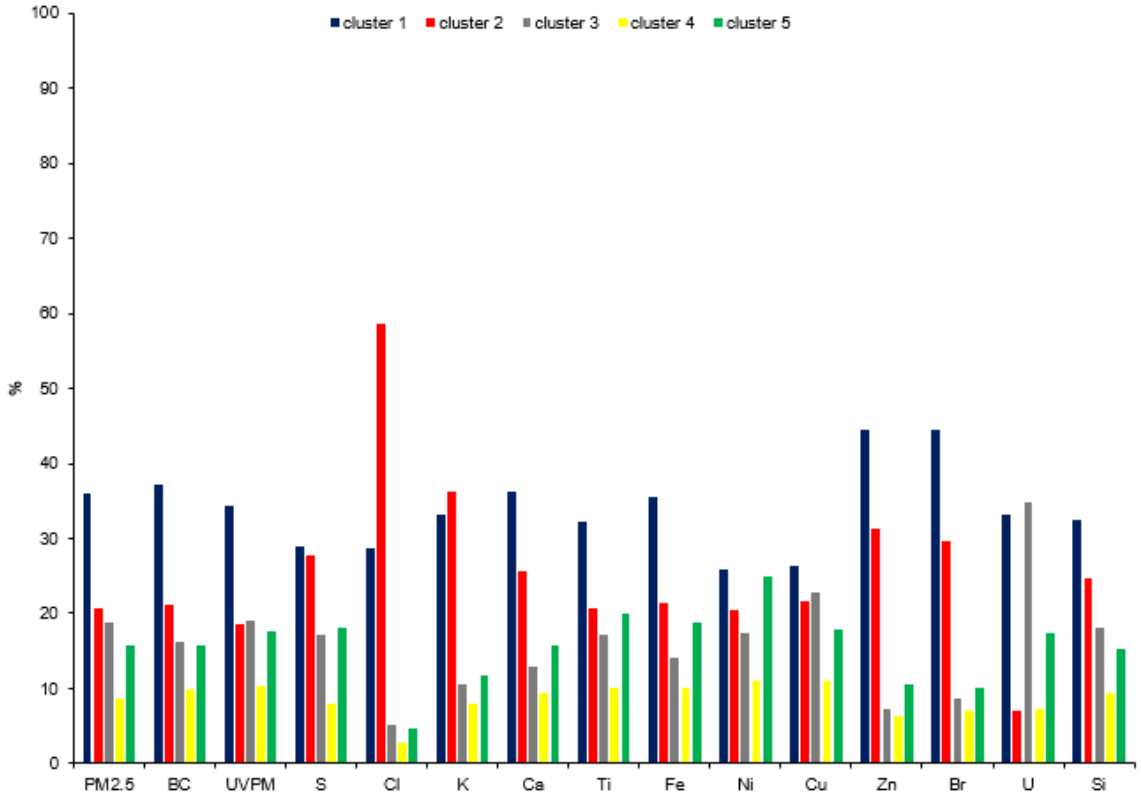


Figure 9.9: Bar graph of the percentage distribution for the five-cluster spectral cluster model for $PM_{2.5}$ and trace element data.

9.3.2. SIX-CLUSTER SPECTRAL CLUSTERING

Figure 9.10 shows the clusters formed using spectral clustering at a six-cluster model. Clusters one, two, three, four, five, and six had $PM_{2.5}$ mean concentration averages of 39.8 µg/m^3, 48.9 µg/m^3, 8.5 µg/m^3, 23.6 µg/m^3, 17.5 µg/m^3, and 14.3 µg/m^3, respectively. Figure 9.11 shows the proportion distribution of the species among the clusters. The percentage distribution for $PM_{2.5}$ for clusters one, two, three, four, five, and six were 27.7%, 23.6%, 11.1%, 14.1%, 12.2%, and 11.3%, respectively. Kruskal Wallis tests shows there was a significant difference in $PM_{2.5}$ among the clusters (p-value < 0.001).

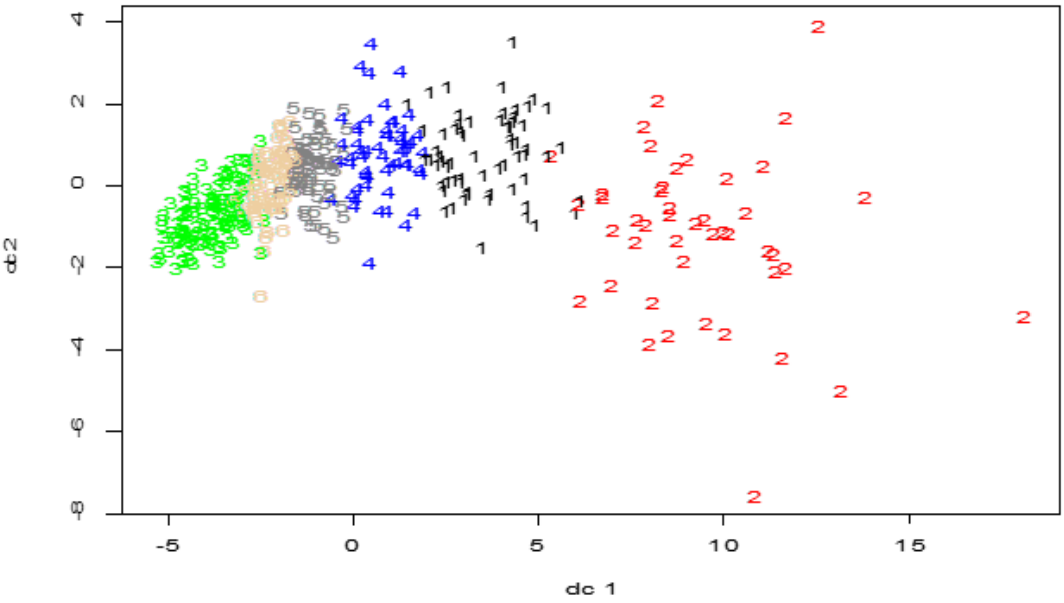


Figure 9.10: Six-cluster spectral cluster model for $PM_{2.5}$ and trace element data.
Dc-data centres.

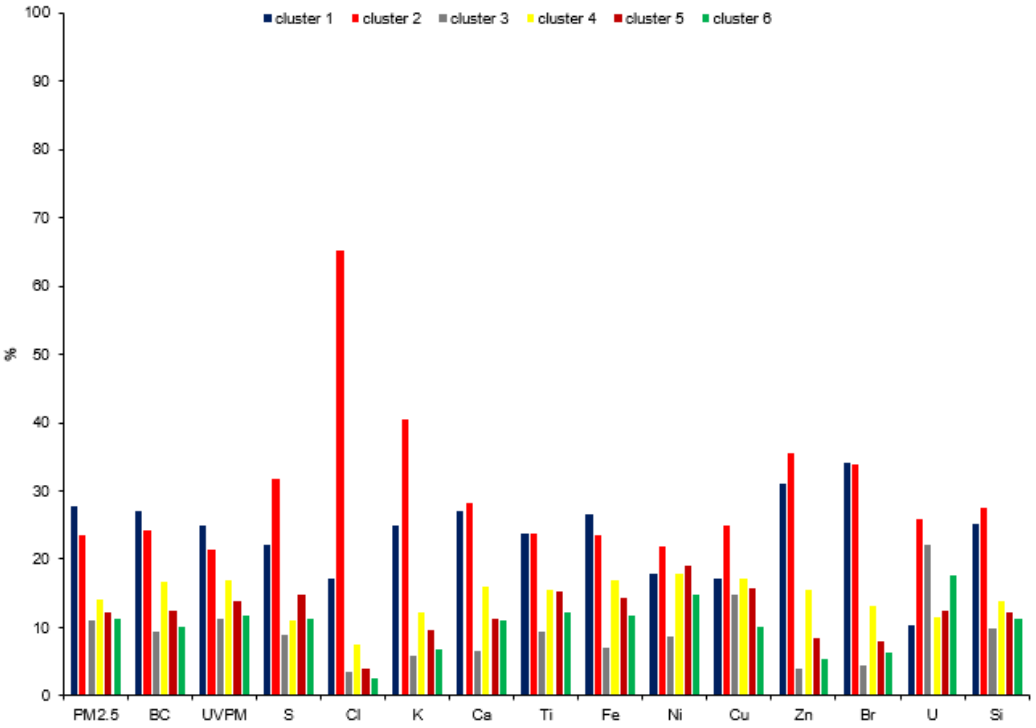


Figure 9.11: Bar graph of the percentage distribution for the six-cluster spectral cluster model for $PM_{2.5}$ and trace element data.

9.3.3. SEVEN-CLUSTER SPECTRAL CLUSTERING

Figure 9.12 shows the clusters formed using spectral clustering at a seven-cluster model. Clusters one, two, three, four, five, six, and seven had $PM_{2.5}$ mean concentration averages of 27.0 µg/m^3, 45.5 µg/m^3, 7.2 µg/m^3, 17.6 µg/m^3, 15.1 µg/m^3, 14.2 µg/m^3, and 10.7 µg/m^3, respectively. Figure 9.13 shows the proportion distribution of the species among the clusters. The percentage distribution for $PM_{2.5}$ for clusters one, two, three, four, five, six, and seven were 22.0%, 42.9%, 5.8%, 10.4%, 6.8%, 6.7%, and 5.5%, respectively. The Kruskal Wallis test shows that there was a significant difference in $PM_{2.5}$ among the different clusters (p-value < 0.001).

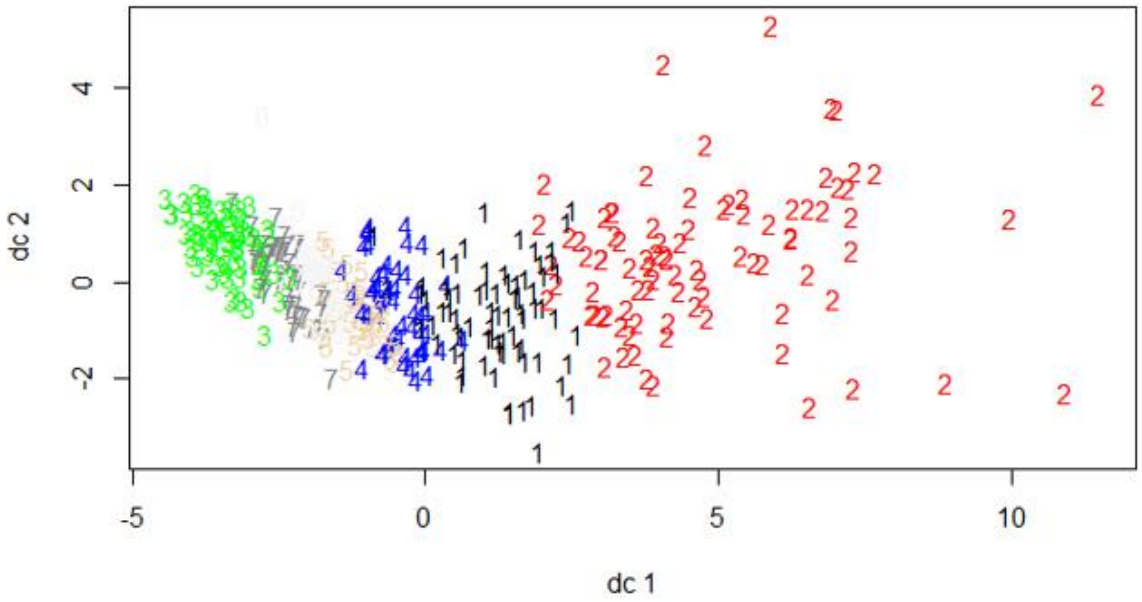


Figure 9.12: Seven-cluster spectral cluster model analysis for $PM_{2.5}$ and trace element data. dc-data centres.

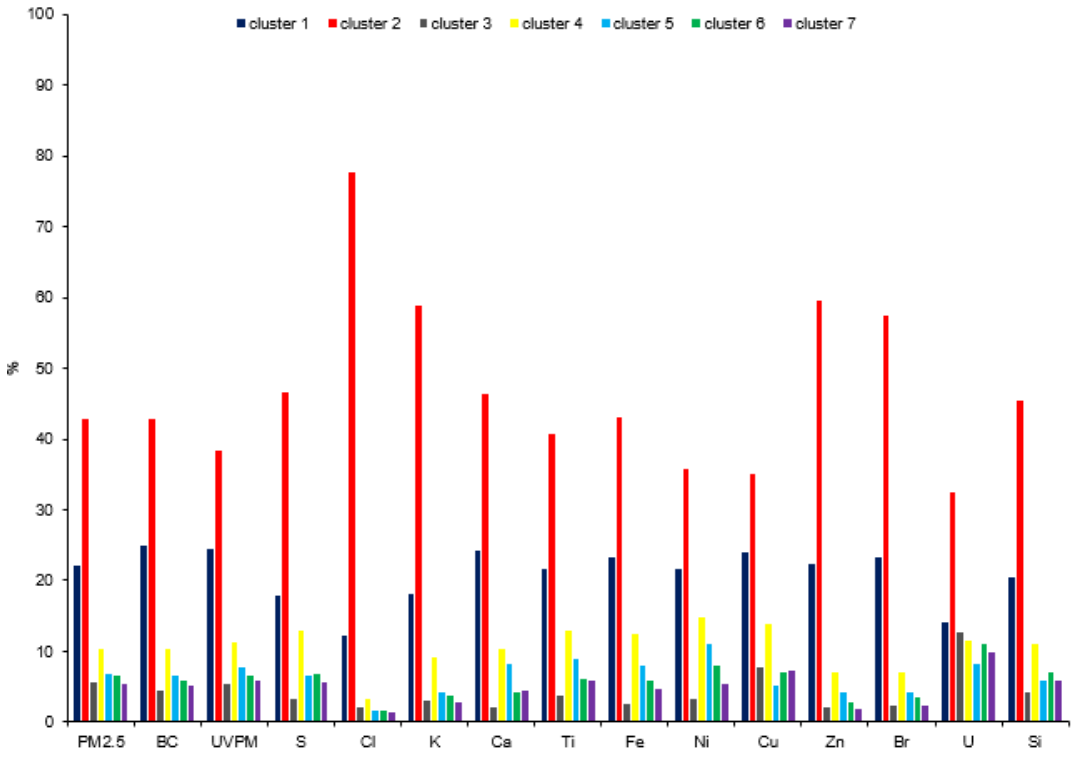


Figure 9.13: Bar graph of the percentage distribution for the 7-clusters spectral cluster model for $PM_{2.5}$ and trace element data.

Table 9.3 shows the centre clusters of each of the three cluster models. The cluster points show to be closer to each other, and in figures 9.8, 9.10, and 9.12 some overlapping of clusters can be seen. Table 9.4 then shows the possible sources

identified from spectral clustering, with the dominant trace metals identified in each possible source.

Table 9.3: A summary of cluster centres for each spectral clustering models, i.e. 5, 6 and 7-clusters, among the different species.

cluster 1	cluster 2	cluster 3	cluster 4	cluster 5		
-1.76E-08	-2.02E-08	-1.90E-08	-2.63E-02	-8.04E-09		
-0.85	-0.97	-0.91	0.011	-0.39		
-0.01	-0.02	-0.02	0.11	0.002		
-0.1	0.21	0.28	-0.77	-0.48		
0.48	-0.11	-0.29	-0.25	0.75		
cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	
-2.01E-08	-2.22E-02	-1.86E-08	-9.38E-09	-1.70E-08	-5.56E-09	
-0.96	-0.1	-0.89	-0.45	-0.81	-0.27	
-0.02	0.09	-0.02	-0.01	-0.02	0.01	
0.21	-0.61	0.27	-0.25	-0.02	-0.52	
-0.11	-0.12	-0.28	0.61	0.33	0.73	
0.02	-0.31	-0.21	0.59	0.45	0.32	
cluster 1	cluster 2	cluster 3	cluster 4	cluster 5	cluster 6	cluster 7
-1.81E-08	-1.44E-08	-1.68E-08	-8.05E-09	-1.92E-08	-2.22E-02	-4.74E-09
-0.87	-0.69	-0.81	-0.39	-0.92	-0.1	-0.23
-0.02	-0.02	-0.02	-0.01	-0.02	0.09	0.003
0.26	0.22	0.17	-0.14	0.25	-0.59	-0.43
-0.25	-0.25	-0.08	0.39	-0.2	-0.12	0.62
-0.17	-0.2	0.03	0.4	-0.1	-0.3	0.29
-0.27	-0.6	0.54	0.64	0.14	0.03	-0.5

Table 9.4: A summary of the main trace elements in each cluster after spectural clustering anlysis.

5	1	Biomass/coal burning	BC, Zn, Br
	2	General exhaust	Cl, K
	3	Base metal	U, Cu, Ni, Si
	4	Residual diesel/ road traffic	Ni, Cu
	5	Industry	Ni, Fe, Ti
6	1	Biomass/coal burning	BC, Zn, Br
	2	Vehicle exhaust	Cl, K, Zn Br
	3	Fossil fuel combustion	U, Cu
	4	Industry	Ni, Cu, Fe, Ca, Ti
	5	Secondary sulphur	Ni, Cu, Ti, S
	6	Base metal	U, Ni, Ti, Fe
7	1	Industry	BC, UV-PM, Ca, Cu, Zn
	2	Coal burning	Cl, Zn, K, Br, Si, BC
	3	Vehicle exhaust	U, Cu
	4	Resuspended dust	Ni, S, Cu
	5	Road traffic	Ni, U, Ti, Ca
	6	Base metal	U, Si
	7	Secondary sulphur	U, CU

9.4. PRINCIPAL COMPONENT ANALYSIS

PCA was run on the data as an additional comparison. Due to the nature of PCA, four runs were executed because of the strong correlation between $PM_{2.5}$, BC, and UV-PM. Figure 9.14 shows results from PCA which included all 15 species. The grouping of the species can be seen in PCA graph (Figure 9.14a). The scree plot (Figure 9.14b) shows the percentage of each principal component. $PM_{2.5}$ is the principal component one at 97.5%.

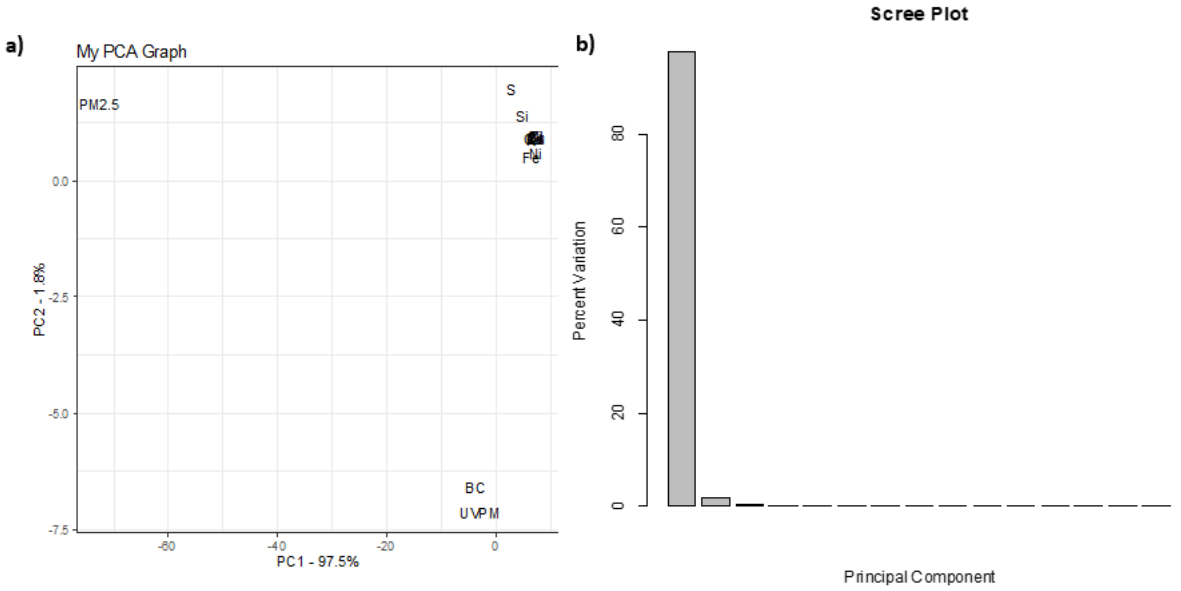


Figure 9.14: PCA run on for $PM_{2.5}$ and trace elements, a) the PCA graph and b) scree plot of component proportion distribution. PC-principal component.

Figure 9.15 shows (a) the PCA graph and (b) the scree plot omitting UV-PM in the analysis. Figure 9.16 shows (a) the PCA graph and (b) the scree plot omitting BC in the analysis. Figure 9.17 shows (a) the PCA graph and (b) the scree plot omitting both BC and UV-PM in the analysis. The PCA analyses showed $PM_{2.5}$ as the principal component in all four runs. It did not show any grouping that would assist in the prediction of possible sources.

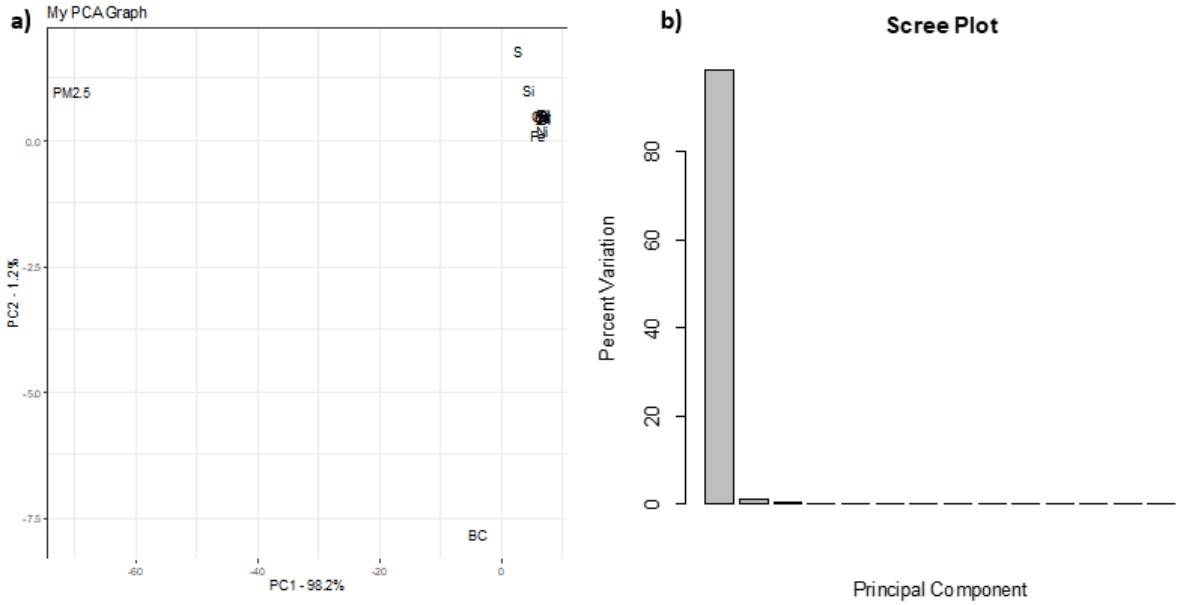


Figure 9.15: PCA omitting UV-PM run on for $PM_{2.5}$ and trace elements, a) the PCA graph and b) scree plot of component proportion distribution. PC-principal component.

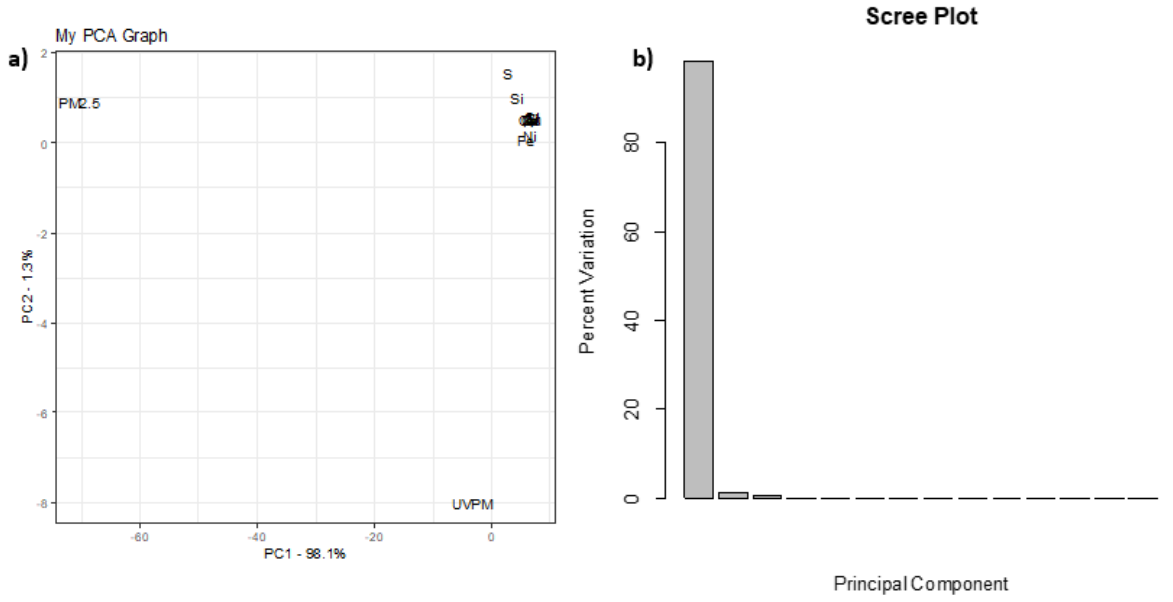


Figure 9.16: PCA omitting BC run on for $PM_{2.5}$ and trace elements, a) the PCA graph and b) scree plot of component proportion distribution. PC-principal component.

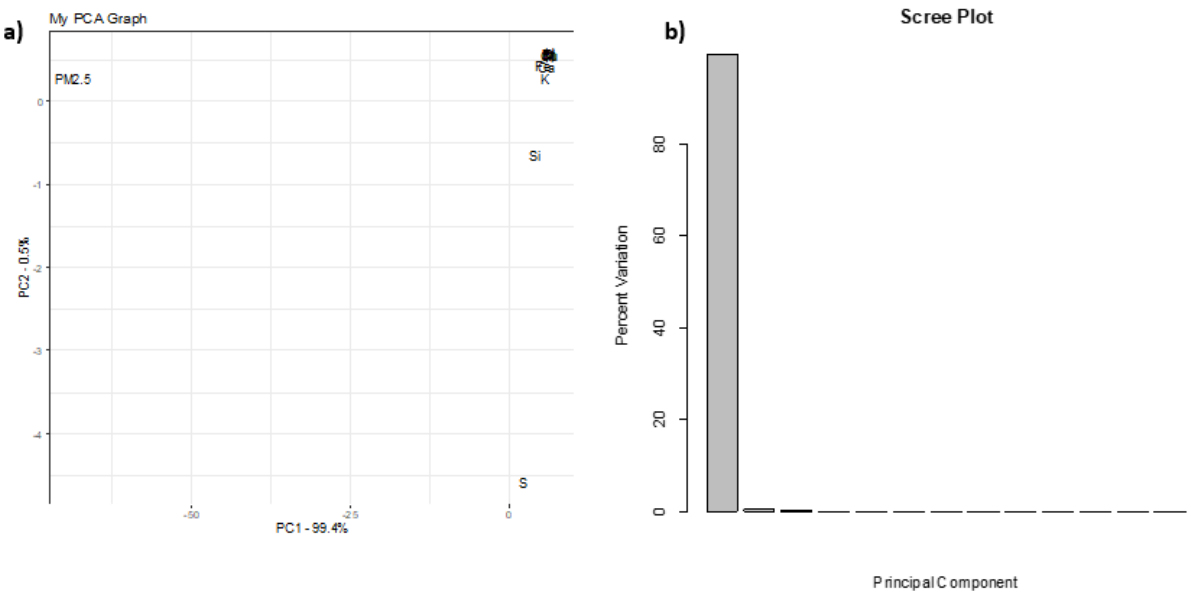


Figure 9.17: PCA omitting BC and UV-PM run on for $PM_{2.5}$ and trace elements, a) the PCA graph and b) scree plot of component proportion distribution. PC-principal component.

9.5. DISCUSSION

The aim of this chapter was to apply unsupervised ML clustering methods for source apportionment of $PM_{2.5}$ data collected from 18 April 2017 to 12 February 2021. The use of clustering techniques is becoming a popular method as it reduces large datasets into logical groups to help researchers analyse their large datasets.[1] This section of the study explores the practicality of these methods for source apportionment as an important part of air pollution studies.[2] K-means based clustering and spectral clustering have been used in previous studies for source apportionment.[2-3] PCA is also a dimension reducing method,[4] however, the results only showed $PM_{2.5}$ as the component with the most influence in the dataset This proved not to be an appropriate method to use for source apportionment.

Determining the optimal number of clusters is a practical step prior to running the cluster models. Different methods suggested the optimal number of clusters ranged from two to three, however, in the PMF analysis (Chapter 8) and previous studies of the same area,[5-7] suggest that there were more than two identified sources in the study area. For this reason, the number of clusters were increased in the cluster models. Using the two-cluster k-means model showed distinct separation in the clusters and no overlapping was seen, but it would not be logical to assume that there are only two

possible sources on $PM_{2.5}$ in the study area of interest. When five and six k-means cluster models were run, overlapping among the clusters and single clusters were observed. Further inspection of their cluster centres showed that they were fairly close together and were not as distinguishable as the two-cluster model. The same single observation was identified in the five- and six-cluster models. Although, this 'cluster' may not seem logical, it had the highest proportion of uranium and could be worth exploring in future studies. Although k-means clustering models in this project did not yield fruitful results, a 2012 study showed k-means clustering in combination with hierarchal clustering to solve the issue of outliers and produce interpretable results based on the chemical, physical, and meteorological data of air pollution data.[8]

Kumar et al,[3] found that using k-means clustering was not appropriate, because of the assumption that k-means clusters data in spherical distribution that may not always be practical. In general, the findings of k-means clustering for source apportionment in this study were found to agree with the findings of Kumar et al.[3] However, another study found k-means clustering to produce reasonable results for classifying geochemical patterns later used for source identification.[9] In this study k-means clustering was not an ideal dimension reduction method for source apportionment. However, Chen et al,[2] used a clustering method that fundamentally uses k-means and suggests that it could possibly be a solution the shortcomings of traditional source apportionment methods. One of these shortcomings is the issue of better handling outliers. However, k-means was found to be sensitive to outliers and the inclusion of outliers can possibly lead to poor quality clusters.[1] The latter could be the reason why a single observation was considered a cluster in the five and six k-means models.

Spectral clustering is able to cluster data without making strong assumptions on the form of the cluster as done in k-means.[10-11] To a certain extent, the clustering did prove to be relatively better outside of the optimal number of clusters, this was because there was reduced overlapping among the clusters although it was still an issue. Visually, the overlapping of the clusters was not as profound as in the k-means cluster models, but the cluster centres were all fairly difficult to distinguish. The five-cluster spectral model was a fairly reasonable model to consider for attributing possible sources in the Pretoria area. Although the possible number of sources was lower that the PMF findings. One very evident issue that arose when using both k-means and spectral

clustering was not so much issue of clustering the data, but rather the method in which one could analyse the data after clustering, in order to appoint a possible source. In this study, although simplistic, comparing the proportion of total species in a cluster was a way to assign possible source to a cluster. There is limited literature that clearly outlines how to analyse clusters for source apportionment. Kumar et al,[3] seem to use the same method of analysing the proportion of the species' contributions to assign possible source.

From these findings using the five, six, and seven spectral clustering models, the five-cluster model was the better fitting model for assigning possible sources. This was because the proportion distribution was better defined, which is possible because there was seemingly less overlap in this model, which helped to distinguish the difference between the clusters. The possible source contributors to $PM_{2.5}$ identified in this model were biomass/coal burning (35.9%), vehicle exhaust (20.7%), base metal (18.9%), residual diesel/road traffic (8.8%), and industry (15.7%). This compared to the possible sources identified in the PMF model: industry/base metal (8.7%), road traffic (11.3%), secondary sulphur (12.1%), mining (43.2%), biomass/coal burning (14.2%), resuspended dust (8.5%), and vehicular exhaust (2.0%). However, the seven-cluster spectral model still remains a plausible option to consider for the distribution of $PM_{2.5}$ industry (22.0%), coal burning (42.9%), vehicular exhaust (5.8%), resuspended dust (10.4%), road traffic (6.8%), base metal (6.7%), and secondary sulphur (5.5%).

Industry, biomass/coal burning, and base metal in the spectral clustering models showed to be large contributing sources of $PM_{2.5}$ in Pretoria, South Africa, which is similar to the findings in the PMF model. The average modelled $PM_{2.5}$ in the PMF analysis was 17.4 µg/m^3, while the unsupervised ML cluster models maintained the actual data, retaining the 21.8 µg/m^3 average. The PMF model accounted for 60% of the data and the spectral clustering model utilised all the data observations in dataset. The 2017 study conducted in the same area also exhibited that the modelled average $PM_{2.5}$ (18.3 µg/m^3) was lower than the actual sampled $PM_{2.5}$ (21.1 µg/m^3).[5] McDuffie et al, [12] found that coal, commercial waste, and open fire combustion highly contribute to the estimated $PM_{2.5}$ in South Africa by 20.5%, 2.7%, 3.9%, and 8.6%, respectively. Other possible sources, such as road transport and wind dust, contributed 6.5% and

8.5% to the country's estimated $PM_{2.5}$. Although there are limited source apportionment studies in South Africa to compare with, the spectral clustering method showed closer proportion distributions to the McDuffie et al,[12] study. This could be an advantage to using unsupervised ML as compared to the traditional PMF source apportionment method however further studies are needed.

There is still a lot to be explored when considering the use of unsupervised ML in source apportionment studies. The main advantages observed in this study was the quick and easy execution of each clustering model to reduce the dataset, in comparison to using PMF analyse. It was also seen that all the data was taken into consideration as opposed to the 0.6 correlation model from PMF analysis. These advantages could confirm the Chen et al,[2] assumption that clustering methods can assist in analysing large multi-dimensional air pollution features in source apportionment studies.

There are still evident limitations to using unsupervised ML clustering methods for source apportionment. One of the major limitations is the lack of external validation for the clusters. PMF analyses allow the researcher to set high error estimates (at 85%) that help to validate the findings from the model.[13] However, the allocation of possible pollution sources still remain a researcher-subjective task in both the PMF and spectral clustering methods.

PCA modelling did not group or cluster the data in a manner that was profitable for source apportionment. The PCA model only showed the principal component within the dataset that had most significance, i.e. $PM_{2.5}$. The process investigates independent linear combinations of the variables that produce the most variance within the exposure data, and within that dataset there cannot be any strong correlation between variables.[4,14] Thus the process was run four times, excluding the strongly correlated variables, i.e. BC, and UV-PM. Although it did not show to be an appropriate source apportionment method in this study, it has been used for source apportionment in combination with other methods, such as Chemical Mass Balance (CMB), heat mapping, and dendrograms.[15-16] The PCA model is both easy and efficient, which are the main advantages, but there is difficulty with results interpretation and a lack of relationship among the components.[4] It could be that in this project PCA analysis, the use of $PM_{2.5}$ mass together with the elements created a double counting of the mass.

Where as in the PMF analysis $PM_{2.5}$ is set as “total mass variable” and this results in the PMF program subtracting all the elemental masses from the $PM_{2.5}$ mass, to prevent double counting of the mass.[17-18]

9.6. CONCLUSION

In conclusion, the utilisation of unsupervised ML for source apportionment is still a relatively new concept and so has to be explored further to become an established source apportionment tool. In comparison with PMF, spectral clustering showed to be a potential dimension reducing tool for source apportionment. These findings are an addition to the limited research on the utilisation of ML in air pollution studies. It is highly recommended that more source apportionment studies are conducted using spectral clustering and unsupervised ML methods in different areas of the country to continue to understand the results produced. This could provide additional information on how to better analyse the clustered data and potentially improve this method as a valuable and validated source apportionment tool for air pollution researchers.

9.7. REFERENCES

1. Mittal M, Goyal LM, Hemanth DJ, Sethi JK. Clustering approaches for high-dimensional databases: A review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2019; 9(3) doi:10.1002/widm.1300.

2. Chen M, Wang P, Chen Q, Wu J, Chen X. A clustering algorithm for sample data based on environmental pollution characteristics. Atmospheric Environment. 2015; 107:194-203. doi:10.1016/j.atmosenv.2015.02.042.

3. Kumar V, Sahu M, Biswas P. Source apportionment of particulate matter by application of machine learning clustering algorithms. Aerosol and Air Quality Research. 2022; 22(3):210240. doi:10.4209/aaqr.210240.

4. Stafoggia M, Breitner S, Hampel R, Basagaña X. Statistical approaches to address multi-pollutant mixtures and multiple exposures: The state of the science. Current Environmental Health Reports. 2017; 4(4):481-90. doi:10.1007/s40572-017-0162-z.

5. Adeyemi A, Molnar P, Boman J, Wichmann J. Source apportionment of fine atmospheric particles using positive matrix factorization in Pretoria, South Africa. Environmental Monitoring and Assessment. 2021; 193(11):716. doi:10.1007/s10661-021-09483-3.

6. Howlett-Downing C. The association between sources of air pollution and respiratory health in Pretoria, South Africa: University of Pretoria; 2022.

7. Howlett-Downing C, Boman J, Molnár P, Shirinde J, Wichmann J. PM2.5 chemical composition and geographical origin of air masses in Pretoria, South Africa. Water, Air, & Soil Pollution. 2022; 233(7) doi:10.1007/s11270-022-05746-y.

8. Austin E, Coull B, Thomas D, Koutrakis P. A framework for identifying distinct multipollutant profiles in air pollution data. Environment International. 2012; 45:112-21. doi:10.1016/j.envint.2012.04.003.

9. Khorshidi N, Parsa M, Lentz DR, Sobhanverdi J. Identification of heavy metal pollution sources and its associated risk assessment in an industrial town using the k-means clustering technique. Applied Geochemistry. 2021; 135:105113.

10. Goubko M, Veremyev A. Bilinear matrix equation characterizes laplacian and distance matrices of weighted trees. Discrete Applied Mathematics. 2021; 305:1-9. doi:10.1016/j.dam.2021.08.025.

11. von Luxburg U, Belkin M, Bousquet O. Consistency of spectral clustering. The Annals of Statistics. 2008; 36(2):555-86.

12. McDuffie EE, Martin RV, Spadaro JV, Burnett R, Smith SJ, O'Rourke P, et al. Source sector and fuel contributions to ambient PM2.5 and attributable mortality across multiple spatial scales. Nature communications. 2021; 12(1):3594.
13. U.S. Environmental Protection Agency. Receptor modeling, positive matrix factorization. 2014. Available from: http://www.epa.gov/heasd/research/pmf.html.

14. Anderson TW. An introduction to multivariate statistical analysis. 2nd ed. New York: Wiley; 1984.

15. Demir S, Saral A, Ertürk F, Kuzu L. Combined use of principal component analysis (PCA) and chemical mass balance (CMB) for source identification and source apportionment in air pollution modeling studies. Water, Air, & Soil Pollution: An International Journal of Environmental Pollution. 2010; 212(1-4):429-39. doi:10.1007/s11270-010-0358-4.

16. Zajusz-Zubek E, Mainka A, Kaczmarek K. Dendrograms, heat maps and principal component analysis-the practical use of statistical methods for source apportionment of trace elements in PM10. Journal of Environmental Science and Health - Part A Toxic/Hazardous Substances and Environmental Engineering. 2019; doi:10.1080/10934529.2019.1670026.

17. Belis CA, Karagulian F, Larsen BR, Hopke PK. Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe. Atmospheric Environment. 2013; 69.

18. US Environmental Protection Agency. Epa positive matrix factorization (PMF) 5.0 fundementals and user guide. 2014. Available from: https://www.epa.gov/air-research/positive-matrix-factorization-model-environmental-data-analyses.

CHAPTER 10: DISCUSSION SUMMARY AND RECOMMENDATIONS

The aim of this project was to assess the applicability of Artificial Intelligence (AI) methods such as Machine Learning (ML) in air pollution epidemiology in a South African context. As far as literature presented in Chapter 2, this is the first public health PhD project that addresses this topic. The first goal of the study was to assess what current attitudes and perceptions were concerning AI use in public health. Thereafter, the study sought to test the practicality of applying unsupervised ML in different air pollution studies, compared to traditional statistical modelling and dimension reduction methods. K-means clustering and spectral clustering were the main unsupervised ML clustering methods applied to investigate the joint effects of air pollution data from the Vaal Triangle Airshed Priority Area (VTAPA), South Africa. These unsupervised ML clustering methods were also used to identify possible sources of $PM_{2.5}$ in Pretoria, South Africa. Thus the main objectives were:

- To assess the perceptions and attitudes regarding AI in public health among postgraduate students registered for the online Postgraduate Diploma in Public Health at the School of Health Systems and Public Health (SHSPH), University of Pretoria (UP).
- To determine the joint effects of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10}, on hospital admissions for respiratory disease (RD) and cardiovascular disease (CVD) in Vereeniging and Vanderbijlpark, Gauteng, using CART analysis. Thereafter, using the unsupervised Machine Learning methods to determine the joint effects of the air pollutants on RD and CVD hospital admission.
- To compare two methods of source apportionment of $PM_{2.5}$ in Pretoria, namely, Positive Matrix Factorization (PMF) and unsupervised Machine Learning clustering methods.

10.1. MAIN FINDINGS

10.1.1. KAP AMONG POSTGRADUATE DIPLOMA STUDENTS

This part of the project addressed one of the nine strategic interventions of the National Digital Health Strategy for South Africa (2019 to 2024), which was “to develop enhanced digital health technical capacity and skilled workforce for digital technology

support and implementation".[1] Different medical and public health fields studies have already used AI and AI-related concepts in South Africa.[2-5] However, the main findings from this part of the project indicated that public health students were not very knowledgeable on common AI terminology. Major concerns were raised about the ethical impact AI may have in public health, as well as the negative effects on job availability in the country.

Students were quite confident that AI would be able to perform objective health tasks, and these findings were similar to a study conducted in Canada.[6] The strong perception that AI in public health is going to affect job availability to public health professionals is possibly a global issue, not only expressed in this project.[6-9] AI in health research is still highly expectant of the capabilities of AI and its different concepts.[10] These high expectations seem to be worrying the general public on their own relevance should AI be introduced within their area of work.

Therefore, in order to rationalise these expectations in future public health professionals, curricular content for AI in public health and medicine can be structured around basic understanding of AI concepts, limitations, and its relevant ethical and legal implications.[6,11-13] This educational strategy could better prepare South African public health students to understand AI. The structuring of such a curriculum can only be informed by more in-depth studies involving more South African tertiary education institutions.

10.1.2. JOINT EFFECTS OF SO_2, NO_2, O_3, $PM_{2.5}$ AND PM_{10} ON RESPIRATORY (RD) AND CARDIOVASCULAR DISEASE (CVD) HOSPITAL ADMISSION USING CART ANALYSIS

The joint effects of SO_2, NO_2, O_3, $PM_{2.5}$, and PM_{10} on RD and CVD hospital admissions were explored through grouping the pollutants in seven mixtures. The seven air pollutant mixtures explored were; (mixture 1) PM_{10}, NO_2, and SO_2; (mixture 2) $PM_{2.5}$, NO_2, and SO_2; (mixture 3) PM_{10}, NO_2, and O_3; (mixture 4) $PM_{2.5}$, NO_2, and O_3; (mixture 5) PM_{10}, SO_2, and O_3; (mixture 6) $PM_{2.5}$, SO_2, and O_3; and (mixture 7) O_3, NO_2, and SO_2. There were 3 346 observations used for CART analyses from January 2011 to February 2020.

Prior to running the CART analyses, imputation was run to address the missing data for both the air pollution and meteorological variables. The proportion of missing data

ranged between ~11% and ~34% across the five pollutants of interest, and between ~10% and ~22% for temperature relative humidity and wind speed. Both univariate and multivariate imputation methods were performed, namely: the Kalman univariate time-series method, multiple imputation by chain equations (mice) method and multiple time-series data imputation (mtsdi) method. The multivariate methods, i.e. mice and mtsdi methods, performed well and were able to capture the time variability in the time-series of the air pollutants. Ultimately, the mice imputed datasets were used in the CART analyses, as evidence showed that mice imputation was an acceptable imputation method. [14-17]

In total 54 822 RD and 22 520 CVD hospital admissions were recorded from the private hospitals in Vereeniging and Vanderbijlpark. The strength of the association between pollutants mixtures and RD and CVD hospital admissions was investigated using statistical modelling via CART model. A time-stratified case-crossover epidemiological design was used that controlled for long-term trend and seasonality as well as apparent temperature.[18-19]

This project investigated the health effects of gaseous pollutant combinations including O_3. The results of the study provide epidemiological evidence that the air pollution concentrations in the VTAPA continue to put community members at risk of RD and CVD hospitalisation. High levels of NO_2 in combination with various levels of SO_2, O_3, $PM_{2.5}$, and PM_{10}, was found to increase the risk of both RD and CVD hospital admissions. The current NAAQIS showed to be lenient in comparison to the WHO guideline for NO_2, for instance, only 14% (475/3 347 days) data exceeded the daily NAAQS but exceeded the daily WHO guideline by 67.5% (2 260/3 347 days). Thus, there may be need to revise and update the current VTAPA air quality management plans.[20] This could increase the control and mitigation of air pollution concentration levels in the area.

10.1.3. JOINT EFFECTS OF SO_2, NO_2, O_3, $PM_{2.5}$ AND PM_{10} ON RD HOSPITAL ADMISSION USING UNSUPERVISED MACHINE LEARNING ANALYSIS

This was the first study in the country, and only limited published studies had applied unsupervised ML to investigate the joint effects of multiple air pollutants on health data,

globally.[21-23] Spectral clustering using the normalised Laplacian matrix showed that exposure to higher levels of SO_2, NO_2, $PM_{2.5}$, and PM_{10} increases the risk of RD hospital admission by 1.04. No cluster mixtures significantly increased the risk of CVD hospital admission. The air pollutant mixtures were set by the unsupervised ML clustering algorithm as opposed to CART analyses, where the air pollutant mixtures are set priori. This was beneficial as it reduced the time taken to run the models. In addition, running the cluster models were not computationally taxing.

Due to the complex mixture of air, setting priori mixtures helps to avoid placing highly correlated pollutants in the same mixtures, e.g. $PM_{2.5}$ and PM_{10}. Initially, this did not seem to be an issue for the unsupervised ML clustering methods. All pollutants, in spite of their correlation could be set through the same analysis. Other studies using unsupervised ML clustering found it relatively easy to run multiple air pollutants in one model.[22-23] However, the inverse relationship of O_3 with other pollutants could be a concern and results from this study and other studies showed significant mixtures only included low levels of O_3. Although DBSCAN clustering was used in the project it proved to be inappropriate as it reduced the sample data and produced unfeasible cluster sizes.

Overall the results provide baseline evidence that unsupervised ML can be useful in joint effects epidemiology studies. However, within the Gartner's AI hype cycle, it seems that it would be premature to assume that research has reached the '*Plateaux of productivity.'* It would be more reasonable to accept that some research is still in the '*Peak of inflated expectations'* entering into the *'Trough of Disillusionment'.*[10] It is evident that there is still more to be done in order to assert ML and AI methods as practical air pollution epidemiology study analysis tools.

10.1.4. SOURCE APPORTIONMENT ON $PM_{2.5}$ IN PRETORIA (2017-2021) USING POSITIVE MATRIX FACTERIZATION (PMF)

A seven-factor PMF model was assigned to $PM_{2.5}$ data collected over a 46-month period in Pretoria, South Africa. The seven contributing sources identified included, mining (43.2%), biomass/coal burning (14.2%), secondary sulphur (12.1%), road traffic (11.3%), industry/base metal (8.7%), resuspended dust (8.5%), and general

exhaust (2.0%) emissions. These sources were similar to those identified in previous studies conducted in the same area.[24-26]

The main focus was to run source identification using PMF. It is very evident as to why PMF is a reliable source apportionment tool used in multiple studies.[27-31] The analysis process is relatively easy to run and interpret, and the biggest advantage is the availability of the different error estimations that provide a validation process for the researcher. Another noticeable advantage is that the results remained consistent. The results from this study showed to be consistent with the two studies conducted in previous years, with little difference in identified sources in the area. Nonetheless, it does come with some limitations. The PMF model only showed a 0.6 correlation with the original dataset, leaving a 40% error margin when making source apportionment assumptions. This also reduced the total $PM_{2.5}$ mean average compared to the original data. Another limitation is that the process can be computationally taxing on medium to large datasets.

10.1.5. SOURCE APPORTIONMENT ON $PM_{2.5}$ IN PRETORIA (2017-2021) USING UNSUPERVISED MACHINE LEARNING CLUSTERING

Literature showed that using unsupervised ML clustering methods for source apportionment was limited.[32-34] This project adds to the limited information on the use of ML in source apportionment studies. Additionally, more than one clustering method was used, which gives an additional comparison.

Spectral clustering using the normalised Laplacian matrix showed to perform well for source apportionment. The clustering showed less overlapping than was seen in the k-means cluster models. The seven-cluster spectral model suggested coal burning (42.89%), industry (22.01%), resuspended dust (10.36%), road traffic (6.78%), base metal (6.69%), general exhaust (5.76%), and secondary sulphur (5.52%) as possible sources. Although the percentage contribution on $PM_{2.5}$ differed from the sources identified using PMF, the possible identified sources did not show major variations in the sources identified. This could suggest that spectral clustering has the potential to be a source apportionment tool in air pollution studies.

The results of using the unsupervised ML clustering seemed to address the limitations identified in the PMF analysis. The clustering methods were fast to run and utilised the

entire dataset in the different cluster models, showing great potential as a dimension reduction tool for large datasets. However, these advantages do not outweigh the prominent setbacks. The optimal number of possible clusters was lower than the anticipated number of possible sources for the study area established from the PMF study. This suggests that, at this stage, unsupervised ML clustering algorithms could lead to an underestimation of the number of possible clusters, showing that the researcher will still need to apply subjective logic before running unsupervised ML models.

10.2. STRENGTHS AND LIMITATIONS OF THE PROJECT

10.2.1. STRENGTHS

Although ML has been used in multiple health studies in South Africa, literature suggests that this PhD project is the first of its kind. It presents baseline information on the perceptions and attitudes of AI among public health students. The project also adds to the limited information on the use of unsupervised ML clustering to investigate joint effects of air pollution on hospital admissions and source apportionment studies.

This study also provided epidemiological evidence of the negative health effects on multiple air pollution mixtures that include O_3, NO_2, and SO_2 in the VTAPA, South Africa. The project utilises imputation to address the common issue of missing data in environmental studies.[14-15] It also shows different methods of imputation that can be applied to missing data under the MAR assumption of missingness. It also included using external meteorological variables like temperature, relative humidity, and wind speed to impute missing values.

10.2.2. LIMITATIONS

This is a public health study and a data science study. Therefore, it showed the application of a few unsupervised ML methods and did not apply an exhaustive list of ML methods. The cross-sectional survey only provided a glance at perceptions and attitudes towards AI in public health, and these results are not generalizable to other public health students in South Africa. The private hospital data used to investigate joint effects of air pollution are also not generalizable to the South African population.

This project only investigates joint effects of air pollution on combined hospital data and did not explore the effects on different groups, e.g. sex and different age groups.

Although imputation was used to address the issue of missing data, the imputation could have introduced measurement error, from the assumption that the measured air pollution and meteorological data were the same across the few air monitoring sites in the VTAPA.

10.3. GENERAL RECOMMENDATIONS

Overall, the results provide baseline evidence that unsupervised ML has the potential to be useful in joint effects epidemiological studies and source apportionment studies. However, more studies are needed to investigate this further. Introducing ML methods in air pollution studies could help to simplify air pollution studies and, perhaps, increase information on air pollution in Africa. Although air pollution and its effects may not be a priority in Africa due to pre-existing societal issues,[35] increasing the ways in which researchers can conduct air pollution studies could be beneficial in promoting more African-based studies and obtaining results that better represent the African population.

Further in-depth studies should include public health students at different tertiary levels, to broaden the scope of the how AI is viewed in public health, in order to produce more generalizable results. The emissions standards for controlled fuels and specifications for fuel compositions has not yet been implemented at national, provincial or municipal level, according to the national air quality act.[36] It is pertinent that the South African government, at all levels, not only enforces these standards, but also revise air quality management plans in VTAPA. Furthermore, local hospitals in South Africa should be encouraged and enabled to capture data electronically. This will provide additional hospital data that can be used in air pollution epidemiology studies, and provide more generalizable findings for the population.

As this project produces baseline results from the use of unsupervised ML methods for air pollution joint effects associative studies and source apportionment, there is a clear need to run more studies using the same methods within different areas of the country. The findings from different areas will enable comparisons within the results and possibly lead to develop unsupervised ML methods as useful tools in air pollution epidemiology studies. Further studies should also encourage collaboration with data scientists to guide public health researchers on modifying the unsupervised ML

methods to improve and enhance the use of ML in public health research. The results show that even though there is an increased use of AI and ML in medicine, the application of AI and ML still needs to be improved if it is to be applied public health research. By performing similar studies a better argument can be determined on whether AI and ML have a space in public health research. If there is a space for AI and ML, the increased studies can determine how these methods can be used to improve public health research.

10.4. REFERENCES

1. Department of Health. National digital health strategy for South Africa (2019-2024). Pretoria, South Africa. 2019.

2. Kim MC, Okada K, Ryner AM, Amza A, Tadesse Z, Cotter SY, et al. Sensitivity and specificity of computer vision classification of eyelid photographs for programmatic trachoma assessment. Plos One. 2019; 14(2):e0210463. doi:10.1371/journal.pone.0210463.

3. Mbunge E, Batani J, Gaobotse G, Muchemwa B. Virtual healthcare services and digital health technologies deployed during coronavirus disease 2019 (COVID-19) pandemic in South Africa: A systematic review. Global Health Journal (Amsterdam, Netherlands). 2022; 6(2):102-13. doi:10.1016/j.glohj.2022.03.001.

4. Moyo S, Doan TN, Yun JA, Tshuma N. Application of machine learning models in predicting length of stay among healthcare workers in underserved communities in South Africa. Human Resources for Health. 2018; 16(1):68. doi:10.1186/s12960-018-0329-1.

5. van Heerden A, Young S, Park CS. Use of social media big data as a novel HIV surveillance tool in South Africa. PLoS One. 2020; 15(10) doi:10.1371/journal.pone.0239304.

6. Mehta N, Harish V, Bilimoria K, Morgado F, Ginsburg S, Law M, et al. Knowledge and attitudes on artificial intelligence in healthcare: A provincial survey study of medical students. medRxiv. 2021; 10(1) doi:10.15694/mep.2021.000075.1.

7. Castagno S, Khalifa M. Perceptions of artificial intelligence among healthcare staff: A qualitative survey study. Frontiers in Artificial Intelligence. 2020; 3:578983. doi:10.3389/frai.2020.578983.

8. Stai B, Heller N, McSweeney S, Rickman J, Blake P, Vasdev R, et al. Public perceptions of artificial intelligence and robotics in medicine. Journal of endourology. 2020; 34(10):1041-8. doi:10.1089/end.2020.0137.

9. Yüzbaşıoğlu E. Attitudes and perceptions of dental students towards artificial intelligence. Journal of Dental Education. 2021; 85(1):60-8. doi:10.1002/jdd.12385.

10. Oosterhoff JH, Doornberg JN. Artificial intelligence in orthopaedics: False hope or not? A narrative review along the line of gartner's hype cycle. EFORT Open Reviews. 2020; 5(10):593-603.

11. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthcare Journal. 2019; 6(2):94-8. doi:10.7861/futurehosp.6-2-94.

12. Kolachalama VB, Garg PS. Machine learning and medical education. NPJ Digital Medicine. 2018; 1 doi:10.1038/s41746-018-0061-1.

13. McCoy LG, Nagaraj S, Morgado F, Harish V, Das S, Celi LA. What do medical students actually need to know about artificial intelligence? NPJ Digital Medicine. 2020; 3(1) doi:10.1038/s41746-020-0294-7.

14. Gómez-Carracedo MP, Andrade JM, López-Mahía P, Muniategui S, Prada D. A practical comparison of single and multiple imputation methods to handle complex missing data in air quality datasets. Chemometrics and Intelligent Laboratory Systems. 2014; 134:23-33. doi:10.1016/j.chemolab.2014.02.007.

15. Hadeed SJ, O'Rourke MK, Burgess JL, Harris RB, Canales RA. Imputation methods for addressing missing data in short-term monitoring of air pollutants. Science of the Total Environment. 2020; 730 doi:10.1016/j.scitotenv.2020.139140.

16. Hajmohammadi H, Heydecker B. Multivariate time series modelling for urban air quality. Urban Climate. 2021; 37 doi:10.1016/j.uclim.2021.100834.

17. Van Buuren S, Groothuis-Oudshoorn K. Mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011; 45(1):1-67.

18. Gass K, Klein M, Chang HH, Flanders WD, Strickland MJ. Classification and regression trees for epidemiologic research: An air pollution example. Environmental Health. 2014; 13(1):1-10. doi:10.1186/1476-069X-13-17.

19. Gass K, Klein M, Sarnat SE, Winquist A, Darrow LA, Flanders WD, et al. Associations between ambient air pollutant mixtures and pediatric asthma emergency department visits in three cities: A classification and regression tree approach. Environmental Health. 2015; 14:58. doi:10.1186/s12940-015-0044-5.

20. Department of Environment Forestry and Fisheries. Draft second generation air quality management plan for Vaal Triangle Airshed Priority Area. Pretoria: Government Gazette. 2020.

21. Mistry S, Riches NO, Gouripeddi R, Facelli JC. Environmental exposures in machine learning and data mining approaches to diabetes etiology: A scoping review. Artificial Intelligence In Medicine. 2023; 135 doi:10.1016/j.artmed.2022.102461.

22. Riches NO, Gouripeddi R, Facelli JC. 73432 assessment of multi-pollutant ambient air composition on type 2 diabetes mellitus using machine learning. Journal of Clinical and Translational Science. 2021; 5(s1):46.

23. Riches NO, Gouripeddi R, Payan-Medina A, Facelli JC. K-means cluster analysis of cooperative effects of CO, NO2, O3, PM2.5, PM10, and SO2 on incidence of type 2 diabetes mellitus in the US. Environmental Research. 2022; 212(Part B) doi:10.1016/j.envres.2022.113259.

24. Adeyemi A, Molnar P, Boman J, Wichmann J. Source apportionment of fine atmospheric particles using positive matrix factorization in Pretoria, South Africa. Environmental Monitoring and Assessment. 2021; 193(11):716. doi:10.1007/s10661-021-09483-3.

25. Howlett-Downing C. The association between sources of air pollution and respiratory health in Pretoria, South Africa: University of Pretoria; 2022.

26. Howlett-Downing C, Boman J, Molnár P, Shirinde J, Wichmann J. PM2.5 chemical composition and geographical origin of air masses in Pretoria, South Africa. Water, Air, & Soil Pollution. 2022; 233(7) doi:10.1007/s11270-022-05746-y.

27. Han F, Kota SH, Wang Y, Zhang H. Source apportionment of PM2.5 in Baton Rouge, Louisiana during 2009–2014. Science of The Total Environment. 2017; 586:115-26. doi:https://doi.org/10.1016/j.scitotenv.2017.01.189.

28. Kim S, Kim TY, Yi SM, Heo J. Source apportionment of PM2.5 using positive matrix factorization (PMF) at a rural site in korea. Journal of Environmental Management. 2018; 214:325-34. doi:10.1016/j.jenvman.2018.03.027.

29. Mooibroek D, Schaap M, Weijers EP, Hoogerbrugge R. Source apportionment and spatial variability of PM2.5 using measurements at five sites in the netherlands. Atmospheric Environment. 2011; 45(25):4180-91. doi:https://doi.org/10.1016/j.atmosenv.2011.05.017.

30. Ogundele LT, Owoade OK, Olise FS, Hopke PK. Source identification and apportionment of PM2.5 and PM2.5−10 in iron and steel scrap smelting factory environment using PMF, PCFA and UNMIX receptor models. Environmental Monitoring and Assessment 2016; 188(10):1-21. doi:10.1007/s10661-016-5585-8.

31. Taghvaee S, Sowlat MH, Mousavi A, Hassanvand MS, Yunesian M, Naddafi K, et al. Source apportionment of ambient PM2.5 in two locations in central tehran using the positive matrix factorization (PMF) model. Science of the Total Environment. 2018; 628-629:672-86. doi:10.1016/j.scitotenv.2018.02.096.

32. Austin E, Coull BA, Zanobetti A, Koutrakis P. A framework to spatially cluster air pollution monitoring sites in US based on the PM2.5 composition. Environment International. 2013; 59:244-54. doi:10.1016/j.envint.2013.06.003.

33. Khorshidi N, Parsa M, Lentz DR, Sobhanverdi J. Identification of heavy metal pollution sources and its associated risk assessment in an industrial town using the k-means clustering technique. Applied Geochemistry. 2021; 135:105113.

34. Kumar V, Sahu M, Biswas P. Source apportionment of particulate matter by application of machine learning clustering algorithms. Aerosol and Air Quality Research. 2022; 22(3):210240. doi:10.4209/aaqr.210240.

35. Okello G, Nantanda R, Awokola B, Thondoo M, Okure D, Tatah L, et al. Air quality management strategies in Africa: A scoping review of the content, context, co-benefits and unintended consequences. Environment International. 2022:107709.

36. Department of Environmental Affairs. National environmental management: Air quality act, 2004 (act no. 39 of 2004) national ambient air quality standards Government Gazette. 2009.

APPENDIX 1: ETHICS APPROVAL FOR QUESTIONNAIRE

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

Institution: The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

- FWA 00002567, Approved dd 18 March 2022 and Expires 18 March 2027.
- IORG #: IORG0001762 OMB No. 0990-0278 Approved for use through August 31, 2023.

Faculty of Health Sciences **Research Ethics Committee**

23 March 2023

Approval Certificate
Annual Renewal

Dear Prof J Wichmann,

Ethics Reference No.: 171/2022 – Line 1
Title: Attitudes and perceptions of postgraduate students towards use of artificial intelligence in public health

The **Annual Renewal** as supported by documents received between 2023-03-01 and 2023-03-15 for your research, was approved by the Faculty of Health Sciences Research Ethics Committee on 2023-03-15 as resolved by its quorate meeting.

Please note the following about your ethics approval:

- Renewal of ethics approval is valid for 1 year, subsequent annual renewal will become due on 2024-03-23.
- Please remember to use your protocol number (171/2022) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, monitor the conduct of your research, or suspend or withdraw ethics approval.

Ethics approval is subject to the following:

- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

On behalf of the FHS REC, Professor C Kotzé
MBChB, DMH, MMed(Psych), FCPsych, Phd
***Acting Chairperson**: Faculty of Health Sciences Research Ethics Committee*

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health)

Research Ethics Committee
Room 4-60, Level 4, Tswelopele Building
University of Pretoria, Private Bag x323
Gezina 0031, South Africa
Tel +27 (0)12 356 3084
Email: deepeka.behari@up.ac.za
www.up.ac.za

Fakulteit Gesondheidswetenskappe
Lefapha la Disaense tša Maphelo

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

APPENDIX 2: STUDY QUESTIONNAIRE

We, Professors Janine Wichmann and Liz Wolvaardt along with Drs Mari van Wyk and Sean Patrick, would like to understand the perception and attitude of students enrolled in the Postgraduate Diploma in Public Health regarding Artificial Intelligence (AI). The results can be applied to guide policy development and education curriculum changes, which may initiate an interest in emerging technologies such as AI. Studies on this are lacking in Africa.

We intend to submit the findings to a journal for publication.

This survey will consist of a series of short questions related to the study that will include open-ended questions and rating answers on a scale.

There are three (3) sections in the questionnaire. (A): Demographics section, (B): Knowledge of Artificial Intelligence and (C) Perceptions of Artificial Intelligence

The survey will take about 15-20 minutes to complete.

Please contact Prof Janine Wichmann, the principal investigator, if you need any clarifications regarding the study questionnaire.

Please click on **NEXT** to proceed to the survey (Qualtrics will be used to collect the data: https://www.qualtrics.com/uk/?rid=ip&prevsite=en&newsite=uk&geo=ZA&geomatch=uk).

***This questionnaire was adapted from this study:**

Mehta N, Harish V, Bilimoria K, Morgado F, Ginsburg S, Law M, et al. Knowledge and attitudes on artificial intelligence in healthcare: A provincial survey study of medical students. medRxiv. 2021; 10(1) doi:10.15694/mep.2021.000075.1

Section A: Demographics

1. What is your gender?

a) Female
b) Male

2. What is your age category?

a) 20-25
b) 26-30
c) 31-35
d) 36-40
e) 41-45
f) 46-50
g) 51-55
h) 56-60
i) 61 or older

3. In what province do you live?

j) Eastern Cape
k) Free State
l) Gauteng
m) Kwa-Zulu Natal
n) Limpopo
o) Mpumalanga
p) Northern Cape
q) North-West
r) Western Cape
s) Do not live in South Africa

5. Do you study full-time or part-time?

a) Full time
b) Part time

6. What is the highest qualification you have obtained?

a) Bachelors
b) Masters
c) Doctorate
d) Other (please specify):_____

7. a What was the programme of your highest qualification? e.g. Masters in Chemistry

_____

7. b What year did you obtain your highest qualification?

_____

8. Do you have a background in computer science?
   a) Yes
   b) No

9. Please describe your training in computer science

   Training/course: _____

   Duration: _____

10. Have you attended or viewed any talks or lectures on artificial intelligence?
    a. No
    b. Yes
       i. Approximately how many? _____

11. Do you have any training in computer programming/coding?
    a. No
    b. Yes
       i. Where did you acquire this training? _____

Section B: Knowledge of Artificial Intelligence

To what extent do you agree with the following statements?

1. I understand what the term “artificial intelligence” means.
    a. Strongly disagree
    b. Disagree
    c. Agree
    d. Strongly agree

2. I understand what the term “machine learning” means.
    a. Strongly disagree
    b. Disagree
    c. Agree
    d. Strongly agree

3. I understand what the term “neural network” means.
    a. Strongly disagree
    b. Disagree
    c. Agree
    d. Strongly agree

4. I understand what the term “deep learning” means.
    a. Strongly disagree
    b. Disagree
    c. Agree
    d. Strongly agree

5. I know what an algorithm is in the context of computer science.
    a. Strongly disagree
    b. Disagree
    c. Agree
    d. Strongly agree

Section C: Perceptions of Artificial Intelligence

Please read this definition: Artificial Intelligence is the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, and decision-making.

In your opinion, **what is the likelihood that artificial intelligence enabled technologies will ever be able to replace human beings in performing these tasks** as well as or better than the average healthcare professional?

**Note: each time you select "extremely likely" or "likely", it will prompt a question regarding the timeline in which you see these innovations happening.*

Section C1: Individual Patient Care

1. Provide patients with preventative health recommendations (e.g. exercise, diet, wellness).
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

2. Analyse patient information to reach a diagnosis.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

3. Analyse patient information to establish possible prognosis.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

4. Read and interpret diagnostic imaging (such as X rays).
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

5. Evaluate when to refer patients to other health professionals.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

6. Formulate personalised treatment plans for patients.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

7. Formulate personalised medication prescriptions for patients.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

8. Provide empathetic care to patients.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

9. Monitor patient compliance to prescribed medications, exercise and dietary recommendations.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

10. Provide psychiatric/personal counselling.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

11. Perform surgery (e.g. robotic surgery).
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

Section C2: Health Systems

12. Provide documentation (e.g., update medical records) about patients.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

13. Assist hospitals in capacity planning and human resource management.
    a) Extremely unlikely
    b) Unlikely
    c) Likely

d) Extremely likely

14. Provide recommendations for quality improvement in practices/hospitals.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

Section C3: Population Health

15. Conduct population health surveillance and outbreak prevention.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

16. Select the best population health interventions.
    a) Extremely unlikely
    b) Unlikely
    c) Likely
    d) Extremely likely

Section C4: Artificial Intelligence impact on Public Health careers

17. Artificial Intelligence will reduce the number of jobs available to me.
    a) Strongly agree
    b) Agree
    c) Disagree
    d) Strongly disagree

18. Artificial Intelligence will reduce the number of jobs in certain public health specialties more than others
    a) Strongly agree
    b) Agree
    c) Disagree
    d) Strongly disagree

19. Artificial Intelligence will/already did impact my choice of public health specialty selection.
    a) Strongly agree
    b) Agree
    c) Disagree
    d) Strongly disagree

Section C5: Artificial Intelligence Ethics

20. AI in public health will raise new ethical challenges.
    a) Strongly agree
    b) Agree
    c) Disagree
    d) Strongly disagree

21. AI in public health will raise new social challenges.
    a) Strongly agree
    b) Agree
    c) Disagree
    d) Strongly disagree

22. AI in public health will raise new challenges around health equity.
    a) Strongly agree
    b) Agree
    c) Disagree
    d) Strongly disagree

23. The South African healthcare system is currently well prepared to deal with challenges related to AI.
    a) Strongly agree
    b) Agree
    c) Disagree
    d) Strongly disagree

Section C6: Artificial Intelligence and public health education

24. My public health education is adequately preparing me for working alongside AI tools
    a) Strongly agree
    b) Agree
    c) Disagree
    d) Strongly disagree

25. My public health training should include training on AI competencies (e.g. what is AI, how will it impact us, what are the challenges it raises).
    a) Strongly agree
    b) Agree
    c) Disagree
    d) Strongly disagree

26. Every public health student should be required to receive training in AI competencies.
    a) Strongly agree
    b) Agree
    c) Disagree
    d) Strongly disagree

27. Training in AI competencies should begin as a:
    a) Undergraduate student regardless of study area
    b) Postgraduate student regardless of study area
    c) No training in AI competencies is necessary

28. Do you have any comments or concerns (good and/or bad) on the topic of AI in public health? (word limit: 100)

29. What is your general reflection on what AI will look like in 5 years within your department? (word limit: 100)

APPENDIX 3: CONSENT FORM

ICD 1A

PARTICIPANT'S INFORMATION & INFORMED CONSENT DOCUMENT

STUDY TITLE: Attitudes and perceptions of postgraduate students towards use of artificial intelligence in public health

PRINCIPAL INVESTIGATOR: Prof Janine Wichmann

COLLABORATORS: Prof Liz Wolvaardt, Dr Mari van Wyk, Dr Sean Patrick

INSTITUTION: School of Health Systems and Public Health, University of Pretoria

DAYTIME AND AFTER HOURS TELEPHONE NUMBER(S):
Daytime telephone numbers: 012 356 3259/ 082 693 7275
Email address: Janine.wichmann@up.ac.za

Dear Prospective Participant

1) INTRODUCTION

You are invited to volunteer in this research study. We are doing research for non-degree purposes at the School of Health Systems and Public Health, University of Pretoria. This information in this document is to help you to decide if you would like to participate. Before you agree to take part in this study you should fully understand what is involved. If you have any questions, which are not fully explained in this document, do not hesitate to ask the principal investigator, Prof Janine Wichmann. You should not agree to take part unless you are completely happy about all the procedures involved.

2) THE NATURE AND PURPOSE OF THIS STUDY

Artificial Intelligence (AI) is an umbrella term used to describe the multidisciplinary approach to use statistical, mathematical, and computer sciences to simulate intelligent behaviour. AI has been a very useful tool in many sectors of society, including medicine and public health. Using the various types of AI, the healthcare sector has been aided with tools to help in for example administration, diagnosis making, treatment planning, medical record mining and drug creation.

However, there are numerous limitations and challenges of AI that need to be scrutinised before formulating and interpreting the AI generated results. It is important that future public health experts are well aware of the opportunities and challenges linked to AI. The aim of this study is to understand the perception and attitude of students enrolled in the online Postgraduate Diploma in Public Health regarding AI.

3) EXPLANATION OF PROCEDURES AND WHAT WILL BE EXPECTED FROM PARTICIPANTS.

There are three (3) sections in the questionnaire, namely (A): Demographics, (B): Knowledge of Artificial Intelligence and (C) Perceptions of Artificial Intelligence. The online questionnaire will be made available after you have given your consent.

The questionnaire will take 15-20 minutes to complete. Should you not feel comfortable to take part in or withdraw from completing the questionnaire you are free to do so at any point in the study.

4) POSSIBLE RISKS AND DISCOMFORTS INVOLVED

There are no medical risks associated with the study. Therefore, there is no significant risk to the participant.

5) POSSIBLE BENEFITS OF THIS STUDY

Although you may not benefit directly, your responses will assist in helping us understand your perceptions and attitude regarding AI. The results can be applied to guide policy development and education curriculum changes, which may initiate an interest in emerging technologies such as AI among postgraduate public health students. Studies on this topic are lacking in Africa.

6) COMPENSATION

Your participation will be entirely voluntary. You can refuse to take part in the study and can withdraw at any point you feel uncomfortable. Withdrawal from the study will not affect you or your treatment in any way.

7) YOUR RIGHTS AS A RESEARCH PARTICIPANT

Your participation in this study is entirely voluntary and you can refuse to participate or stop at any time without stating any reason.

8) ETHICS APPROVAL

The research study protocol was approved by the Faculty of Health Sciences Research Ethics Committee, University of Pretoria (Ref number 171/2022). The contact details of the Committee are: Telephone numbers 012 356 3084 / 012 354 1330 or email: manda.smith@upco.za. The study was structured in accordance with the Declaration of Helsinki (last update: October 2013). A copy of the Declaration may be obtained from the principal investigator Prof Janine Wichmann should you wish to review it.

9) INFORMATION

If you have any questions concerning this study, you should contact: Prof Janine Wichmann, principal investigator, 012 356 3259/082 693 7275

10) CONFIDENTIALITY

All information you provide will be kept strictly confidential. Once we have analysed the information no one will be able to identify you. Research reports and articles in scientific journals will not include any information that may identify you.

11) CONSENT TO PARTICIPATE IN THIS STUDY

- I have also received, read and understood the above written information about the study.
- I have had adequate time to ask questions and I have no objections to participate in this study.
- I am aware that the information obtained in the study, including personal details, will be anonymously processed and presented in the reporting of results.
- I understand that I will not be penalised in any way should I wish to discontinue with the study and that withdrawal will not affect my further treatments.
- I am participating willingly.
- I have received an electronic version of this informed consent agreement and I signed electronically with my name and date.

Participant's name: Will not be collected (use Qualtrics to collect data: https://www.qualtrics.com/uk/?rid=ip&prevsite=en&newsite=uk&geo=ZA&geomatch=uk)

Date: Will not be collected (use Qualtrics to collect data)

	Yes	No
I hereby give my consent to participate in this study	○	○

APPENDIX 4: FIRST AAC APPROVAL

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

19 May 2021

NS Mwase
Student number: 17242496
Email: nandisisa@gmail.com

Dear Nandi

SUCCESSFUL PHD PROTOCOL DEFENCE

SUPERVISOR: Associate Prof J Wichmann, Faculty of Health Sciences, SHSPH
SUPERVISOR: Prof Washington Junger (Rio de Janeiro State University)

TITLE: Artificial Intelligence applications in air pollution epidemiology in South Africa: supervised and unsupervised machine learning

You had a successful virtual PhD protocol oral defence on **10 February 2021** at the School of Health Systems and Public Health, chaired by Dr Joyce Shirinde. All three reviewers indicated that revisions were required to your protocol and that they recommended your protocol for PhD Epidemiology [10260409] studies at the School of Health Systems and Public Health. Herewith the names of reviewers:

Internal Reviewer: Prof Inger Fabris-Rotelli, Department of Statistics, University of Pretoria
External Reviewer: Prof Patrick De Boever MSE, PhD, Hasselt University / University of Antwerp
External External Reviewer: Justin Lessler, Johns Hopkins Bloomberg School of Public Health

You addressed all the reviewers feedback and the SHSPH Academic Advisory Committee approved your updated protocol. You may now proceed and submit for ethical approval. The AAC strongly recommends that you also attach all three review forms in case the Faculty Research Ethics Committee questions the quality of your PhD protocol.

Yours sincerely

Dr Joyce Shirinde
Chairperson
SHSPH Academic Advisory Committee

Prof K Voyi
Chairperson:
School of Health Systems and Public Health

Cc: Prof J Wichmann

AAC members: Dr Joyce Shirinde (Chairperson), Prof Janine Wichmann, Dr Neo Ledibane, Prof Halina Rollin, Dr Flavia Senkubuge, Annette Welman (student admin), Dr S Patrick, Dr A Musekiwa, Dr T Kruger, Dr E Webb, Mrs Kathy Pieterse

AAC email address: kathy.pieterse@up.ac.za
http://shsph.up.ac.za
www.up.ac.za

APPENDIX 5: SECOND AAC APPROVAL LETTER (AFTER TITLE CHANGE)

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

18 June 2021

NS Mwase
Student number: 17242496
Email: nandisisa@gmail.com

Dear Nandi

SUCCESSFUL PHD PROTOCOL DEFENCE

SUPERVISOR: Associate Prof J Wichmann, Faculty of Health Sciences, SHSPH
CO-SUPERVISOR: Prof Washington Junger (Rio de Janeiro State University)

TITLE: Supervised and unsupervised machine learning in air pollution epidemiology in South Africa: Artificial intelligence subset application

You had a successful virtual PhD protocol oral defence on **10 February 2021** at the School of Health Systems and Public Health, chaired by Dr Joyce Shirinde. All three reviewers indicated that revisions were required to your protocol and that they recommended your protocol for PhD Epidemiology [10260409] studies at the School of Health Systems and Public Health. Herewith the names of reviewers:

Internal Reviewer: Prof Inger Fabris-Rotelli, Department of Statistics, University of Pretoria
External Reviewer: Prof Patrick De Boever MSE, PhD, Hasselt University / University of Antwerp
External External Reviewer: Justin Lessler, Johns Hopkins Bloomberg School of Public Health

You addressed all the reviewers' feedback and the SHSPH Academic Advisory Committee approved your updated protocol. You may now proceed and submit for ethical approval. The AAC strongly recommends that you also attach all three review forms in case the Faculty Research Ethics Committee questions the quality of your PhD protocol.

Yours sincerely

Dr Joyce Shirinde
Chairperson
SHSPH Academic Advisory Committee

Prof K Voyi
Chairperson:
School of Health Systems and Public Health

Cc: Prof J Wichmann

AAC members: Dr Joyce Shirinde (Chairperson), Prof Janine Wichmann, Dr Neo Ledibane, Prof Halina Rollin, Dr Flavia Senkubuge, Annette Welman (student admin), Dr S Patrick, Dr A Musekiwa, Dr T Kruger, Dr E Webb, Mrs Kathy Pieterse

AAC email address: kathy.pieterse@up.ac.za
http://shsph.up.ac.za
www.up.ac.za

APPENDIX 6: FIRST ETHICS APPROVAL FOR PHD STUDY

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

Institution: The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

- FWA 00002567, Approved dd 22 May 2002 and Expires 03/20/2022.
- IORG #: IORG0001762 OMB No. 0990-0279 Approved for use through February 28, 2022 and Expires: 03/04/2023.

Faculty of Health Sciences **Research Ethics Committee**

26 August 2021

Approval Certificate
New Application

Dear Miss NS Mwase

Ethics Reference No.: 433/2021
Title: Supervised and unsupervised machine learning in air pollution epidemiology in South Africa: Artificial Intelligence subset application

The **New Application** as supported by documents received between 2021-07-21 and 2021-08-25 for your research, was approved by the Faculty of Health Sciences Research Ethics Committee on 2021-08-25 as resolved by its quorate meeting.

Please note the following about your ethics approval:

- Ethics Approval is valid for 1 year and needs to be renewed annually by 2022-08-26.
- Please remember to use your protocol number (433/2021) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, monitor the conduct of your research, or suspend or withdraw ethics approval.

Ethics approval is subject to the following:

- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

On behalf of the FHS REC, Professor Werdie (CW) Van Staden
MBChB, MMed(Psych), MD, FCPsych(SA), FTCL, UPLM
***Chairperson**: Faculty of Health Sciences Research Ethics Committee*

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health)

APPENDIX 7: SECOND ETHICS APPROVAL LETTER

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

Institution: The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.
- FWA 00002567, Approved dd 18 March 2022 and Expires 18 March 2027.
- IORG #: IORG0001762 OMB No. 0990-0278 Approved for use through August 31, 2023.

Faculty of Health Sciences **Research Ethics Committee**

19 January 2023

Approval Certificate
Annual Renewal

Dear Miss NS Mwase,

Ethics Reference No.: 433/2021 – Line 2
Title: Supervised and unsupervised machine learning in air pollution epidemiology in South Africa: Artificial Intelligence subset application

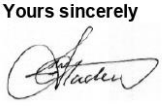
The **Annual Renewal** as supported by documents received between 2022-12-02 and 2023-01-18 for your research, was approved by the Faculty of Health Sciences Research Ethics Committee on 2023-01-18 as resolved by its quorate meeting.

Please note the following about your ethics approval:
- Renewal of ethics approval is valid for 1 year, subsequent annual renewal will become due on 2024-01-19.
- Please remember to use your protocol number (433/2021) on any documents or correspondence with the Research Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, monitor the conduct of your research, or suspend or withdraw ethics approval.

Ethics approval is subject to the following:
- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely



On behalf of the FHS REC, Professor Werdie (CW) Van Staden
MBChB, MMed(Psych), MD, FCPsych(SA), FTCL, UPLM
***Chairperson**: Faculty of Health Sciences Research Ethics Committee*

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health)

Research Ethics Committee
Room 4-60, Level 4, Tswelopele Building
University of Pretoria, Private Bag x323
Gezina 0031, South Africa
Tel +27 (0)12 356 3084
Email: deepeka.behari@up.ac.za
www.up.ac.za

Fakulteit Gesondheidswetenskappe
Lefapha la Disaense tša Maphelo

APPENDIX 8: FINAL ETHICS APPROVAL LETTER

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Health Sciences

Institution: The Research Ethics Committee, Faculty Health Sciences, University of Pretoria complies with ICH-GCP guidelines and has US Federal wide Assurance.

- FWA 00002567, Approved dd 18 March 2022 and Expires 18 March 2027.
- IORG #: IORG0001762 OMB No. 0990-0278 Approved for use through August 31, 2023.

Faculty of Health Sciences **Research Ethics Committee**

18 May 2023

Approval Certificate
Amendment

Dear Miss NS Mwase

Ethics Reference No.: 433/2021 – Line 3
Title: Unsupervised machine learning in air pollution epidemiology in South Africa: Artificial Intelligence subset application

The **Amendment** as supported by documents received between 2023-04-18 and 2023-05-17 for your research, was approved by the Faculty of Health Sciences Research Ethics Committee on 2023-05-17 as resolved by its quorate meeting.

Please note the following about your ethics approval:

- Please remember to use your protocol number (**433/2021**) on any documents or correspondence with the Reseasrch Ethics Committee regarding your research.
- Please note that the Research Ethics Committee may ask further questions, seek additional information, require further modification, monitor the conduct of your researc, or suspend or withdraw ethics approval.

Ethics approval is subject to the following:

- The ethics approval is conditional on the research being conducted as stipulated by the details of all documents submitted to the Committee. In the event that a further need arises to change who the investigators are, the methods or any other aspect, such changes must be submitted as an Amendment for approval by the Committee.

We wish you the best with your research.

Yours sincerely

On behalf of the FHS REC, Dr R Sommers
MBChB, MMed (Int), MPharmMed, PhD
Deputy Chairperson *of the Faculty of Health Sciences Research Ethics Committee, University of Pretoria*

The Faculty of Health Sciences Research Ethics Committee complies with the SA National Act 61 of 2003 as it pertains to health research and the United States Code of Federal Regulations Title 45 and 46. This committee abides by the ethical norms and principles for research, established by the Declaration of Helsinki, the South African Medical Research Council Guidelines as well as the Guidelines for Ethical Research: Principles Structures and Processes, Second Edition 2015 (Department of Health).

Research Ethics Committee
Room 4-60, Level 4, Tswelopele Building
University of Pretoria, Private Bag x323
Gezina 0031, South Africa
Tel +27 (0)12 356 3084
Email: deepeka.behari@up.ac.za
www.up.ac.za

Fakulteit Gesondheidswetenskappe
Lefapha la Disaense tša Maphelo

APPENDIX 9: DECLARATION FROM PROOFREADER

DECLARATION FROM PROOFREADER

This is to state that the PhD (Epidemiology) submitted to me by Miss Nandi Sisasenkosi Mwase (student no u17242496) of the University of Pretoria, South Arica, has been proofread by me according to the tenets of academic discourse.

Mrs Dené Hees, MA (Drama and Film Studies); Cert. of Copy-editing and Proofreading (SAWC); TEFL cert. (Level 5, QUALIFI).

98 Papillon

624 Farm Road

Equestria, Pretoria

Gauteng

0184

060 970 8425

denejvrb@gmail.com

26 April 2023