UNIVERSITI TEKNOLOGI MARA

ASSESSING POTENTIAL DOMINANT FACTORS OF TRAFFIC RELATED AIR POLLUTION (TRAP) USING FACTOR AND CLUSTER ANALYSIS BASED ON DIFFERENT LOCATION CATEGORIES

HARITH FARHAN BIN HAMDAN

MSc

February 2022

UNIVERSITI TEKNOLOGI MARA

ASSESSING POTENTIAL DOMINANT FACTORS OF TRAFFIC RELATED AIR POLLUTION (TRAP) USING FACTOR AND CLUSTER ANALYSIS BASED ON DIFFERENT LOCATION CATEGORIES

HARITH FARHAN BIN HAMDAN

Dissertation submitted in partial fulfilment

of the requirements for the degree of

Master of Science in
(Applied Statistic)

Faculty of Computer and Mathematical Sciences

February 2022

AUTHOR'S DECLARATION

I declare that the work in this thesis was carried out in accordance with the regulations of Universiti Teknologi MARA. It is original and is the results of my own work, unless otherwise indicated or acknowledged as referenced work. This thesis has not been submitted to any other academic institution or non-academic institution for any degree or qualification.

I, hereby, acknowledge that I have been supplied with the Academic Rules and Regulations for Postgraduate, Universiti Teknologi MARA, regulating the conduct of my study and research.

Name of Student	:	Harith Farhan bin Hamdan
Student I.D. No.	:	2020841058
Programme	:	Master of Science (Applied Statistic) – CS702
Faculty	:	Computer and Mathematical Science
Proposal Title	:	Assessing Potential Dominant Factors of Traffic Related Air Pollution (TRAP) using factor and cluster analysis based on different location categories
Signature of Student	:	....
Date	:	February 2022

APPROVED BY:

DR. HASFAZILAH BINTI AHMAT

Supervisor

Faculty of Science Computer and Mathematical Sciences

Universiti Teknologi Mara

ABSTRACT

Traffic related air pollution (TRAP) is an air pollution that comes from vehicles. Most of the greenhouse gas emission are from transportation such as petroleum-based fuels from vehicles. The strong growth of economic and other factors leads to the increase of transport sector among people hence increasing TRAP indirectly. This is very concerning due to the health impact that can affect towards the society such as lung cancer, stroke and other more. Although there is no denying that air pollution studies were widely used among researcher but there is still lack of study which focusses on TRAP itself particularly in Malaysia. Therefore, the study focusses on assessing the potential dominant of traffic related air pollution using both factor and cluster analysis to obtain more information on the pollution itself. The dataset was obtained from Department of Environment (DOE) and the variables involved are meteorological factors (wind speed, temperature, humidity)and other gaseous concentration that relates with traffic emission (NO_2,NOx,O_3,CO,PM_{10},$PM_{2.5}$). Petaling Jaya (industrial area), Shah Alam (urban area), Banting (sub-urban area) and Jerantut (background station) were the four categories of location used in this study. Factor analysis was used to examine the underlying or latent relationship between the potential of dominant factors in transportation related air pollution while cluster analysis was used to classify and compare the cluster or group made from in the context of transportation along with the result obtain by factor analysis. Spearman correlation analysis also were used to illustrate the strength of Carbon Monoxide (CO) among other factors based on different location. From the result, it was identified that differences in categories of location brings different outcome towards air pollution. Industrial area has the highest impact when comes to traffic emission meanwhile Jerantut has the lowest impact respectively. Each pollutant differs by each location as its main sources of traffic emission. For example, Shah Alam main pollutant is NO_2 while Jerantut main pollutant for traffic emission is CO. CO were mostly highly correlated with another pollutant which includes in traffic emission. Finally, cluster and factor analysis produce the same group of pollutant in Jerantut and Shah Alam meanwhile differs with Petaling Jaya and Banting. Although, Banting and Petaling differs among each other, it is just only act as a comparison that could be a reference for the future study.

ACKNOWLADGEMENT

Foremost, our greatest and ultimate gratitude to Allah S.W.T the Most Merciful for giving us the opportunity to complete this long and challenging project promptly. Aside from that, I would like to express my gratitude and appreciation to my supervisor, Hasfazilah Ahmat (DR), for her continual advice, assistance, and support throughout the development of this research. Without her assistance, I may not be able to complete the course work on time.

Furthermore, I want to express my thanks to staff member of Department of Environment, Malaysia (DOE) for providing a complete set of data that is related towards the topic in finishing my study successfully.

Next, I want to thank my classmates for the fascinating talks and all of the fun we had in order to complete this final year project on time during the class session. Thank you for all of your moral support and untold assistance as we work to complete our project respectively.

Finally, I would like to thank my parents, particularly my mother Lili Rozita binti Abdul Manaf, for everything she has done for me during this hectic semester with so many things going on. Without their prayers and support, I may not be able to fulfil my work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

Symbols

mg/m^3	Micrograms per cubic metre
m/s	Miles per second

LIST OF ABBREVIATIONS

Abbreviations	
PCA	Principal Component Analysis
HCA	Hierarchical Clustering Analysis
TRAP	Transport Air Related Pollution
TSP	Total Suspended Particles
WHO	World Health Organization
DOE	Department of Environment
VOCs	Volatile Organic Compound Concentration
FA	Factor Analysis
CA	Cluster Analysis
DA	Discriminant Analysis
O_3	Ozone
CO	Carbon Monoxide
SO_2	Sulphur Dioxide
NO_2	Nitrogen Dioxide
HFC	Hydrocarbon
N_2O	Nitrogen Oxides
PM_{10}	Particulate pollutant that is 10 microns or smaller in size.
$PM_{2.5}$	Particulate pollutant that is 2.5 microns or smaller in size.

CHAPTER ONE
INTRODUCTION

1.1 Research Background

World Health Organization (WHO) classifies the first six contaminants listed "classic" air pollutants as the main factor for air pollution. The six listed are total suspended particles (TSP), Ozone (O_3), Carbon Monoxide (CO), Sulphur Dioxide (SO_2), Nitrogen Dioxide (NO_2) and other air toxins emissions. Most of the greenhouse gas emission are from transportation such as petroleum-based fuels from vehicles. This shows that bad air quality in many cities is due to the increase concentration of human activities which emit damaging pollutants to the environment (Rossi et al., 2020). Malaysia is one of the world's highest rates of auto ownership per capita which resulting in extremely high levels of carbon monoxide pollution in all its main cities (Makmom et al., 2012). The study are focussing on traffic-related air pollution (TRAP) since it is linked to several heart disease problems, particularly hypertension. Traffic related air pollution is an air pollution that were mainly contributed by transportation such as the movement of goods through automobiles, trucks and other more.

Malaysia was ranked in the 50th place amongst all the countries of the world, with a particle that have diameter less than 2.5 micrometres ($PM_{2.5}$) rating of 19.36 $\mu g/m^3$, putting its yearly average into the 'moderately' polluted range in 2019. However, Department of Environment (DOE) of Malaysia is always busy enacted to several number of policies in eliminating and reducing pollution in accordance with their objectives of the Environmental Quality Act of 1974. Thus, reducing air particles from transportation are very important in achieving their objectives. Reducing the usage of vehicles for a couple of days on the road could be one of the strategies in reducing transport related air particles effectively (Farda & Balijepalli, 2018). The effectiveness of this strategy is very clear when it shows a significant decreasing in damaging pollutant for the environment in Malaysia during the influence of Malaysian Movement Control Order (Othman & Latif, 2021) . During the time, people were advised to stay at home due to Covid-19. Since then, there are significant reduction of air pollutants in the main cities. For example, the significant reductions in cities such as Kuala Lumpur

(54 %), Ipoh (58 %), Seremban (50 %), Kuantan (54 %), and Kota Kinabalu (54 %) of NO_2 once the MCO was imposed. Table 1.1 provides a better insight on the air quality levels in Malaysia that were published by Department of Environment (DOE) Malaysia.

Table 1.1
Description of air quality levels in Malaysia

AQI	**Air Pollution Level**
0 - 50	Good
51 -100	Moderate
101-150	Unhealthy for Sensitive Groups
151-200	Unhealthy
201-300	Very Unhealthy
300+	Hazardous

Source: Department of Environment Malaysia, 2021

Several factors are reviewed from previous studies which are be focusing on the factors that influence traffic emission which degrade the air quality such as meteorological factors, location of study and air particles related to TRAP. Air pollution concentration have a close relationship with meteorological influence. Some may affect it in a negative way while some also may affect in a positive way in producing good air quality (Cuhadaroglu, Burhan & Dermichi, 1997). Levels of air pollution rose steadily and significantly from local to moderate, massive, and metropolitan areas (Liu et al., 2018). The approach and method that employed in this paper are clearly explained in chapter 2 and 3 such as using factor and cluster analysis related to the interest topic. In view of the problem too, the study is meant to recognize potential dominant factors towards TRAP by its characteristic to identify global warming potential in different category of locations. The purpose of this study is to classify the importance for each type of air pollution in analysing the limitations of current air pollution assessments, especially in the context of transportation since traffic emission is the main contributor towards air quality

1.2 Problem Statement

Transportation is known to emit several harmful air pollutants. Air pollution, particularly near roads and highways has been related to unfavourable health effects on everyone who lives in metropolitan settings. For example, Sulphur dioxide (SO_2), which is released in comparison to the amount of Sulphur in fuel affects lung function in asthmatics and amplifies respiratory symptoms in people who are susceptible to it. The problem grew worse when the World Health Organization announced that the transportation industry had surpassed agriculture as the fourth major source of global GHG emissions which accounts for 25% of total emissions. Nearly 7% of Europe's population was exposed to nitrogen dioxide concentrations over the WHO limit while 44% of Europe's population was exposed to PM_{10} concentrations above the WHO guideline.

Several research have been carried out to investigate air pollution and determine air quality. However, majority of the research is primarily focused on the context of air particle concentrations in industrial areas or the trend of air pollution as a whole. It does not go into detail as to what is going on with transportation-related pollution right now. As a result, it is possible to conclude that most of the studies does not focus extensively on traffic emissions because each research has its own set of goals to pursue. Although there have been a few studies on traffic pollution, they have tended to focus on health issues rather than overall air pollution. Thus, the researcher could not classify and say on what is really happening in assessing the dominant particles on TRAP itself.

Cluster analyses are used in the study to partition of dominant factors of TRAP based on different location. The result obtained from the analysis could bring better insight on how each location react to TRAP since different location has its own characteristic such as the density of population, category of location and other more. Factor analysis also are used in the study to classify the underlying or latent relationship between the potential of dominant factors in Carbon Monoxide (CO) since it is the major factors in TRAP based on previous study meanwhile spearman correlation analysis is used in looking at the strength of the relationship between CO and other pollutants for each factor within different categories of location. Both of this analysis were often used in past studies when relates with air pollution. However, the reason these methods were used on this study is because that most of the previous study related on traffic pollution uses different approach and method to achieve their respective objectives. The approach includes multiple linear regression, Chi-square, Bayesian modelling and other more.

1.3 Research Questions

The research questions for this study on accessing potential dominant factors of traffic related air pollution (TRAP) based on different location categories are as follows:

a) What is the latent relationship between the potential of dominant factors in traffic related air pollution (TRAP) in Shah Alam, Petaling Jaya, Banting and Jerantut?

b) What is the strength of relationship between CO and other pollutants for each factor at different categories of location?

c) What are the classification of dominant factors of traffic related air pollution (TRAP) in Shah Alam, Petaling Jaya, Banting and Jerantut ?

1.4 Research objectives

The research objectives for this study on accessing potential dominant factors of traffic related air pollution (TRAP) based on different location categories are as follows:

a) To examine the underlying or latent relationship between the potential dominant of TRAP pollution within Shah Alam, Petaling Jaya, Banting and Jerantut

b) To determine the strength of the relationship between CO and other pollutants for each factor within different categories of location

c) To classify potential dominant factors of transport related air pollution (TRAP) in Shah Alam, Petaling Jaya, Banting and Jerantut using factor and cluster analysis

1.5 Significance of Study

The finding of the study is beneficial to the academician since the topic covers up the scope on transport related air pollution (TRAP). Academician can define the problem deeply on TRAP pollution and can propose a model or mechanism to overcome air quality problem . The study is significant as it can help the academician to discover the real problem regarding the increasing rate of traffic emission more specifically. The Department of Environment (DOE) of Malaysia is the ministry of Government Malaysia responsible for the any kind of pollution towards the environment such as water pollution, air pollution and other more. Therefore, the study which focus on traffic pollution is related to their fieldwork. The study could ease the Department of Environment (DOE) to monitor and define the actual problem regarding the main contributor towards air pollution. The research is also critical for the society because it is them who are most affected by air pollution caused by transportation. Thus, by spreading awareness on what is currently happening, they can begin taking personal action to improve air quality such as utilising public transportation. Private company also can emphasis more on their workers to work from home rather than making them work in their company if not necessarily Finally, this study also can be useful for town planner in different location of areas. For example, they can construct more stations related to public transportation to encourage society to use less transportation that contributes more pollutants that cause poor air quality. They can evaluate the potential impacts of emissions reduction policies; air pollution impacts and health impacts.

1.6 Scope of the study

The study uses secondary data obtained from the Department of Environment (DOE).There are 4 location groups which are industrial area, urban area, sub-urban area and rural Area. Those location for the research study is monitoring stations in Petaling Jaya, Shah Alam, Banting and Jerantut respectively. Method used for the study is the combination of factor analysis and cluster analysis. Cluster analyses are used to partition the factors based on different categories of location while factor analysis is used to identify the relationship between factors among each location. All data analysis is carried out in SPSS Software, which are made available for use in interpreting complex analyses. The data obtained are data hourly from January 2018 until December 2019

CHAPTER TWO
LITERATURE REVIEW

2.1 Introduction

This chapter talks about a study on air quality and transportation emissions as well as accessing dominant factors influencing TRAP based on past research. Furthermore , this chapter discuss about the variables that the researcher is studying on and the methods used for the study.

2.2 Global Air Quality

Dust fall-out, suspended specific matter, and lead are among the components detected in the surrounding atmosphere along congested roads that are primarily attributed to motor vehicles. Particulate matter (PM), nitrogen oxides (NOx), and volatile organic compounds are all pollutants from these pollution makes a significant contribution to bad air quality (VOCs) and bring bad impact towards society. For example, a study by Boschi (1999) and Kappos et al. (2014) stated that it is important to note that individuals suffering from asthma, pneumonia, diabetes, respiratory and cardiac diseases are particularly sensitive to the effects of $PM_{2.5}$, followed by PM_{10} which are strongly linked to a variety of respiratory system diseases due to their ability to pierce interior spaces. Not only that, polluted air was responsible for 6·4 million deaths worldwide in 2018 which are 2·8 million from household air pollution and 4·2 million from ambient air pollution (Landrigan, 2017). Thus, it shows that air pollution is very dangerous since it could harm humans in terms of health. The study of air quality is becoming popular since carbon emission has become the third largest source of emission due to fast urbanization (Guo et al., 2021). China which is one of the most significant sources of pollution is becoming a major concern particularly in several big cities such as Beijing . Locations with NO_2 concentration levels above the annual limit value (40 g/m^3) are spreading throughout Europe as in the past years . Thus, from here it is possible to say that TRAP is one of the major causes of bad air quality.

2.3 Air Quality in Malaysia

Afroz et al (2003) stated that land transportation, industrial emissions, and open burning sources are the primary contributors of air pollution. Among them, land transportation is one that contributes the most to air pollution. Transportation-related air pollution contributes to poor air quality which have negative effects on the health and welfare in Malaysia. Since Malaysia is still a developing country, study by Azmi et al (2010) showed that vehicles are the primary cause of air pollution in most developing countries' urban areas. Findings shows that the increasing volumes of CO_2 and NO_2 levels reported in Sunway Malaysia have been influenced by the area's saturation traffic volume. The CO_2 concentration was significantly greater in the morning, whereas the NO_2 concentration was higher in the afternoon (Li et al., 2018). Since 1980s, land use areas have gone through a series of urbanisation. Since then, the PM_{10} concentrations in Kuala Lumpur's air have been surpassed by the World Health Organization (WHO) where regulations for ambient PM_{10} concentrations is (50 g/m^3). This is due to high industrial and the strong growth of economic which leads to the increase of transport sector among people (Halim et al., 2020). The preferred option to use private cars is common in Malaysia leading into bad air quality. Air population is increasing in Malaysia especially in urban areas. As a result, polluted air has recently become a major problem in Malaysia. This is due to research indicating that Malaysia's environmental health concerns are mostly driven by air pollution, water pollution, global warming, and ozone layer depletion. Finally, there are several factors that need to take into consideration in focusing on traffic emission which are meteorological factors, air pollutants from vehicles and other air pollutants.

2.4 Meteorological Factors

Study in Malaysia shows that meteorological factors also play an important part in increasing air quality since Malaysia has its hot and wet season depends on its monsoon. Malaysia is subject to two monsoon wind seasons: The Southwest Monsoon and the Northeast Monsoon. The Southwest Monsoon occurs from June to September while the Northeast Monsoon occurs from August to March. Since it originates in China and the north Pacific, the Northeast Monsoon produces more rainfall than the Southwest Monsoon. Meteorological factors such as air temperature, wind velocity, and moisture, were also found to be significantly related to pollution. Study by Chen et al. (2015)

proves that meteorological factors play an important role since many factors influence air pollutant concentrations, including pollution source emissions amounts, geographic location, and meteorological conditions. Meteorological variables must be considered when determining possible sources of transportation-related air pollution since they can lead to poor air quality such as air temperature, humidity, wind direction and speed. Thus, the most significant element in improving long-term air quality was emission reductions such that meteorological conditions play a vital part on it.

2.4.1 **Temperature**

Temperature is measured unit used to measure how hot or cold something is. It is the expression of latent heat which exists in all things and serves as the source of heat. PM_{10} levels remain highly concentrated during the south-west monsoon , whereas CO levels are highly concentrated during the north-east monsoon (Rahman et al., 2015). By regulating the quantity of space heating required, air temperature and sun radiation have an impact on the amount of pollution emitted. Air pollution was lower in coastal regions than in inland areas that were affected by open sea atmospheric transmission and coastal pollutants temperature (Hsu and Cheng,2019). Furthermore, $PM_{2.5}$, PM_{10}, SO_2, NO_2, and CO concentrations were highest in winter and lowest in summer, while O_3 had the opposite seasonal variation (Liu & Yoon, 2019). The duration of sunlight has a secondary impact on Air Quality Index change in Beijing, as it aids in the dissipation of air pollutants. The presence of a cooler air has a significant impact on weather and the diffusion of atmospheric pollutants, obstructing rapid air movement and the diffusion of smoke resulting in reduced air visibility and severe pollution (Zhang, 2019). Finally, extreme heat, heavy precipitation, and low O_3 concentrations all increased the risk of Hand–foot–mouth disease (HFMD). This shows that, cooler air is important in improving air quality based on some constant factors (Ang et al., 2009)

2.4.2 **Humidity**

The concentration of water vapor in the atmosphere is referred as humidity. Water vapor or water in its gaseous state is typically invisible to the naked eye. Humidity gives an indication of precipitation, dew or fog. Higher air humidity stimulates the conduction of atmospheric particulate matter on water vapour causing haze and suspending in the air. It could result in the accumulation of air pollutants and worsening air pollution. Nonetheless, humidity in the air rises enough to generate effective rainfall such that it exerts a scouring and scavenging impact on air contaminants. (Zhang,2019). There are other studies that demonstrate that areas with high humidity tend to be foggy and have a high concentration of contaminants in the atmosphere (Li et al, 2021). Finally, humidity plays part in saving people's life. There has been some biological evidence to back up these findings where if relative humidity is too low, bioaerosol microbial activity is inhabited because microorganisms' metabolism and other physiological activities are suppressed in a dry environment.

2.4.3 **Wind Speed and Direction**

Wind speed is the velocity at which weather-related air flows from one area to another whereas wind direction is the direction of wind origin. The outcome of air pollutants is influenced by air movement and speed. As a result, any study of air pollution should include an examination of local weather patterns (meteorology). Whenever high turbulent winds blow, the pollutant concentrations may disperse fast resulting in reduced pollutant concentrations. Strong wind speed in comparison to average wind speed is by far the most important factor influencing the volatility of the Air Quality Index in Beijing (Zhang ,2019). Extreme wind speed has a major effect on air quality over the next 1–3 days, with the second day having the greatest influence. Besides that, findings show that contribution its assessment of the potential effects of mean wind speed and temperature on the quality of large-scale PM_{10} (Liu & Yoon,2019). The influence of airspeed and wind direction on a human's thermal conditions and airflow around the body (Oh & Kato, 2018). The wind direction also has an impact on air pollution. If somehow the air is moving from an industrial area towards an urban area, pollution levels are likely to be higher within the city or town than if the wind is blowing from another direction.

2.5 Air Pollutants from Vehicles

There are several pollutants that could affect air quality in terms of combustion from transportation and also from other sources. Fossil fuels consisting mainly of carbon, hydrogen, nitrogen, sulphur, and oxygen produce the following products during combustion: The primary pollutants are Carbon Monoxide (CO), Sulphur (SO_2), Mixture of Nitrogen (NO_2, N_20, NO_x) and Volatile organic compounds (VOCs), These gases are very harmful to the society since it could affect humans' health and the environment

2.5.1 Carbon Monoxide Concentration (CO)

Carbon Monoxide is mainly produced by the combustion of various fossil fuels for power generation and transportation. Motor vehicle emission levels are a major source of anthropogenic carbon monoxide CO (Payus et al, 2019). The major source of CO from vehicles is incomplete combustion of gasoline in engine cylinders. The fuel-oxidation process which essentially means vehicle combustion entails converting the fuel to lower-molecular-weight intermediate hydrocarbons such as olefins and aromatics, then to aldehydes and ketones and finally CO. CO emissions differed across regions, indicating significant regional heterogeneity in the mechanisms of urbanization process influence on regional CO emissions. In the view of Department of Environment Malaysia, Carbon Monoxide was identified as an indicator on traffic emission. It means that if the concentration of CO is higher, the area is saturated with traffic.

2.5.2 Sulphur Dioxide Concentration (SO_2)

Sulphur is a biologically active component of crude oil that can be found in gasoline and diesel. Sulphur dioxide (SO_2) or sulphate fine particles is emitted when these fossil fuels are burnt. Just because a low-sulphur fuel is used, the accumulated amount of sulphur in the air mixture can be quite high due to the large amount of fuel used and the vehicle's long lifetime. If a fuel containing 10 wt.-ppm sulphur is used, this would result in a total exposure of approximately 5 kg of sulphur (Dahlin et al., 2019). Thus, sulphur dioxide needs to consider as transportation emission although only a low sulphur fuel was used. Besides that, it is a toxic gas that causes the odour of burnt matches.

2.5.3 Mixture of Nitrogen Concentration (NO_2 / N_2O / NOx)

When a car engine is turned on, it tends to cause a type of combustion by rapidly heating up. This process serves as a catalyst, combining nitrogen (NO_2) and oxygen (O_2) to form nitric oxide (NO) or nitrogen dioxide (NO_2). The generic term for both is nitrogen oxide (NO_x). Nitrogen Dioxide (NO_2) is one of a class of reactive gases known as nitrogen oxides or oxides of nitrogen (NO_x). The primary source of NO_2 in the atmosphere is the combustion of fuel. NO_2 is produced by emissions from automobiles, trucks, and buses, as well as power plants and off-road vehicles. NO_2 may aggravate respiratory symptoms in the presence of a coexisting infection. Furthermore, nitric oxide (N_2O) and NO_2, or NO_x as they are colloquially known, are precursors of ground-level ozone. NOx emissions are produced by both diesel and gasoline-powered vehicles (Gwilliam et al., 2014).

2.5.4 Volatile Organic Compound Concentration (VOCs)

Volatile organic compounds can be found in many different of products, including liquid paints, printing inks, a wide range of commercial products, organic solvents, and petroleum products. Vehicles and vessels also emit VOCs, which contribute to air pollution and smog. In urban environments, on-road vehicles are a significant source of VOCs (Zhao et al., 2016, Zhao et al., 2017). Although individual VOCs such as alkanes and polycyclic aromatic hydrocarbons (PAHs) have been confirmed in numerous studies (Schauer et al., 2001, Schauer. 2012), total-VOCs emitted by vehicles have received more attention. Gwilliam et al., (2014) study showed that the mainly two precursors of ozone are nitrogen oxides (NO_x) and chemically reactive volatile organic compounds (VOCs). NOx is emitted by both gasoline and diesel-powered vehicles, while VOCs are mostly emitted by gasoline-powered vehicles. Thus, decreasing the amount of diesel from transportation is important in avoiding the country suffered from high ambient concentration of ozone so that the society could not harm in terms of health problems.

2.6 Other Air Pollutants

Other air pollutants will be taking into consideration too since it also influences air quality. The other pollutants are Ozone (O_3) and Total Suspended Particles which are $PM_{2.5}$ and PM_{10}.

2.6.1 Ozone (O_3)

Ozone is a gas made up of three oxygen atoms. Ozone exists both in top and bottom atmospheres of the Earth. Ozone can be beneficial or harmful depending on how it is discovered. Ozone (O_3) is a highly reactive gas that is comprised of three oxygen atoms. It is both an organic and man-made component that occurs in the Earth's upper atmosphere (the stratosphere) and lower atmosphere (the troposphere). There is also finding showing that O_3 has a significant relationship with pollutants from traffic emission (Rahman et al.,2015).

2.6.2 Total Suspended Particles (PM_{10}, $PM_{2.5}$)

Total suspended particles (TSP) is an antiquated result presented of particulate matter (PM) volume concentration in community air. It was defined by the (unintended) size-selectivity of the particle-collection filter's inlet. Total suspended particles (TSP), also known as suspended particulate matter (SPM), is a composite of many sub-types of pollutants that exist both for liquid and solid forms. WHO emphasizes particulate matter diameter of 10 microns (m) in diameter (PM_{10}) while those smaller than 2.5 m ($PM_{2.5}$) are known as fine or respirable particulate matter. Finally, Gwilliam et al., (2014) stated that other air toxin emissions of primary concern in vehicle exhaust include benzene and poly-aromatic hydrocarbons (PAHs), both well-known carcinogens. Air toxin emissions such as benzene depend mostly on fuel composition and catalyst performance. Exhaust PAHs are due primarily to the presence of PAHs in the fuel itself; in the case of gasoline, they are also formed by fuel combustion in the engine.

2.7 Characteristics of study area

A person's or a group's socioeconomic status is their social rank or class. Education, wealth, and employment are frequently used to calculate it. Examining socioeconomic status frequently reveals discrepancies in resource access, as well as concerns of privilege, power, and control. Socio economic and characteristics developments have significantly shaped the distinctive characteristics of cities in urbanized areas, potentially leading to ambiguity in interpretations of the relationship between urbanization and environment issues due to transportation.(Wang et al., 2021). Ding et al. (2017) found that the trend of these influence factors' effect on regional CO emissions varied across different provinces, which means there are significant regional heterogeneity in the influence mechanisms of urbanization process on regional CO emissions. This can prove by a study from Zakaria et al. (2017) shows that the number of motor vehicles, industries, and other activities in Shah Alam grows, so does the concentration of NO_2, O_3, and PM_{10}, as well as CO emissions since Shah Alam is a growing city from time to time. Cities with high and low socioeconomic status have varying levels of hazardous air pollution. It was discovered that the occurrence of polluted and non-polluted days is affected by PM_{10} concentrations with a high probability that the polluted day returns every 4 to 5 days in Shah Alam and every 2 to 3 days in Jerantut (Mohamad et al.,2018). This is due that most of the pollution came from transportation in Shah Alam while the main contribution for air pollution in Jerantut is biological sources. Jerantut is surrounded by natural forest and agricultural areas, as well as traditional Malaysian villages. The reason why it is polluted although the city is not as growing in Shah Alam is due to local open burning, soil dust and a low number of motor vehicles (Azmi et al., 2010).

The rise of metropolitan areas reduces cities' environmental resistance even further, resulting in intense regional stresses as a result of this "urban syndrome" will cause the rise of bad air pollution eventually (Fang and Ren, 2017). For example, Selangor's fast urbanisation has resulted in drastic changes in the state's land use pattern. Between 1991 and 2002, the state's saturated land usage increased from 33,680 hectares to 127,591 hectares which includes Shah Alam, Klang, Banting, Petaling Jaya and Kuala Selangor. Selangor relates to roadways (Federal Highway, MRR2, etc.) and trains within its districts, province, and neighbouring area (KTM, ERL). The study also stated that even if the progress is beneficial to the country, the impact on the environment, particularly urban air quality, is concerning. Not only that, Kuala Selangor can be classified as the cleanest city among the five research locations while the number of good API days in Klang and Banting was the lowest during the period since Kuala Selangor is less urbanized than any other states (Zaini et al., 2021).

2.8 Air Pollution Analysis

This chapter talks about the method used by previous study when comes to air pollution. The method includes in the study are factor analysis, cluster analysis and spearman correlation analysis. Details of how the analysis work are explained below.

2.8.1 Application of Factor Analysis (FA) in air pollution studies

Factor analysis is a statistical technique used frequently in psychology and the social sciences. Even with the development of supercomputers, factor analysis and other multivariate approaches are now available to a much bigger population (Kline,2014). Factor analysis is a method of reducing a large number of variables to a smaller number of factors. The highest common variance from all elements is extracted and converted into a single score using this procedure. This score can be used as an index of all factors for further investigation.

Malaysian academics have also frequently employed factor analysis in their studies on air pollution. Several examples of past research study are explained so that the researcher could compare the result and interpret it in the context of transportation. Firstly, factor analysis was utilised to reduce the number of independent variables to determine which factors relates with. Jamil et al, (2019) uses factor analysis to investigate the factors that contribute towards the air pollution in Penang. It was found out that there are three factors were identified as either implicated in Penang. CO, NO_2, and PM_{10} are among the pollutants emitted by traffic emissions. Furthermore, factor analyses were utilised to establish which factors in this study were the most dominant. Factor analysis was also used in the Klang Valley to identify main sources of air pollution (Rahman et al., 2015). Motor vehicles were thought to be the primary contributors to pollution production and the study concluded that there are three factors which are related to transportation, weather characteristics, and 0_3. CO and NO_2 are among the pollutants emitted by traffic in the Klang Valley. Factor analysis is one of the most common and useful statistical approaches for determining the potential structure of a set of variables. Azid et al. (2015) conducted factor analysis to focus on finding possible sources of air quality differences across the study area. Potential gaseous pollutants and non-gas pollutant were discovered to be two possible factors. CH_4, NmHC, and THC are examples of gas pollutants while O_3 and PM_{10} are examples of non-gas pollutants. Then, this method is also important in identify linear combinations of the original variables that are useful in accounting for the variation in those variables. For example, factor analysis was used to conduct a preliminary assessment to identify pollution sources in the Klang Valley (Mohamad et al., 2015). Pollutants such as CO, NO_2 and PM_{10} are emitted by traffic while meteorological parameters such as temperature, humidity and wind speed are recorded. Meteorological has high variance compared to traffic emission since it influences air quality in Klang the most. Asmarini (2015) attempts to identify the main factors and causes of air pollution levels within four Malaysian air monitoring stations. The investigation revealed three factors: Organic Pollution Factor, Meteorological Factor, and the Fuel Factor. High concentrations of CO, NO_2, SO_2, and PM_{10} make up the fuel component which contribute the most in air pollution. The geographical factors of the 5-air quality monitoring site layout are investigated in a study (Isiyaka et al.,2015). The study's findings indicate that there are two important factors. Heavy industrial operations and photochemical pollution are the two factors . PM10 and NO2 are among the components of factor 1 while NO_x, CO, and VOCs are among the components of factor 2. Finally, factor analysis also used to extract the most significant parameters by eliminating the less significant parameters with minimal loss of the original variables. It was used to identify probable causes of air pollution in different location (Abdullah et al., 2015).

The approach has also been used globally to measure air quality based on the researcher's objective. Keresztes and Rapo (2017) used factor analysis to investigate the variations in air pollution concentrations over a two-year period. Factor analysis was used to summarizes the correlation patterns. Findings found out that factor 1 comprises of automobile transportation whereas factor 2 incorporates O_3 and meteorological conditions. Factor 1 is made up of NO_x, NO_2, CO, SO_2, PM_{10}, and temperature while factor 2 is made up of O_3 as well as climatic components like wind speed, relative humidity, and sun radiation. Factor analysis is employed to identify the major air pollutants in different season (Zhang et al., 2016). The results demonstrate that the first factor in spring is NO_2, PM_{10}, CO, and $PM_{2.5}$ whereas the first factor in winter is NO_2, PM_{10}, $PM_{2.5}$, and SO_2. It illustrates that the burning of fuel varies depending on the season since meteorological elements play a significant role in air pollution. A study by Das et al (2015) shows that factor analysis was used to determine that there are 3 different factors for PM_{10} and $PM_{2.5}$. Aside from that, the levels of air pollutants in Mecca of Saudi Arabia were examined using factor analysis and important sources were found (Nayebere et al., 2018). It indicates that there are 4 major factor in mecca when comes to air pollution. The factors are Fossil Fuel Combustion, Industrial Emission, Vehicular emission and agriculture soil industry emission while traffic emission includes NO_3-, C_2O_4 , V and BC.

According to the review, most of the studies do not focus on TRAP because they interpret air quality broadly rather than focused on transportation. Despite from the issue, it is possible to compare the study results with these past reviews but will be interpreting it in the context of TRAP. Hence, in this study will be focussing on the potential factors that may affect TRAP itself. Appendix 1 below summarizes the method use in the study related to Air Pollution.

2.8.2 Application of Cluster Analysis (CA) in air pollution studies

Cluster analysis is a popular method for analyzing multivariate data. Its use is widespread and rapidly expanding (Kline,2014). Cluster analysis which known as clustering is the process of grouping a set of elements so that elements in the same group are much more like those in other groups (Kettenring, 2006). Cluster analysis, also known as "clustering," is a method for grouping similar observational data, data points, or feature vectors based on their similarities (Jain et al., 1999). It is defined as "the art of finding groups in data" by Kaufman and Rousseeuw (1990).

Cluster Analysis also been often used by the researchers from Malaysia in the study on air pollution. Several examples of past research study are explained so that the researcher could compare the result and interpret it in the context of transportation. Firstly, cluster analysis is important since it is necessary to determine air quality patterns of observed air pollutants. For example, Cluster analysis was used to analyse the air quality of Langkawi Island, using data from the Malaysian Department of Environment spanning 13 years (1999–2011). Throughout the investigation, three clusters were identified. One of the cluster were corresponds to traffic emissions from early morning to early evening, resulting in greater concentrations of PM_{10}, CO, NO, NO_x, NO_2, and SO_2 (Halim et al., 2018). Besides that, cluster analysis is a technique for grouping or clustering observations based on their similarities or differences. By using cluster analysis, two clusters were identified which one with a moderately populated area and one with a less populated area (Isiyaka et al.,2015). Cluster with moderately populated area is heavily associated with commercial and industrial operations as well as severe traffic congestion whereas cluster with less populated area is regarded as a residential neighbourhood, tourist destination and has less industrial activity. Not only that, this method also is able to identify the origins of a big group of data that was also divided into smaller groups based on their similarities and form a cluster (Hua, 2018). There are around hundred thousand of observation obtain from 8 station, but it is possible to detect three clusters and which station were most affected by transportation and industrial pollution Awang et al. (2016) also intend to identify the key sources of air pollutants that influence the ozone Critical Conversion Point. Cluster analysis was responsible for categorising the nine air monitoring sites into three distinct clusters based on the parameters chosen. Cluster 1 was highlighted as the primary city centre, and it has a high concentration of carbon monoxide since it is surrounded by residential areas, commercial areas, and other amenities. The method also is important in illustrate dissimilarity from a large group of samples to profiling based on the pattern. Sahrir et al. (2019) use cluster analysis to discover air quality patterns on a yearly and monthly basis since transportation is a major issue in air quality. There were two cluster form from the output. The cluster were mainly affected by its location of area, concentration of Carbon Dioxide and PM_{10} also plays a major factor.

Cluster analysis is also often used method in environmental studies. A study by Wu et al. (2019) aims to examined on the spatiotemporal characteristics of major pollutants as well as city clusters with identical air quality. The findings by using cluster analysis in the study shows that several cluster identified. These clusters were influence by the contaminants pollute Chinese air to differing degrees based on each respective city. Shekarrizfard et al (2016) aims to test the relationships between emissions and exposures in different Montreal areas by using cluster analyses. Two cluster were identified and results also found out that people who reside in the suburbs locations emit more NOx but are exposed to lower NO_2 concentrations at home and during their everyday activities. A study by Zhang et al. (2016) aims to investigate the principal air contaminants, as well as their geographical and seasonal distribution in 74 Chinese cities by using cluster analysis. For example, cities in Cluster 5 are the most severely polluted in the area by having high concentration of SO_2, CO, PM_{10}, and $PM_{2.5}$ while O_3 emission sources would have a significant impact on air quality. Cluster analysis were used in several monitoring stations across Shanghai such to correlate air pollutants (Li et al, 2019). The results revealed that the spatial variation differed depending on the pollutant as did the urban form $PM_{2.5}$ concentrations were high in Shanghai western suburbs and low in the city's eastern outskirts. Cluster analysis were also used to classify the synoptic weather patterns (Hsu and Cheng, 2019). It found out that meteorological factors play an important role in high concentration of TRAP gasses. For example, cluster 3 impacted by weak synoptic weather and has the lowest wind speeds, as well as the highest $PM_{2.5}$ and PM_{10} levels. Finally, a study by Chuang et al, (2018) aims to evaluate the relationships between $PM_{2.5}$ bioreactivity in vitro and emission sources. Findings found out that $PM_{2.5}$ level influenced by the long-range transported and major industrial emissions. $PM_{2.5}$ level mainly contributed by traffic and farming industrial emissions. From the review, most of the study can be observed that very few have focusses on classifying the interested matter depending by the by its location, time , and the concentration of gasses. This shows that less study was focus on TRAP itself. Despite from the issue, it is possible to compare the study results with these past reviews but will be interpreting it in the context of TRAP. Hence, in this study will be focussing on the potential factors that may affect TRAP itself. Appendix 2 below summarizes the method use in the study related to Air Pollution in the context of transportation.

2.8.3 Application of Spearman Correlation relates to air pollution studies

Spearman correlation analysis is a test statistic that determines the statistical link, or association, between two continuous variables. It is known as the best approach of quantifying the relationship between variables of interest. It reveals the size of the association, or correlation, as well as the relationship's direction. The strength of relationship can be anywhere between -1 and +1. The stronger the correlation, the closer the correlation coefficient comes to ±1. If the coefficient is a positive number, the variables are directly related. (Mukaka, 2012). This analysis is suitable if the continuous data between two variables is not normally distributed.

Spearman correlation have been often used by the researchers from Malaysia in the study on air pollution. Several examples of past research study are explained so that the researcher could compare the result and interpret it in the context of transportation. Spearman analysis was often used by previous researcher to determine the relationship between the most significant air pollutant and other factors. However, CO have been focus too since it is an indicator variable for treaffic emission. For example, CO were significant with SO_2, NO_2, O_3, PM_{10} and Humidity (Zakaria et al., 2017). CO was highly related with NO2 among other pollutants with a correlation coefficient of 0.732 indicating a strong positive relationship. The sources of these contaminants which originated from industrial and traffic activities show a link between them. According to several studies, motor vehicle emissions are the primary source of CO and NO2 (Isiyaka et al., 2014 ; Dominicik et al., 2012). Furthermore, Sofwan et al., (2021) used this approach and identified that CO were significant with NO_2, SO_2, humidity and temperature. CO was found to be strongly linked to NO_2 as well as other pollutants with a correlation coefficient of 0.635 indicating a moderately positive relationship. The significant correlation between CO and NO_2 is owing to their cumulative effects on respiratory admissions. The spatial assessment of ambient air quality status was then assessed using this method (Shafii et al., 2017). CO levels were shown to been significant with wind speed, humidity, temperature, SO_2, NO_2, PM_{10} and O_3. CO was strongly linked to NO_2 due to both pollutants being caused by motor vehicle emissions in Klang (Azmi et al.,2010).This approach were also used by Kamaruzzaman et al., (2017) to evaluate the air pollution in Putrajaya. It was found out that CO were significant with SO_2, NO_2, O_3 and PM_{10}. CO were highly associated with NO_2 compare with other pollutants due by the released from engine boats and motor vehicles, especially when the engines turned on, thus leads to the incomplete combustion

The approach has also been used globally to measure air quality based on the researcher's objective. Parveen et al., (2021) employed Spearman correlation analysis to investigate the relationship between meteorological variables and air pollutants. In New Delhi, India, it was discovered that CO were significant with meteorological factors such as with humidity, temperature, wind speed and rainfall. CO and NO_2 were strongly linked to other pollutants which it owings to emissions from vehicles and most of it were from motorbikes. This approach is used by Liu et al., (2021) to investigate the relationship between $PM_{2.5}$, PM_{10}, and gaseous pollutants in California, USA. CO levels were significant with PM_{10} and $PM_{2.5}$. It also indicates a moderately positive association suggesting that when CO concentrations rise, PM_{10} and $PM_{2.5}$ concentrations rise as well. Finally, Fang et al., (2021) employed this approach to find relationships between air contaminants and climate. In Yancheng, China, CO levels were related to temperature and humidity. Due to automobile and industry emissions, CO was also strongly linked to $PM_{2.5}$. From the review, it was compared with and interpret it in the context of TRAP pollution although most of the study focus air quality in a broad way. Hence, in this study, the researcher is focussing on the potential factors that may affect TRAP itself . Appendix 3 below summarizes the method use in the study related to Air Pollution in the context of transportation

2.9 Concluding Remarks

A literature review is a thorough summary of prior research on a particular subject. The literature review examines scholarly articles, books, and other sources that are pertinent to a specific study topic. This previous study should be enumerated, described, summarised, objectively evaluated, and clarified in the review. Therefore, from this study, the literature review covers the starting point of air quality that relates with transportation related air quality until the end where the researcher accessing the potential factors that contribute to TRAP. Based on the previous study, the researcher could say that TRAP pollution is very dangerous since it could harm people health. Nevertheless, lack of study on TRAP pollution has been focus among the researchers. Since it is very dangerous, the researcher would like to learn more on TRAP pollutions to assess the dominant factors on the pollution itself. Therefore, from here the researcher would like to assess the pollution by looking into consideration on each type of particle of air and other factors.

CHAPTER THREE
RESEARCH METHODOLOGY

3.1 Introduction

This chapter contains in-depth descriptions of the methodologies used in this study. This approach covers every aspect of performing this research, including data collecting, data preparation, and data cleaning. Research Design as well as the Data Analysis and Methods are the focuses of this chapter too.

3.2 Research Design and Methods

The type of research used in this study would be determined by the study's aims as well as the nature of the data. The study is quantitative and explanatory research as it aims to investigate a relative unknown field to gain a better insight on the issue. In the context of this study, the purpose is to obtain better knowledge on TRAP itself.

3.3 Data Acquisition

Data on hourly concentrations of various air contaminants are received from the Department of Environment (DOE) Malaysia. The source of data used in this study is a secondary data on air quality. These data is recorded hourly in a span of two years (2018-2019) for certain location. Table 3.1 below displays the summary of the dataset used in this study.

Table 3.1
Description of the dataset

Location	Shah Alam , Petaling Jaya, Jerantut and Banting
Dataset Span	2018 - 2019
Type of data	Continuous data
Air Pollution Parameter that are focussing on	Carbon Monoxide (CO)

Source : Department of Environment (DOE) Malaysia,2021

The Department of Environment (DOE) uses a network of 51 stations to monitor the country's ambient air quality. These monitoring stations are deliberately placed in residential, commercial, and industrial locations to identify any substantial change in air quality that could be damaging to human health and the environment. These locations are chosen because it was used to compare TRAP itself in different category of location. There are four category of location which are Industrial station, Urban station, Sub-urban station and Background station. All these locations are look at so that the researcher could have better understanding of traffic emission based on these different categories of location. Thus, 4 different categories of location are chosen for the study. The information is obtained by Rahman et al. (2017) and Latif et al. (2014) in the interpretation of location that are chosen are described as Table 3.2 below

Table 3.2
Description of chosen air quality station

Air Quality Station	**Coordinate**	**Area Category**	**Area Description**
Shah Alam	N 3.1 E 101.5	Urban	A station where is focussing on mixed housing area with commercial centres or offices. Usually, this type of area is a highly common traffic area.
Banting	N 3.32 E 101.2	Sub-Urban	A station where is focussing on agriculture area and less common traffic area than in urban area.
Jerantut	N 3.8 E 103.3	Background	A station where it usually made to be the reference among other type of station. The location has good air quality and ambient air.
Petaling Jaya	N 3.1 E101.5	Industrial	A station where it is solely focus on air pollution that comes from industrial area and highly congested area.

Finally, the stations of each location are shown as Figure 3.1 as below to have a bigger picture on how these stations are explored.

Table 3.3 displays the variables and a description of each variable with different types of measuring scales used in this study

Table 3.3
The description of each variable

Factors	Variables	Description	Measure
Meteorological factors	Temperature	Temperature of the location(°C)	Numeric
	Humidity	Humidity of the location (%)	Numeric
	Wind direction	The direction of the wind (°)	Numeric
	Wind Speed	The speed of the wind m/s	Numeric
Air pollutants from Vehicles	Carbon Monoxide	Concentration of Carbon Monoxide (CO) (mg/ m^3)	Numeric
	Nitrogen Dioxide	Concentration of Nitrogen Dioxide (NO_2) (mg/ m^3)	Numeric
	Nitrogen Oxides	Concentration of Nitrogen Oxides (NO_x) (mg/ m^3)	Numeric
	Sulphur Dioxide	Concentration of Sulphur Dioxide (SO_2) (mg/ m^3)	Numeric
Other Air Pollutants	Ozone (O_3)	Concentration of Ozone (O_3) (mg/ m^3)	Numeric
	PM_{10}	Concentration of PM_{10} (mg/ m^3)	Numeric
	$PM_{2.5}$	Concentration of $PM_{2.5}$ (mg/ m^3)	Numeric

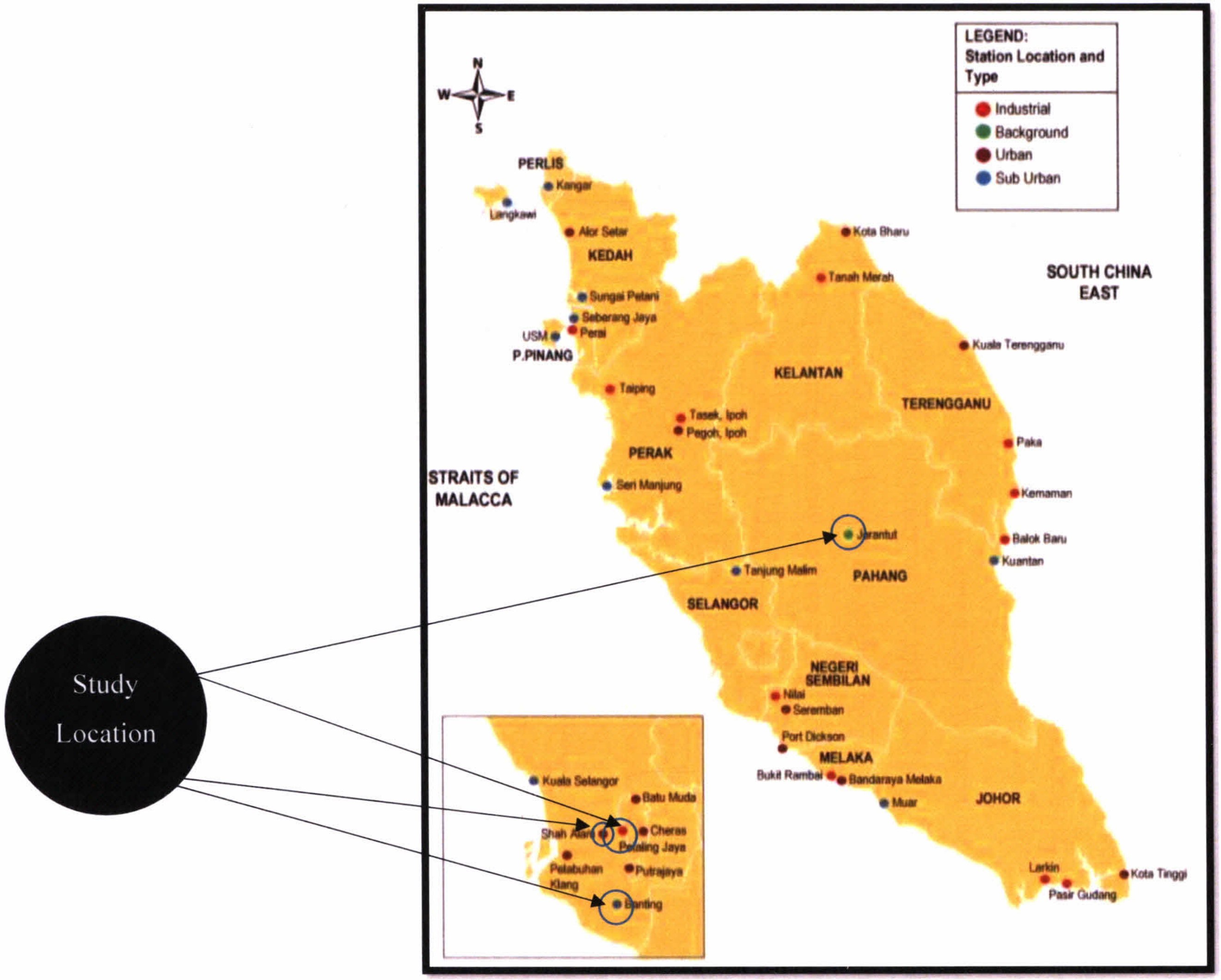


Figure 3.1 Map of each station location

3.4 Flow Chart of Research Process

The research framework lays out the steps for conducting the analysis in this study, which ultimately aid in reaching all of the research goals. Figure 3.2 shows the flow research of each process. The first step is data acquisition on each variable that need to be obtain for the study. The data are obtained from Air Quality Division, Department of Environment (DOE) Malaysia. Data understanding involves where the researcher should know the data from all aspects such as the identifying the dependent variables to the independent variables. For example, the dependent variable is carbon monoxide since Department of Environment (DOE) used this pollutant as an indicator to measure the concentration of traffic emission. If the concentration of traffic emission is higher in a particular area, therefore it has high concentration of traffic emission.

Then, data preparation is done since it is crucial when comes to data analysis such it could affect the accuracy of the analysis itself. Data preparation involves in dealing with the data since the data is large. The chances of a large data have missing values, outlier, multicollinearity and incorrect value are extremely high. This could the the accuracy of the analysis if there is no countermeasure on the given problem. If the data is already prepared for further analysis but does not satisfy in diagnostic checking, data preparation is to be made again.

After preparing the data, descriptive analysis and multivariate analysis are done here. Descriptive analysis is used to summarize about the sample of data and measure while multivariate analysis is a more complicated analysis to obtain better insight on the result. For example, in multivariate analysis, the researcher uses cluster and factor analysis in dealing with our objective 1 and 3 respectively meanwhile objective 2. This objective was put accordingly to illustrate the flow of this study in interpretating air quality in the context of TRAP. Therefore, each of this process are explained one by one on each subtopic from data acquisition until achieving those objectives below.

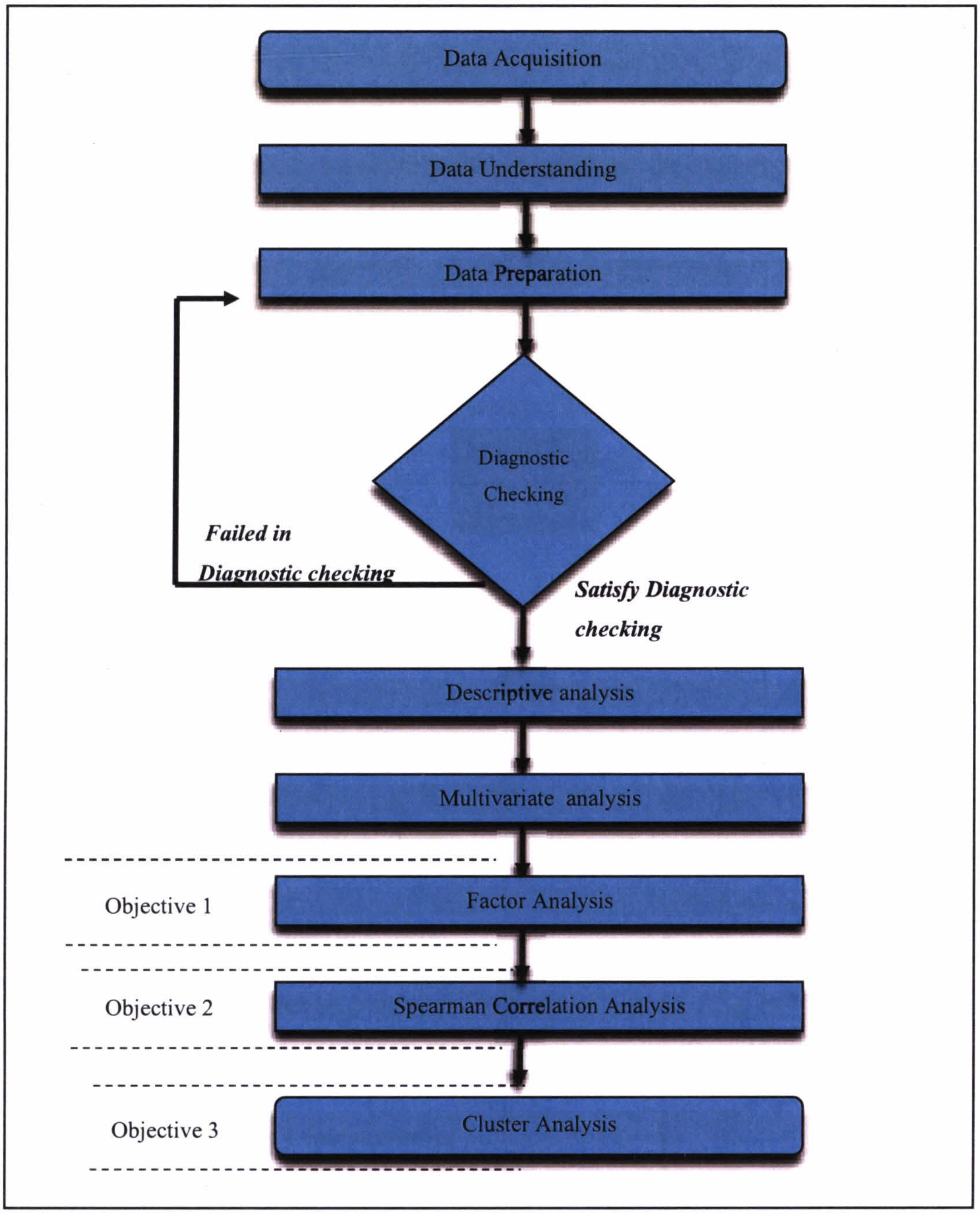


Figure 3.2 Flow of Research Process

3.4.1 Data Pre-processing

After the data has been obtained, some time are spent understanding the information contained in the data. The meaning of each value in each column are determined by scanning each column in the dataset. Understanding the data is a critical step in ensuring that the analysis uses the appropriate variable Prior to conducting any analysis, the data undergo pre-processing stage which the data were cleansed to make it readily available for further analyses. The pre-processing of the data is further discussed in next section 3.6 (Data Preparation) since the data need to be diagnosed

3.5 Conceptual Framework

The conceptual framework below depicts the relationship between the independent variables (i.e., Meteorological factors, Air pollutants from vehicles and other pollutants) and the dependent variable (i.e., Concentration of Carbon Monoxide). Since the Department of Environment Malaysia has focused on the concentration of carbon monoxide in determining the level of transportation emissions in Malaysia, carbon monoxide is the dependent variable. Thus, the researcher could say if the concentration of CO is high , the level of traffic emission is bad at a particular area.

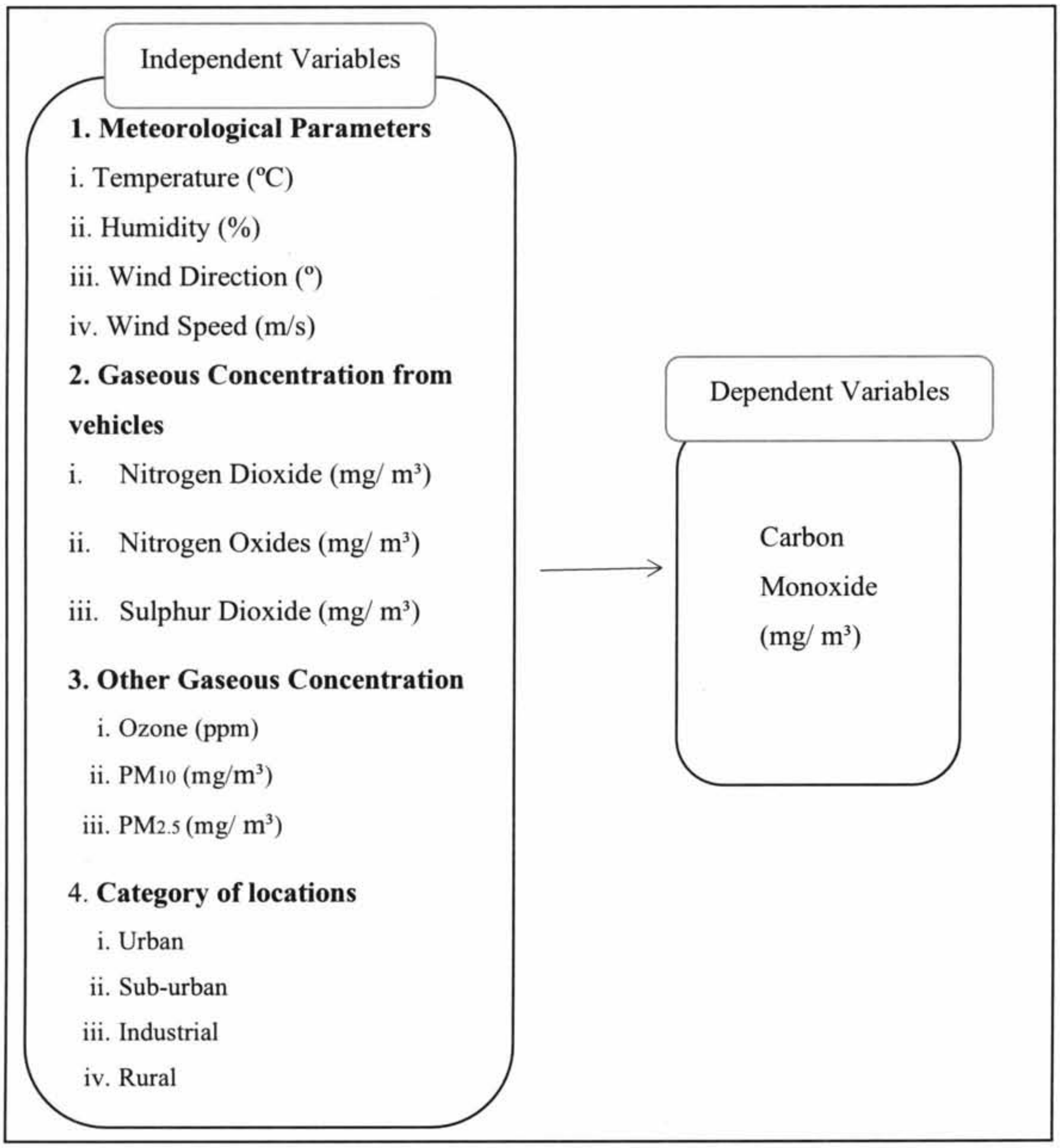


Figure 3.3 Conceptual Framework

3.6 Data Preparation

The process of cleaning and altering raw data prior to processing and analysis is known as data preparation. During this phase, it involves in checking the normality of the data, imputing missing values and checking multicollinearity.

3.6.1 Normality of the data

Normality test will be performed by the researcher before proceeding or continue with replacing missing data. This is because, the researcher needs to identify which method is to be used in replacing the missing data. By using SPSS software, the researcher could gain or determine the distribution of the by looking at the pattern of the histogram for each variable. If it shows a bell curve shaped, the data is normally distributed. However, Kolmogorov-Smirnov test was used to identify the distribution of the data. The Kolmogorov-Smirnov test is used to test the null hypothesis that a set of data comes from a normal distribution(Costello and Osborne, 2005). Although Shapiro-Wilk is also one of the method could be used in checking the normality of the data, Kolmogorov-Smirnov test is suitable to be used because Kolmogorov-Smirnov test is suitable for big data set where sample is more than 50 observation (Paul,2010). If the result is significant, the data shows a not normally distributed data for the given variable.

3.6.2 Imputing Missing Values

In this research, SPSS is used to impute missing values from data. The variables with missing values are ready to be imputed by substituting with other values. The SPSS output reveals the number of missing values in this study's variables. For continuous data, the SPSS employed a default imputation approach that replaced the missing value with the mean value and for categorical data, it replaced the missing value with the mode value. The major disadvantages of missing data are reduced statistical power because it limits the number of samples n and greater standard errors in estimations. However, if the continuous data is not normally distributed, median imputation are used to replace the missing values. Although mean imputation is often used by other researcher to replace missing values, median imputation is also another simple method often appropriate for highly skewed data and may yield better results compared to mean imputation (Junger and De Leon, 2015; Miettinen, 1985). Evaluation by Hadeed et al. (2020) shows the best imputation by comparing the error in Table 3.4.

Table 3.4

Top three method of imputation of missing values

Error Metrics	**Percent Missing**			
	Less than 20	40	60	80
MAE	1. Median	1. Median	1. Median	1. Markov
	2.Kalman	2.Kalman	2.Kalman	2. Median
	3.LOCF	3.LOCF	3.Markov	3. Mean
RMSE	1. Median	1.Kalman	1. Markov	1. Mean
	2. Kalman	2. Median	2. Random	2. Random
	3. Markov	3.LOCF	3. Mean	3. Markov

3.6.3 Multicollinearity

Multicollinearity is a scenario in which the predictor variables are significantly associated with one another (Paul, 2010). When two independent variables are discovered to be correlated with each other, multicollinearity is said to present in the dataset. The independent variables must not be connected to each other in order to proceed with the analysis. The data are verified to confirm this by checking at the Variance Influence Factor (VIF) values or Tolerance value. If the VIF value is greater than 10 and the Tolerance value is less than 0.1, it implies that a multicollinearity problem occurs (Tranmer and Elliot, 2008). If there is multicollinearity, the most correlated variable was dropped to provide an accurate answer. This process is very important in achieving the assumption of factor analysis where the determinant value should be higher than 0.001

3.6.4 Outlier Detection

An outlier is a value that deviates abnormally from other values in a population random sampling. For the purpose of this study, the data will test for the outlier. If the data contain the outlier, the observation will be removed for this study. The data will be used mahalanobis distance which is one of a multivariate technique to look for an outlier in this study. Mahalanobis Distance (MD) is an effective distance metric that finds the distance between point and a distribution (Leys et al., 2018). Therefore, in detecting outliers using MD, the distance between each point and the centre in n-dimension data is calculated where outliers are found by taking these distances into account. If the data contains an outlier, the observation will be eliminated such that the analysis will continue with the outliers excluded. If the probability value for mahalanobis distance is less than 0.001, then it shows that there exists an outlier on it (Dashdondov et al.,2021). The formula for the mahalanobis distance is shows as follow where D^2 is the square of mahalanobis distance, x is the vector of row observation, m is the mean value for each variable and C is the covariance matrix in the independent variable.

$$\mathbf{D}^2 = (\mathbf{x}-\mathbf{m})\text{^}\mathbf{T} * (C^{-1}) * (x-m) \tag{3.1}$$

3.7 Method of Analysis

As stated in an earlier chapter, the study's aims are to find the latent relationship between the potential of dominant factors in transportation related air pollution (TRAP) and partition prospective dominant factors of transport related air pollution (TRAP) based on different category of regions to classify and make comparison from the result. Thus, the analysis is discussed clearly and explained on how Factor Analysis (FA) and Cluster Analysis (CA) really works. The reason both of this method is use in the study is because that both methods are often used by other researchers too. Thus, it is possible to compare the result of the study among other past studies to interpret TRAP itself. Some may argue in why factor analysis is to be used instead of principle component analysis (PCA). There are differences on factor analysis and principal component analysis. Factor Analysis (FA) assumes that the measured responses are based on the underlying factors while in PCA are based on the measured responses . FA is the most suitable method to be used when the study is interested in making statements about the factors that are responsible for a set of observed responses, while PCA supposed to be use when the study is simply interested in performing data reduction (Yong et al,2013). Since, the study would like to see other variables are responsible in the concentration of Carbon Dioxide, therefore factor analysis is used in the study

3.7.1 Descriptive Analysis

Descriptive analyses are used to describe or summarize the characteristic and the features of the data in the study. Simple summarize are provided about the sample of data and measure. Every basic of quantitative analysis of the data are form with the simple graphic analysis. For example, bar chart, pie chart, histogram and other more are conduct to this study within certain variables. The data also were explained according to the result obtain. The central tendency measure also is explained in this section. This includes descriptive statistics like mean and median, standard deviation, kurtosis, and histogram as well as the box plot. Not only that, the skewness value is reported too in identifying the distribution of the data as well. Data are considered as normally distributed since the value of skewness is between -3 and 3 (Jones,1969) Thus, it is important to explore the data first to have a better insight on the data itself before proceeding towards the objectives.

3.7.2 Factor Analysis (FA)

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved variables called factors. Factor analysis is a technique for condensing many variables into a smaller number of factors (Chatfield and Collins, 1981). This method extracts the maximum common variance from all variables and converts it into a single score. Even though Principal Component Analysis could be used for the study, factor analysis is much suitable to use for exploratory research design to gain better information on the data itself (Costello and Osborne, 2005). The goal of factor analysis is to condense many individual items into a smaller number of dimensions. Factor analysis can be used to reduce the number of variables in regression models. For example, in this study, the researcher uses factor analysis to examine the underlying or latent relationship between the potential of dominant factors in transportation related air pollution. A variety of Likert scales are examples of ordinal variables that are often utilized in PCA. A better insight of how Factor analysis is shown in Figure 3.4 to show how does Factor analysis works in the study.

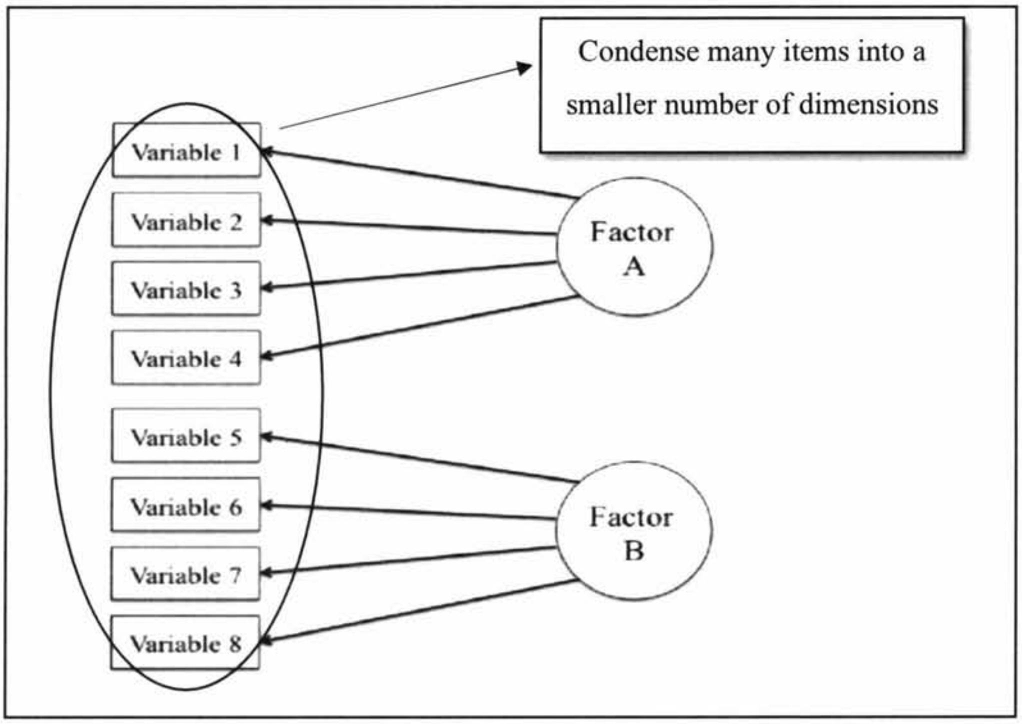


Figure 3.4: Description factor analysis in a figure

3.7.2.1 Assumption Checking

The assumption for factor analysis includes determinant in correlation matrix, bartlett's test of sphericity and The Kaiser Mayer Olkin (KMO). The assumption was needed to be satisfied before proceeding with the actual analysis.

a) Determinant in correlation matrix

Multicollinearity can be detected by the determinant of the correlation matrix. If the determinant is greater than 0.001, then there is no multicollinearity (Field, 2000). If the determinant is lesser than 0.001, therefore the data is not appropriate for a factor analysis. This test must be significant where the correlation matrix is an identity matrix, there would be no correlations between the variables

b) Bartlett's test of sphericity

Bartlett test was used to determine whether the correlation between items is good enough for factor analysis to be conducted. The hypothesis that the correlation matrix is an identity matrix is tested by Bartlett's test of sphericity, which indicates that the variables are unrelated and thus unsuitable for structure detection. Data reduction should be possible with the data. For variables to be reduced to a smaller number of components, the researcher must have an acceptable correlations between them. SPSS Statistics employs the Bartlett's test of sphericity to detect this matter. The correlation matrix is an identity matrix that is used to test the hypothesis .If the value of the test statistic for sphericity is large and the associated significance level is low, the population correlation is unlikely to be an identity. Because the observed significance level is too high to reject the hypothesis that the population correlation matrix is an identity matrix, the use of the factor model should be reconsidered (Fruchter,1954). The test of the equation is as follow such that n is the number of observations, p is the number of components and λj represents the eigenvalue for each component.

$$\boldsymbol{\lambda} = \sum_{j=k+1}^{p} \frac{\lambda j}{p-k} \tag{3.2}$$

c) The Kaiser-Meyer-Olkin (KMO)

It is a metric for comparing the magnitudes of observed and partial correlation coefficients. The researcher should have sampling adequacy, which essentially means that large enough sample sizes are required for PCA to generate a trustworthy conclusion. Many different approaches have been offered. These are mostly differentiated by whether an absolute sample size is proposed or a multiple of the number of variables in your sample is employed. A minimum sample size of 150 cases, 5 to 10 instances per variable has been advocated. Nevertheless, the researcher uses The Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy for the Overall Data Set is one of several approaches for detecting sampling adequacy. The closer the KMO measure is to one, the greater the sampling adequacy (0.8 and higher are excellent, 0.7 is adequate, 0.6 is mediocre, and less than 0.5 is unacceptable). A good factor analysis requires values that are sufficiently large (Kline,2014). A factor analysis of the variables may not be a good idea if the KMO value is low. To make it simpler, the value of KMO is explained as Table 3.5

Table 3.5
Description value of KMO

KMO Value	Degree of Common Variance
0.90 to 1.00	Marvelous
0.80 to 0.89	Meritorious
0.70 to 0.79	Middling
0.60 to 0.69	Mediocre
0.50 to 0.59	Miserable
0.49 and below	Do not factor

The KMO formula were given as below such that *rij* is the simple correlation coefficient and *aij* is the partial correlation coefficient between *i* and *j*.

$$\boldsymbol{r} = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n(\sum x^2)-(\sum x)^2][n(\sum y^2)-(\sum y)^2]}} \tag{3.3}$$

d) Factor Extraction (Scree Plot)

A scree plot is a line plot showing the eigenvalues of factors or principal components in an analysis. The scree plot is used to determine the number of factors to maintain in an exploratory factor analysis (FA) or the number of principal components to keep in a principal component analysis (PCA). Any factor with an eigenvalue ≥ 1 explains more variance than a single observed variable. Thus, the factors or components that included are those factors that has more the 1 in terms of eigenvalue based on the scree plot (Rummel,1987). The scree plot offers a visual representation of the overall variation associated with each component.

e) Factor Extraction (Total Variance Explained)

The total variance explained is a table shows the similar result such as scree plot. The difference between Total variance explained is basically, the results show in a table form while scree plot usually shows in a graphical form. The eigenvalue, or amount of variation in the original variables accounted for by each component, is given in the Total column. The percent of Variation column displays the proportion of variance accounted for by each component to the total variance in all variables (Rummel,1967). The percentage of total variance are obtained as follows.

$$\%\, \boldsymbol{of\ variance} = \frac{\boldsymbol{eugenvalue\ of\ the\ factor}}{\boldsymbol{total\ eugenvalue}} \tag{3.4}$$

f) Factor Loadings

Factor Loadings are used to examine how a set of variables measures a certain category. Factor Loadings are simply coefficients that tell us how strong the link is between the variable and the factor. They are scaled from 0 to 1. In other words, As a general guideline, the variable should have a rotated factor loading of at least |0.5|, which means that the value should be more than 0. 5 or less than –0. 5 onto one of the factors in order to be deemed significant (Rummel,1967). 0.3 is consider appropriate to some study but 0.5 and above is the safest guideline in order to factor a component (Malviya et al, 2013). Therefore, this research focusses on the factor loading which is higher or equal than 0.5. It is considered as a high factor loading; thus, it can define a factor. Lastly, the factor loading was obtained as follows where L is the matrix of factor loading and $(\lambda_i e^i)$ is the eigenvalue and eigenvector pairs.

$$\boldsymbol{L} = \sqrt{(\boldsymbol{\lambda_1 e^1})}(\boldsymbol{\lambda_2 e^2})\ldots(\boldsymbol{\lambda_3 e^3}) \tag{3.5}$$

3.7.3 Cluster Analysis

There are 2 types of clustering method used for the study such as Hierarchical and K-means Clustering. K-means clustering is a simple and widely used unsupervised machine learning technique. As previously reported by Carslaw and Beevers (2013), sources generating by air pollution should be identified and clustered using a k-means clustering technique. K-means clustering technique was employed to enable post-processing to better understand potential source characteristics in the neighbourhood. The researcher can target number k, which refers to the number of centroids required in the data, based on this analysis. A centroid is an imaginary or real point that represents the cluster's centre. Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Hierarchical clustering is a strong approach that enables researchers to construct tree structures based on data similarities. From this method, the researcher could classify and compare the cluster or group made from in the context of transportation. Finally, there is no assumption to satisfy in using both analysis too.

a) Silhouette Measure (K-Means Clustering)

The K-means algorithm clusters data by attempting to break samples into as many groups with equal variance as possible to minimise a criterion known as the inertia or within-cluster sum-of-squares. Aside from that, the K-Means technique requires the researcher to specify the number of clusters ahead of time. The number of clusters must be given for this algorithm. The silhouette measure is a technique that is used to evaluate the validity of clustering from a range of clusters ranging from 2 to 5 for each location. The silhouette is built to pick the optimum number of clusters using a ratio scale data that is appropriate for clearly separated clusters. The silhouette value varies from -1 to 1, with a high value indicating that the object is well matched to its own cluster. As a result, the clusters were chosen based on the greatest value of the Silhouette metric for each site and if the value is below 0, the cluster is not suitable to use (Browne,2001).The silhouette measure formula are obtained as follow where a is mean intra cluster distance and b is mean nearest cluster distance

$$\mathbf{S(o)} = \frac{\boldsymbol{b-a}}{\boldsymbol{max(a,b)}} \tag{3.6}$$

b) Final cluster centres (K-means Clustering)

The final cluster centres are calculated as the mean of each variable inside each final cluster. The final cluster centres represent the characteristics of the typical situation for each cluster. The final cluster centres are determined by the optimum value obtained through silhouette measurement (Browne,2001) . As a result, the researcher refers to the final cluster centres for each place to compare the trend of pollution along with both objective 1 and 2.

c) Dendrogram (Hierarchical Clustering)

A dendrogram is a branching diagram that depicts the similarities between a group of items. Each branch is referred to as a group. A group can have an infinite number of leaves. The key to dendrogram interpretation is to focus on the height at which any two things are connected together (Browne,2001). In the context of traffic emissions, the researcher would like to examine and emphasise the similarity of the results obtained through factor and cluster analysis.

3.7.4 Spearman Correlation Analysis

Spearman's correlation measures the strength and direction of monotonic association between two variables. The requirement of Spearman correlation is both variables must be quantitative despite the data is not normally distributed. For the study, the variable that test for spearman correlation is between the gaseous concertation and CO. A correlation coefficient of zero indicates that no linear relationship exists between two continuous variables, and a correlation coefficient of -1 or +1 indicates a perfect linear relationship. The strength of relationship can be anywhere between -1 and +1. The stronger the correlation, the closer the correlation coefficient comes to ±1. If the coefficient is a positive number, the variables are directly related. (Mukaka,2012). Table 3.6 shows the description and interpretation of spearman value by Zakaria et al., (2017).

Table 3.6:
Spearman value and interpretation

Value	Interpretation
Below than 0.5	Weak relationship
0.5 – 0.69	Moderate relationship
0.7-1.00	Strong relationship

Spearman rank correlation analyses are formula are obtained as follow where *p* is the spearman rank coefficient, *di* is the difference between the two ranks of each observation and *n* is the number of observation.

$$p = 1 - \frac{\sum 6di^2}{n(n^2-1)} \tag{3.7}$$

3.8 Software

The analysis is to factoring concentration of air using several software. The analysis took places using SPSS and Power BI for different analysis purposes. In this section we discussed on the uses of these software. SPSS is short form for IBM SPSS Statistics, and it can be used in various kinds of research for complex statistical data analysis. SPSS is a software has easy capabilities to produce outputs. In this study, SPSS were used for data imputation, check on multicollinearity, data description and Pearson correlation. Microsoft Power BI is a business analytics service. It intends to deliver dynamic visualizations and business intelligence capabilities, as well as an interface that is easy enough for end users to construct their own reports and dashboards. This software is to ensure that our descriptive analysis is much better and easier in interpreting by the audience

3.9 Summary Research Method and Analysis

The summary for research method and analysis are shown as Table 3.7. Elaboration on each analysis were elaborate on 3.10 (Concluding Remarks)

Table 3.7
Summary of Research Objectives and Methods

Research objectives	Analysis used in the study
To examine the underlying or latent relationship between the potential dominant of TRAP pollution within Shah Alam, Petaling Jaya, Banting and Jerantut	Factor analysis
To determine the strength of the relationship between CO and other pollutants for each factor within different categories of location	Spearman Correlation
To classify potential dominant factors of transport related air pollution (TRAP) in Shah Alam, Petaling Jaya, Banting and Jerantut using factor and cluster analysis	Cluster Analysis

3.10 Concluding Remark

The flow of the researcher was shown as Figure 3.2. The data is obtained from Department of Environment where the data were listed as in Table 3.3. Since the data obtained is large, data preparation is done because cleaning process was complicated as each of the variables had to be reviewed to see if any were missing or incorrect to proceed with.

As summarize in Table 3.7, each objective was mainly done by each method. Objective 1 involves factor analysis in examine and compare the underlying or latent relationship between the potential dominant of TRAP pollution within Shah Alam, Petaling Jaya, Banting and Jerantut. It could examine these pollution by reducing the number of variable into several factors. Those factors were reviewed and interpreted in the context of traffic emission.

After obtaining the factors from objective 1, objective 2 deals with finding the correlation and strength of the relationship between CO and other factors for each location. In this part, the researcher could identify which factor has the highest impact and weather it is a positive impact or negative impact. This method includes spearman correlation analysis since the data is not normally distributed

Finally, cluster analysis was used to compare the result obtained by factor analysis and spearman correlation analysis for each location. Even though the result may not be the same compared along with objective 1 and 2, it will not be a problem since there is no performance indicator demonstrate which method is better to be use. Therefore, this final objective is just purely for comparison so that other researcher could highlight the findings and improve it for the further future

CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 Introduction

This chapter is the checkpoint of discussing the result obtain from the analysis. A step-by-step analysis of the result are shown in this topic. Initially, this chapter tells us the overview of the descriptive for the important variables such as the mean, median and other more. Data imputations are described in this section, thus a descriptive analysis of before and after imputation are showed. Finally, the result that are being based by the objectives were showed here such as the result from factor and cluster analysis. Detail explanation for each result is described on the related sub-topic.

4.2 Descriptive statistics (Before Imputation)

This section the original state of the data before doing data preparation. On this section, the focus is to look on the missing values, normality of the data and outlier and the descriptive summary of the overall dataset.

4.2.1 Missing values

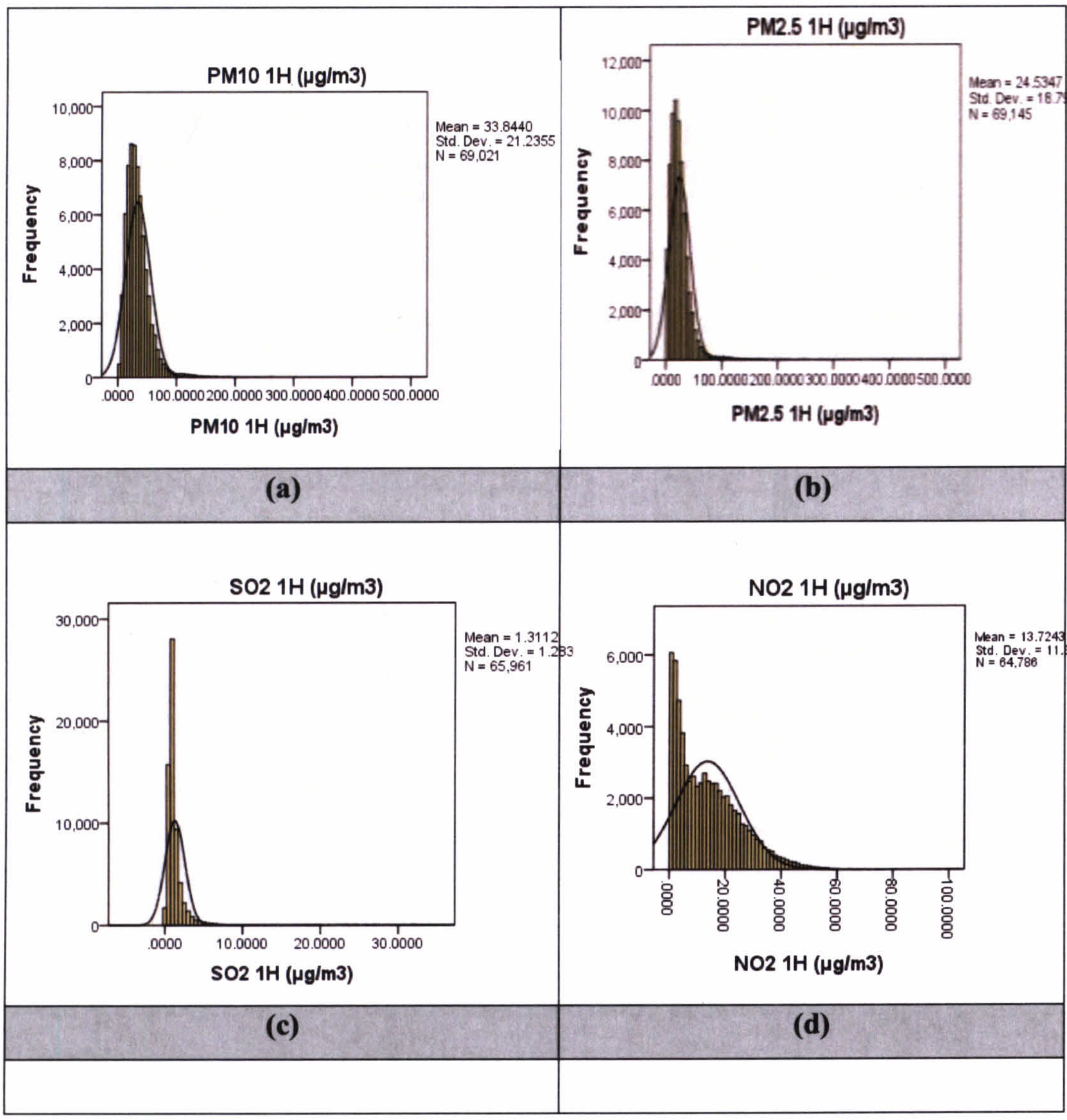
Table 4.1
Missing Values based on important variables

Variables	Total Observation	Valid	Missing	Percentage Missing Values (%)
PM_{10}	69,997 Observation	69,021	976	1.394
$PM_{2.5}$		69,145	852	1.217
Wind Speed		69,237	760	1.086
Humidity		69,506	491	0.701
Temperature		68,732	1265	1.807
SO_2		65,961	4036	5.766
NO_2		64,786	5211	7.445
NO_x		64,380	5167	7.382
O_3		66,206	3791	5.416
CO		65,876	4121	5.887

Table 4.1 show the missing values based on the important variables. These important variables are vital in achieving the objectives. Therefore, there are 69,997 observation in each variable. Valid means real data consist of the variables while missing values occurs if there is no data on the observation. The highest percentage missing value is variable NO_2 while the lowest is humidity. Since the variables are numerical, it can be replaced with mean or median of each variable through imputation.

4.2.2 Distribution of the data

Based on figure 4.1, it shows the distribution of each variables. The distribution can either be positively skewed, negatively skewed, normally skewed or uniform. It also can show an insight towards the data itself.



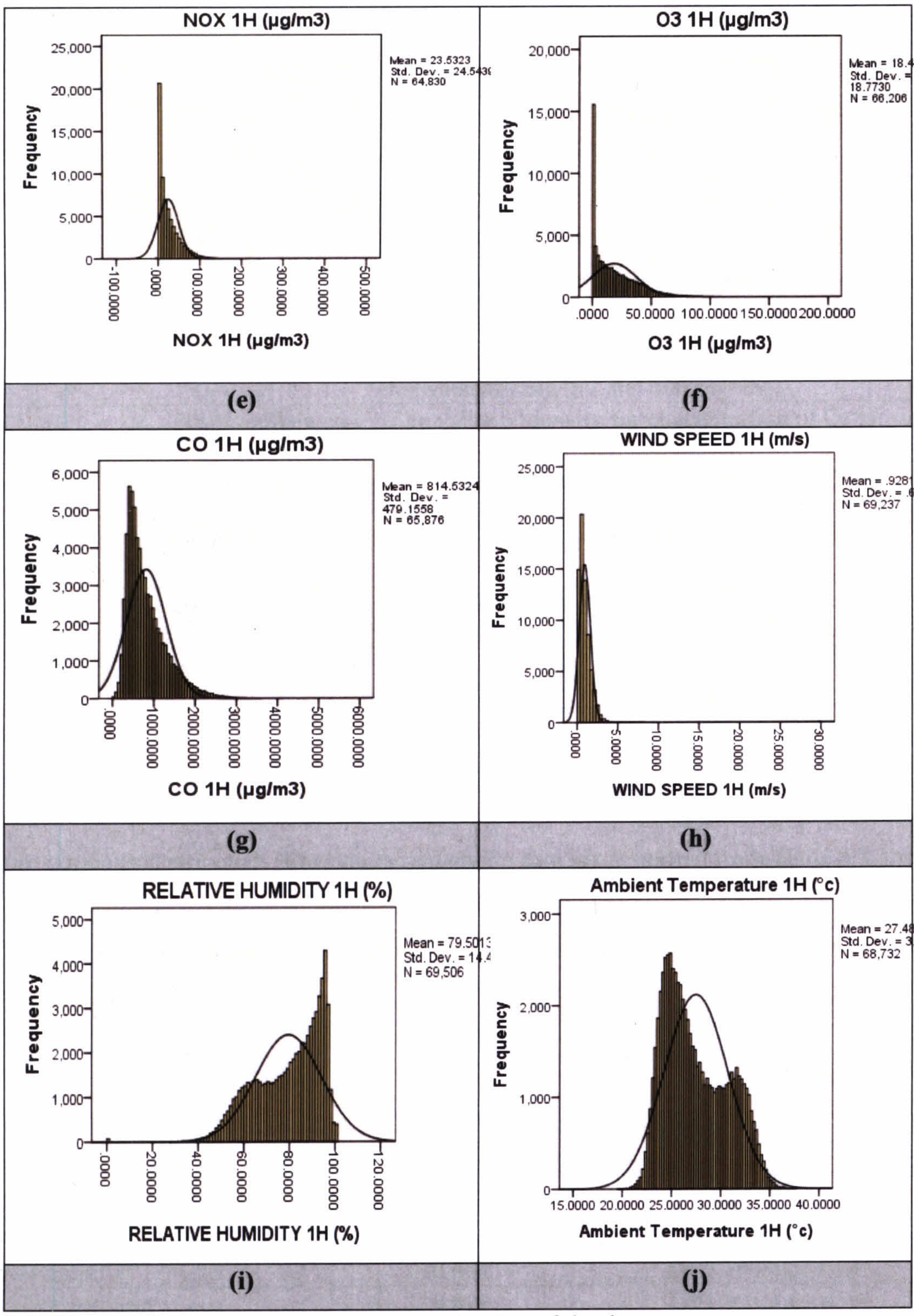


Figure 4.1 Distribution of the data

Table 4.2 showed below is to simplify the results obtained from figure 4.1. Each variable's insight is detailed in depth too.

Table 4.2
Distribution of the data

Figure	Distribution	Insight
a	Positive Skewed	The value of density for concentration of PM_{10} is between 0 mg/ m³ and 500 mg/m³
b	Positive Skewed	The value of density for concentration of $PM_{2.5}$ is between 0 mg/ m³ and 500 mg/ m³
c	Positive Skewed	The value of density for concentration of SO_2 is between 0 mg/ m³ and 30 mg/ m³
d	Positive Skewed	The value of density for concentration of NO_2 is between 0 mg/ m³ and 100 mg/ m³
e	Positive Skewed	The value of density for concentration of NOx is between -10 mg/ m³ and 100 mg/ m³
f	Positive Skewed	The value of density for concentration of O_3 is between 0 mg/ m³ and 200 mg/ m³
g	Positive Skewed	The value of density for concentration of CO is between 0 mg/ m³ and 6000 mg/ m³
h	Positive Skewed	The value of density for wind speed is between 0 m/s and 30 m/s
i	Negative Skewed	The value of density for Humidity is between 20% and 100%
j	Negative Skewed	The value of density for temperature is between 20 °C and 40°C

Kolmogorov-Smirnov test were conduct as in Table 4.3 to further prove that the data is not normally distributed for all the variables. If the p-value is less than 0.05, the data is not normally distributed. However, although the data is not normal from Table 4.3, the skewness value for the descriptive summary after imputation is between -3 to 3. The data were considered as normally distributed since the value of skewness is between -3 and 3 (Jones,1969) .

Table 4.3
Kolmogorov-Smirnov Test

Variables	Sig value	Interpretation
PM_{10}	0.001	Data is not normal
$PM_{2.5}$	0.001	Data is not normal
Wind Speed	0.001	Data is not normal
Humidity	0.001	Data is not normal
Temperature	0.001	Data is not normal
SO_2	0.001	Data is not normal
NO_2	0.001	Data is not normal
NO_x	0.001	Data is not normal
O_3	0.001	Data is not normal
CO	0.001	Data is not normal

4.2.3 Outlier Detection

Based on Table 4.4 below, it shows the summary of outlier for the dataset by observation. The method uses in detecting the outlier is mahalanobis distance based on each location for each observation. If the probability value of mahalanobis distance is lesser than 0.001, then there exists outlier for the given observation in the dataset. The summary of outlier is shown as follow.

Table 4.4
Summary of outlie for the given dataset

Total observation (Outlier)	**Total Observation by each location (Outlier)**			
	Jerantut	Shah Alam	Petaling Jaya	Banting
2633 observation	133 observation	619 observation	1095 observation	786 observation

Petaling Jaya consist of the highest value of observation that consist of outlier at 1095 observation while the lowest is Jerantut at 133 observation. The total observation that exists outlier for the overall dataset is 2633 observation. Therefore, this observation will be deleted before further analysis proceeds.

4.2.4 Summary of the dataset.

Table 4.5 above shows the descriptive analysis for each variable. From the meteorological factors of view which are (Wind Speed, Humidity and Temperature), the highest mean is Humidity while the lowest is Wind Speed . For the maximum value, the lowest is Wind Speed while the highest is Humidity. The lowest variance is wind speed (0.4m/s) and the highest is Humidity (209.0). By comparing air particles also (PM_{10}, $PM_{2.5}$,SO_2,NO_2,NOx,O_3 and CO), the highest mean is CO while the lowest is SO_2 .Therefore, the concentration of CO is higher compared to any other air particles. Concentration for the maximum value is lowest for SO_2 while the highest is CO (5267μg/ m^3). Finally, the lowest variance for air particles is SO_2 while the highest is CO.

Table 4.5
Central tendency for the original dataset

Variables	Mean	Median	Maximum
PM_{10}	33.8	29.9	447.2
$PM_{2.5}$	24.5	21.0	409.5
Wind Speed	0.9	0.7	29.7
Humidity	79.5	82.8	100.0
Temperature	27.4	26.7	37.6
SO_2	1.3	1.0	31.8
NO_2	13.7	11.5	96.8
NOx	23.5	15.2	448.5
O3	18.4	12.9	155.9
CO	814.5	682.0	5267.0

4.3 Data Preparation (Analysis)

Details in data preparation for analysis are shown here below to have an accurate analysis before further proceed with. The process involved is handling missing values, multicollinearity and outlier.

4.3.1 Handling Missing values

Missing data are common when undertaking short-term (24 h) monitoring of air contaminants with real-time monitors, especially in resource-constrained settings. There are several missing cases for each important variable based on Table 4.1 .

Since the distribution of the data is mostly highly skewed either it is to the right or left, median is a better representation of replacing missing data. Therefore, median is the best method in replacing the missing data. The data are replaced as follows in Table 4.6. Table 4.6 shows the value will be replaced for each variable in the missing cases.

Table 4.6
Replacing the missing data

Variables	Missing	Missing value replace by
PM_{10}	976	Median (29.9)
$PM_{2.5}$	852	Median (21.0)
Wind Speed	760	Median (0.7)
Humidity	491	Median (82.8)
Temperature	1265	Median (26.7)
SO_2	4036	Median (1.0)
NO_2	5211	Median (11.5)
NO_x	5167	Median (15.2)
O_3	3791	Median (12.9)
CO	4121	Median (682.0)

4.3.2 Handling Multicollinearity

If the VIF value is greater than 10 and the Tolerance value is less than 0.1, it implies that a multicollinearity problem occurs. The value of VIF are shows as below in Table 4.7 in detecting multicollinearity.

Table 4.7
Value of VIF.

Variables	Value of VIF
PM_{10} 1H ($\mu g/m^3$)	27.0
$PM_{2.5}$ 1H ($\mu g/m^3$)	26.5
WIND SPEED 1H (m/s)	1.3
RELATIVE HUMIDITY 1H (%)	7.3
Ambient Temperature 1H (°c)	7.7
SO_2 1H ($\mu g/m^3$)	1.1
NO_2 1H ($\mu g/m^3$)	3.1
NO_x 1H ($\mu g/m^3$)	2.9
O_3 1H ($\mu g/m^3$)	2.5

Since PM_{10} and $PM_{2.5}$ has high values of VIF which is greater than 10, it means that both variables were connected to each other. Thus, one of the variable were needed to remove for further analysis. Therefore, $PM_{2.5}$ are dropped from the study when comes to factor analysis. However, $PM_{2.5}$ are still being used for descriptive analysis. Table 4.8 shows the value of VIF if variables of $PM_{2.5}$ was eliminated

Table 4.8
New VIF value

Variables	Value of VIF
PM_{10} 1H ($\mu g/m^3$)	1.2
WIND SPEED 1H (m/s)	1.3
RELATIVE HUMIDITY 1H (%)	7.3
Ambient Temperature 1H (°c)	7.7
SO_2 1H ($\mu g/m^3$)	1.1
NO_2 1H ($\mu g/m^3$)	3.0
NO_x 1H ($\mu g/m^3$)	2.9
O_3 1H ($\mu g/m^3$)	2.5

4.3.2 Handling Outlier

Since the outlier were obtained by using mahalanobis distance in sub-section 4.2.3, these observation that consist of outlier by each location will be deleted. The summary of the overall dataset for each location were shown in Table 4.9 after deleted this observation since the location were taken into consideration when detecting outlier.

Table 4.9
Summary of overall dataset after removing outlier

Location	Total observation	
	Before removing outlier	After removing outlier
Jerantut	17520 observation	17387 observation
Shah Alam	17512 observation	16892 observation
Petaling Jaya	17481 observation	16386 observation
Banting	17484 observation	16697 observation
Total Observation :	69,997 observation	67362 observation

4.4 Data Descriptive (After Imputation)

In this summary, it shows the summary of the data after imputation and an insight on what is really happening on our dependent variable based on certain categories

4.4.1 Descriptive summary (Before vs After Imputation for each location)

Table 4.10 below shows the descriptive analysis for each variable after and before imputation. **A** were labelled as after imputation while **B** is before imputation

Table 4.10
Descriptive summary of data before and after imputation

Location	Variables	Mean		Median		Max		Skewness	
	Imputation	A	B	A	B	A	B	A	B
Jerantut	PM_{10}	22.4	22.9	19.2	19.1	120.1	447.3	1.9	3.584
	$PM_{2.5}$	15.0	15.4	11.7	11.6	106.5	409.6	2.1	4.016
	Wind Speed	1.1	1.1	0.9	0.9	4.1	6.3	1.1	1.325
	Humidity	80.5	80.5	84.6	84.7	97.4	97.4	-0.7	-0.731
	Temperature	27.0	27.0	26.2	26.2	36.8	36.8	0.5	0.515
	SO_2	0.9	0.9	0.9	0.9	5.5	5.5	1.9	1.922
	NO_2	3.1	2.7	2.3	2.1	22.2	22.2	1.7	1.495
	NOx	4.0	3.4	2.9	2.7	60.9	60.9	2.4	3.218
	O_3	17.3	17.6	14.7	15.6	59.9	66.8	0.7	0.673
	CO	443.5	432.0	426.0	419.0	1670.0	1670.0	0.8	1.150
Shah Alam	PM_{10}	36.7	38.2	33.6	34.4	135.0	264.6	1.1	2.4
	$PM_{2.5}$	27.3	28.7	24.6	25.1	115.9	252.3	1.4	3.0
	Wind Speed	0.8	0.9	0.7	0.7	3.8	7.9	1.1	1.5
	Humidity	78.5	78.2	82.0	81.5	98.0	98.0	-0.6	-0.5
	Temperature	27.4	27.5	26.7	26.8	36.2	36.3	0.5	0.4
	SO_2	1.2	1.3	1.0	1.0	7.4	31.8	2.4	5.8
	NO_2	17.2	17.9	15.3	16.7	61.6	71.0	0.7	0.7
	NOx	27.0	28.4	20.9	23.1	133.4	186.8	1.3	1.4
	O_3	20.4	21.8	12.9	14.0	98.5	155.9	1.1	1.2
	CO	914.7	940.6	829.0	860.	3508.0	3508.0	0.9	1.0
Petaling Jaya	PM_{10}	36.2	38.1	33.6	34.6	131.0	229.5	1.4	2.6
	$PM_{2.5}$	26.3	27.9	23.8	24.5	115.0	219.3	1.7	3.3
	Wind Speed	0.7	0.8	0.6	0.6	3.6	29.7	1.4	9.5
	Humidity	76.9	76.9	79.6	79.5	98.0	98.0	-0.5	-0.5
	Temperature	28.1	28.1	27.4	27.4	37.7	37.7	0.4	0.4
	SO_2	1.5	1.7	1.0	1.1	7.6	27.1	2.0	4.1
	NO_2	22.3	23.7	21.2	22.3	60.9	96.8	0.5	0.7
	NOx	43.1	46.4	39.6	42.4	138.6	448.5	0.6	1.4
	O_3	14.7	14.6	9.0	7.5	95.1	109.8	1.4	1.4
	CO	1127.6	1195.	1030.0	1095.	4520.0	4538.0	0.8	0.9
Banting	PM_{10}	34.7	36.2	30.6	31.4	134.2	304.0	1.3	2.5
	$PM_{2.5}$	24.8	26.2	21.0	21.5	116.7	279.6	1.5	2.7
	Wind Speed	1.0	1.0	0.8	0.8	4.1	8.4	0.9	1.3
	Humidity	83.0	82.4	86.8	86.6	100.0	100.0	-0.7	-1.3
	Temperature	27.2	27.3	26.6	26.4	36.4	37.2	0.5	0.5
	SO_2	1.2	1.4	1.0	1.0	7.5	23.1	2.4	4.9
	NO_2	10.7	10.8	10.8	10.0	47.6	53.9	0.7	0.8
	NOx	16.0	16.3	15.1	13.7	122.5	127.4	1.2	1.4
	O3	18.9	19.6	12.9	13.9	93.9	102.8	0.9	0.9
	CO	677.2	686.3	649.0	625.	5267.0	5267.0	1.4	1.5

From the Table 4.10 above, it shows the result between before and after imputation for each location. The mean after imputation is less than before the mean before imputation for Banting, Shah Alam and Petaling Jaya is less than before imputation for Banting, Shah Alam and Petaling Jaya but not in a significant way. The result also applied to its median too. However, both mean and median has much higher value at after imputation compared to before imputation.

The maximum value for each pollutant on each location does not differ from one another for both before and after imputation. The skewness of the data decrease after the imputation and can be considered as normally distributed since the value of skewness is between -3 and 3 (Jones,1969) However, the skewness value given is to show that the data preparation is effective since the skewness value decrease hence limiting the extreme value for the given dataset. Therefore, it is safe to say that the data is in good condition to proceed with factor analysis,

4.4.2 Concentration of CO based on different location

Figure 4.2 shows the concentration of CO in different categories of location. From the figure, Jerantut has the least concentration of CO at 445.8 while Petaling Jaya has the highest concentration of CO at 1169.1 . The reason why concentration of CO is less in Jerantut is due to less traffic on the road since it is background area. Petaling Jaya has the highest CO is due to that the is surrounded by many industries, residential and commercial areas (Azmi et al., 2010).

Even Abdullah et al., (2012) has highlighted that High levels of CO in Petaling Jaya are probably related to motor vehicle emissions. This shows that motorcycle plays an important part in increasing the concentration of CO in a particular area. Motorcycles and passenger cars are the dominant classes of motor vehicles, with 6,572,366 units in 2004 that increased by 35% to 11,989,591 units in 2015 (Shafie et al., 2020).

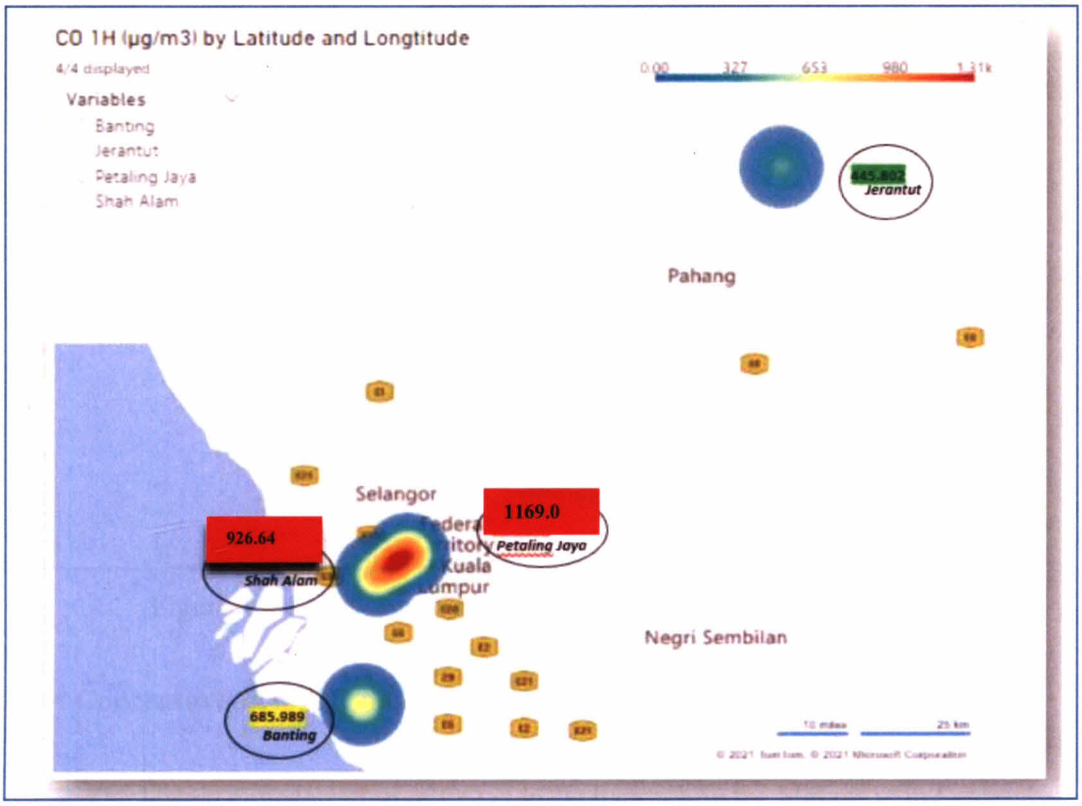


Figure 4.2 Concentration of CO based on different location

4.4.3 Concentration of CO based on different period

The largest amount of CO is between the period of 9pm to 11pm and 12am until 2am. Surprisingly, the concentration is very high during that time due to that it corresponds to the emissions from by early morning and late evening traffic. (Abdullah et al., 2015) The gas emission was accumulating during that time and were influence by meteorological factors when the station recorded the value.

Not only that, the researcher could compare the concentration of CO during lunchtime, people going to work in the morning and people coming back from work in the late evening since traffic is saturated during that time. The researcher could identify that the concentration is higher in people going to work in the morning accumulate about 14% of the total concentration of CO while the lowest is during lunchtime.

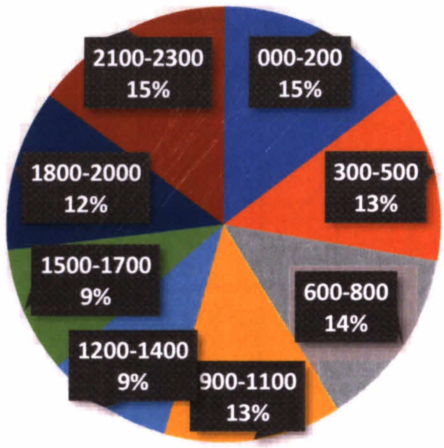


Figure 4.3 Concentration of CO based on different time period

4.4.4 Concentration of CO based on different month

The lowest concentration of CO is in January while the highest is in September. CO is affected by different month due to meteorological factors. For example, Malaysia is subject to two monsoon wind seasons: The Southwest Monsoon and the Northeast Monsoon. The Southwest Monsoon occurs from June to September while the Northeast Monsoon occurs from August to March.

The south-west monsoon wind which occurs during the months from June to September transport the air pollutants from the south-west corner of Southeast Asia and influences the air quality in Malaysia (Khan et al., 2016) . That is the reason why concentration of CO starts to increase from June until it reaches the peak in September. Not only that, highest combustion of vehicles or haze were most likely to occur during the months of June to September every year, when the weather is dry (Latif et al., 2014)

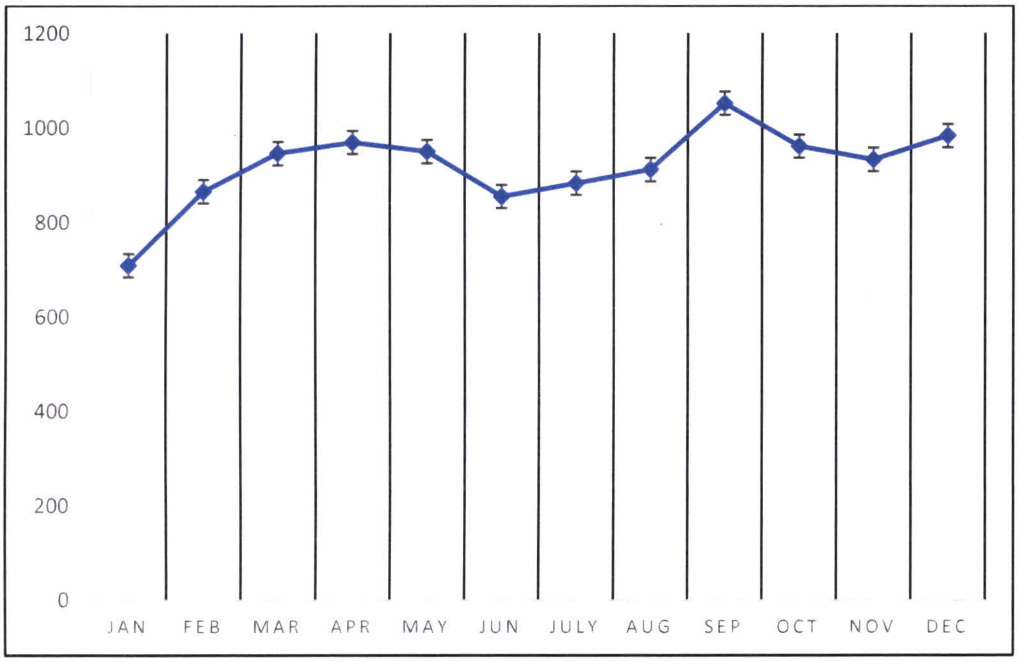


Figure 4.4 Concentration of CO based on different months

4.5 Factor Analysis

4.5.1 Determinant (Analysis)

One of the assumption is that the determinant (located under the correlation matrix) should be more than 0.0001. The value of determinant by each category of location are as in Table 4.11. Since the value is more than 0.001 for each location, the factor analytics can be obtained. Therefore, the assumption is satisfied for each location.

Table 4.11
Value of determinant in each location

Location	Value of Determinant
Jerantut	0.002
Shah Alam	0.002
Petaling Jaya	0.002
Banting	0.004

4.5.2 Bartlett's test of sphericity (Analysis)

The Bartlett test should be significant which means that the variables are correlated highly enough to provide a reasonable basis for factor analysis. The Bartlett test is significant where significance value (sig value < 0.05) which is good and indicates that the correlations is not near zero. The bartlett's test of sphericity for each location are shown as Table 4.12.

Table 4.12
Value of bartlett's test for each location

Location	Significance value
Jerantut	0.001
Shah Alam	0.001
Petaling Jaya	0.001
Banting	0.001

Since the sig value is less than 0.05, it means that the matrix is not identity. Therefore, the assumption is met in proceeding with factor analysis for each location.

4.5.3 Kaiser-Meyer- Olkin (Analysis)

The value of KMO is given in table 4.13 for each location. The value of KMO for each location are as follows.

Table 4.13
Value of KMO in each category of location.

Location	KMO Value	Degree of common variance
Jerantut	0.660	Mediocre
Shah Alam	0.799	Middling
Petaling Jaya	0.748	Middling
Banting	0.819	Meritorious

Since KMO value higher than 0.05, the value the overall data set are considered good enough for the sampling adequacy. Thus, the final assumption for factor analysis is met before factor analysis were used to reduce variables into factors.

4.5.4 Scree Plot (Analysis)

The scree plot shows the number of factors and their related eigenvalues. The eigenvalues are sorted from largest to smallest in the scree plot. The eigenvalues of the correlation matrix equal the variances of the elements when no rotation is performed. Factors are based on the component number which has the value of eigenvalue that is more than 1. It means that the factors component account for the most of the total variability of the data. Component that has eigenvalue less than 1 are considered not important to classify it as a factor since very small part of the variability taken place and most likely insignificant for the study. Explanation for the scree plot is explained in Table 4.14 based on Figure 4.5 .

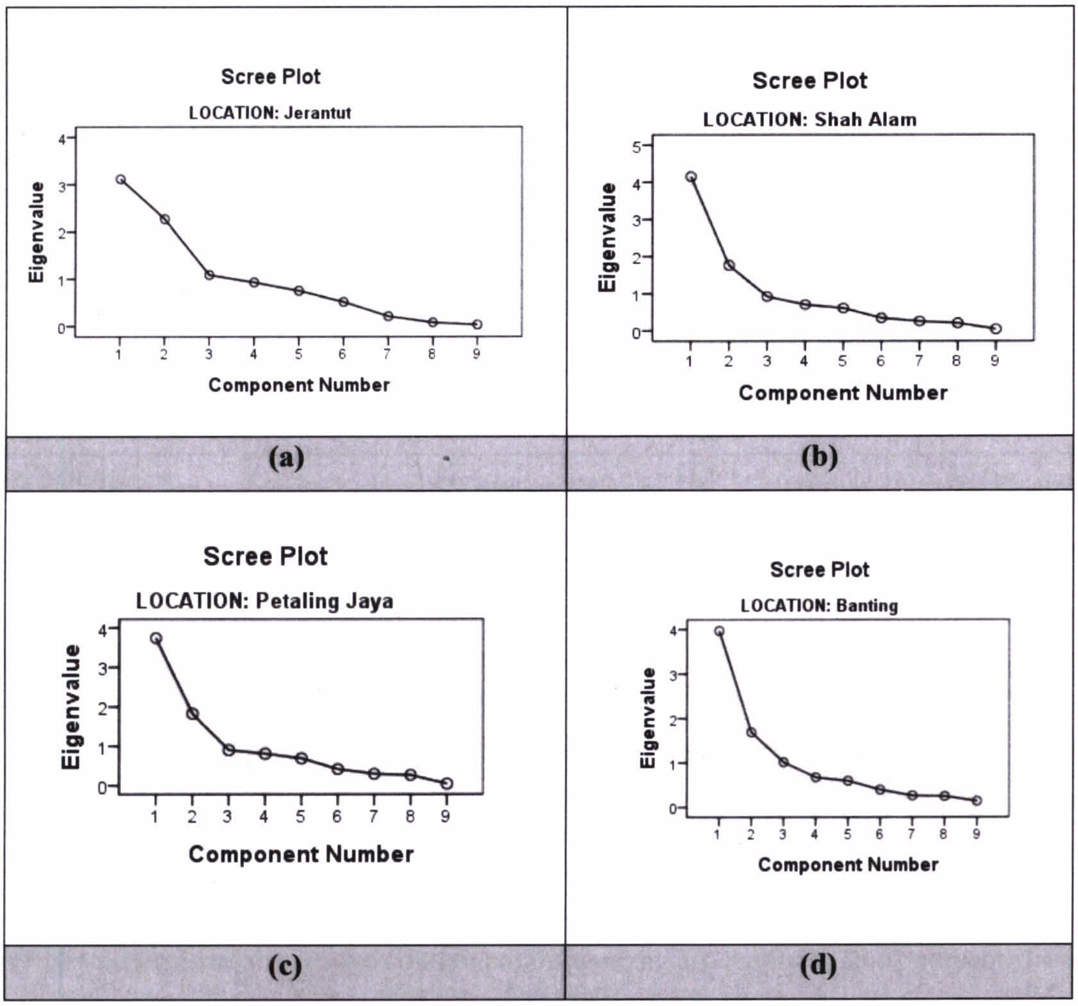


Figure 4.5 : Scree plot based on each location

Table 4.14
Scree plot interpretation

Figure 4.5	Interpretation
a	There are 3 factors were taken into consideration in Jerantut
b	There are 2 factors were taken into consideration in Shah Alam
c	There are 2 factors were taken into consideration in Petaling Jaya
d	There are 2 factors were taken into consideration in Banting

4.5.5 Total variance explained (Analysis)

Total variance explained is a more detail explanation from the scree plot. The specific eigenvalue, percentage for each variance components and cumulative percentage of variance are obtained here. The outcome is shown in Table 4.15 below for each location based on the number of factors gathered

Table 4.15
Total variance explained for each location

Location	Components	Initial Eugenvalue		
		Total Eigenvalues	% of Variance	Cumulative % Variance
Jerantut	1	3.13	34.85	34.85
	2	2.26	25.21	60.07
	3	1.07	11.93	72.00
Shah Alam	1	4.30	47.80	47.80
	2	1.76	19.61	67.41
Petaling Jaya	1	3.92	43.64	43.64
	2	1.89	21.10	64.74
Banting	1	4.21	46.82	46.82
	2	1.69	18.83	65.65

Table 4.15 demonstrates the results after the extraction was done for each location. Jerantut have three factors which has a total variance of 72.0% . Form those total variance, 34.85% influenced the first component, 25.21% are influenced by the second component and 11.93% are influenced by the last component. Shah Alam has two factors with a total variance of 67.41%. The first factor component comprises at 47.801 percent and the second component influencing air quality at 19.61 percent. Two factors in Petaling Jaya were discovered to have a total variance of 64.74 percent. It shows that the first component influencing 43.64% percent and the second component influencing 21.10 percent. Finally, two of Banting's factors were discovered to have eigenvalues greater than one which are 4.21 and 1.69. It demonstrates that the first component influencing 46.82 percent and the last component influencing 18.33 percent.

4.5.6 Factor loadings (Analysis)

The Rotated Factor Matrix table is essential for comprehending the analysis's findings. Factors are rotated so that they may be concluded more easily. Rotation is used as much as feasible to explain various items or variables by different underlying factors. Furthermore, each component explains more than one item. Table 4.16 shows the result of Rotated factor matrix obtained from the overall dataset based on each location. The table only displays the value for each variable with a value greater than 0.5 in making it easier to determine which variables should be included in each factor.

Table 4.16
Rotated Factor Matrix obtained from each location

Location	**Jerantut**			**Shah Alam**		**Petaling Jaya**		**Banting**	
Factors	1	2	3	1	2	1	2	1	
PM_{10} 1H			0.682		0.713		0.643		0.782
WIND SPEED	0.604			0.645		0.560		0.548	
Relative humidity	-0.924			-0.933		-0.934		-0.916	
Ambient Temperature	0.916			0.948		0.935		0.924	
SO_2 1H			0.649				0.542		
NO_2 1H		0.950			0.790		0.803		0.710
NO_x 1H		0.957			0.762		0.729		0.734
O_3 1H	0.899			0.841		0.872		0.831	
CO 1H			0.781		0.724		0.701		0.728

The primary results of factor analysis for each site based on the 10 major air pollutants are shown in Table 4.13. Three primary factors were discovered for the location Jerantut which is located in a small town and district in Pahang. These factors explained 72.0 percent of the variance in the data. Wind speed, O_3, temperature, and humidity are the variables included for the first factor in Jerantut. Humidity has the biggest effect in factor 1 while Wind Speed is the lowest . NO_2 and NO_x are the second factor in Jerantut. NO_x has the biggest contribution in factor 2 while NO_2 is the lowest. CO, SO_2, and PM_{10} make up the third factor in Jerantut. CO has the biggest contribution in the third factor while SO_2 has the lowest impact. Factor 1 is the most important factor influenced by O_3 and meteorological parameters accounting for 34.8 percent of the total. Since 0_3 requires meteorological conditions as well as the presence of sunshine to react, it is a trustworthy result (Lam et al.,2012). Furthermore, as noted by Geddes et al.,(2009), meteorological conditions such as local temperature at each sampling station have the potential to enhance ozone levels in the atmosphere. Additionally, meteorological elements such as wind speed, temperature, humidity and UV play a crucial part in air pollution because these factors influence the concentration of air pollutants in Jerantut. The Southwest monsoon has been found to be capable of transporting air pollutants from urbanised areas on the west coast of the Malaysian Peninsular including Kuala Lumpur city centre to the rural background research area (Latif et al., 2014). Humidity is the most important component in this factor since it directly controls the dispersion of air contaminants in the environment (Isiyaka & Azid,2015). Factor 2 is classified as air pollution that were influenced by power plants. Pahang have several power plant that could alter the concentration of air pollution in Jerantut. Biogas Power Plant 1.5mw Cenergi FJP Sdn Bhd in Jerantut, Ulu Jelai Hydroelectric Power Plant in Cameroon Highland and Jengka Advance Renewable Energy Plant in Bandar Tun Razak were some examples of power plant located not far away from Jerantut. Power plant boilers produce about 40% of the NOx emissions from stationary sources(Cox, 1999). Although NO_2 is mainly comes from transportation but it is also a source coming from power plant. A study by Metia et al., (2018) stated that NO_2 is evident near emissions sources from power plant that could affect air quality. Finally, the last factor is traffic emission which consist of CO, SO_2 and PM_{10}. The 'least important' group is influenced by motor vehicles since the results of Latif et al., (2014) strongly imply that the increasing volume of traffic in Jerantut has not really to substantially damage the air quality level because the air pollution indicators for motor vehicles were classed in the low-moderate groups. Not only that, it also stated that traffic emission from Jerantut were caused by traffic in Jerantut town. CO was the primary contributor from this factor since it was mainly produced by the incomplete combustion of motor vehicle fuels (Dominick et al.,2012). According to Rahman et al., (2015), the transportation sector accounted for 77 percent of global CO emissions.

Furthermore, Shah Alam have 2 factors identified from the analysis. These factors explained 67.4 percent of the variance in the data. Wind speed, O_3, temperature, and humidity are the variables included for the first factor in Shah Alam. Temperature has the biggest effect in factor 1 while wind Speed is the lowest .CO,PM_{10}, NO_2 and NOx make up the final factor in Shah Alam. NO_2 has the biggest contribution in the factor while PM_{10} is the lowest. The first factor includes O_3 and meteorological parameters. This factor is the most important in Shah Alam since it accounts for 47.8 percent of the total variance. There are a few reasons why meteorological elements play a role in Shah Alam air pollution. According to the findings of Ahmat et al. (2019), different meteorological parameters influenced the concentrations of air pollutants in Shah Alam and Klang. Humidity was found to influence PM_{10} and CO concentrations in Shah Alam. Not only that, high temperatures combined with high Ultraviolet (UV) levels in Shah Alam contribute to higher concentrations of traffic emissions such as PM_{10}, CO, and other more. (Payus et al., 2013). O_3 was included as meteorological factors since O_3 requires meteorological conditions as well as the presence of sunshine to react. Temperature has the greatest contribution to this factor as Barmpadimos et al. (2011) highlighted that the concentration of some pollutants appears to rise during high temperatures and may be inferred from different monsoon directions. The final factor for Shah Alam is related to traffic emission. The traffic emission in Shah Alam consist of CO,PM_{10}, NO_2 and NOx. Shah Alam stations are placed on major roadways in industrial and high-density residential districts, which are regularly affected by traffic-related pollution (Ahamad et al., 2012). The highest contribution of traffic emission in Shah Alam is NO_2. Previous research has revealed that NO_2 is a prominent source of outdoor air pollution in Shah Alam (Rahman et al., 2015). It is primarily due to emissions caused by traffic congestion (Isiyaka et al., 2015). As a result, the main source of NO_2 in high-traffic areas was motor vehicle emissions. This claim was backed up by a study conducted by Dominick et al (2012) who concluded that NO_2 is a consequence of heavy traffic and is released by motor vehicles because traces of nitrogen impurities in fuel can oxidise into NO_2 which that the amount increases with vehicle load and speed.

Also, the investigation found two factors for Petaling Jaya. The first factor includes wind speed, O_3, temperature, and humidity. Temperature has the biggest effect in factor 1 while wind speed is the lowest . CO,PM_{10}, NO_2 and NOx make up the final factor in Petaling Jaya. NO_2 has the biggest contribution in the factor while SO_2 is the lowest. The first factor consists of O_3 and meteorological parameters. All pollutants have a strong relationship with meteorological variables such as relative humidity, ambient temperature, and wind speed in Petaling Jaya (Rahman et al., 2015). Since temperature is the most major contributor in Factor 1, Verma and Desai (2008) concluded that high temperatures contribute to a high amount of air pollution dispersion. Not only that, Shaharuddin et al. (2008) found that high temperature with low frequency is associated with high PM_{10} and NO_2 levels in Petaling Jaya. Furthermore, high temperatures create favourable conditions for the reaction of nitrogen oxides (NOx) and other VOCs. The final factor for Petaling Jaya is related to traffic emission. Traffic emission includes CO,PM_{10}, NO_2 ,NO_x and SO_2. These pollutants are always high at Petaling Jaya which is due to influence of heavy traffic (Azmi et al., 2010). There were 4.6 million registered motor vehicles in Kuala Lumpur and 2.2 million registered motor cars in Selangor with the majority concentrated in major cities such as Petaling Jaya. SO_2 is often produced by industrial operations (Isiyaka & Azid, 2015) but in Petaling Jaya, the main source of SO_2 is predicted to be motor vehicles specifically diesel-engined trucks and buses (Mohamad et al., 2015). ANOVA tests revealed a significant relationship between SO_2 levels and the volume of cars travelling through the area. However, NO_2 contributes the most to traffic emissions. Due to traffic emissions, the most affected areas' NO_2 and NO_x distributions were all in the centre of the KLEMUR, where Petaling Jaya station is situated (Halim et al., 2020). NO_2 in Petaling Jaya indicates that the oxidation of gasoline from nitrogen oxides with vehicle load and speed may suggest that traffic congestion in this area is a likely factor to the increase in air pollution (Mohamad et al., 2015). Finally, two factors relating to air quality were discovered in Banting. Wind speed, temperature, humidity and O_3 are the first factors to consider. Temperature has the greatest effect whereas wind speed is the least. The last factor for banting emission which comprises for CO, NO2, NOx and PM_{10}. With a factor loading of 0.78 for PM_{10} make it the most contribution for factor 2 whereas NO_2 is the lowest.

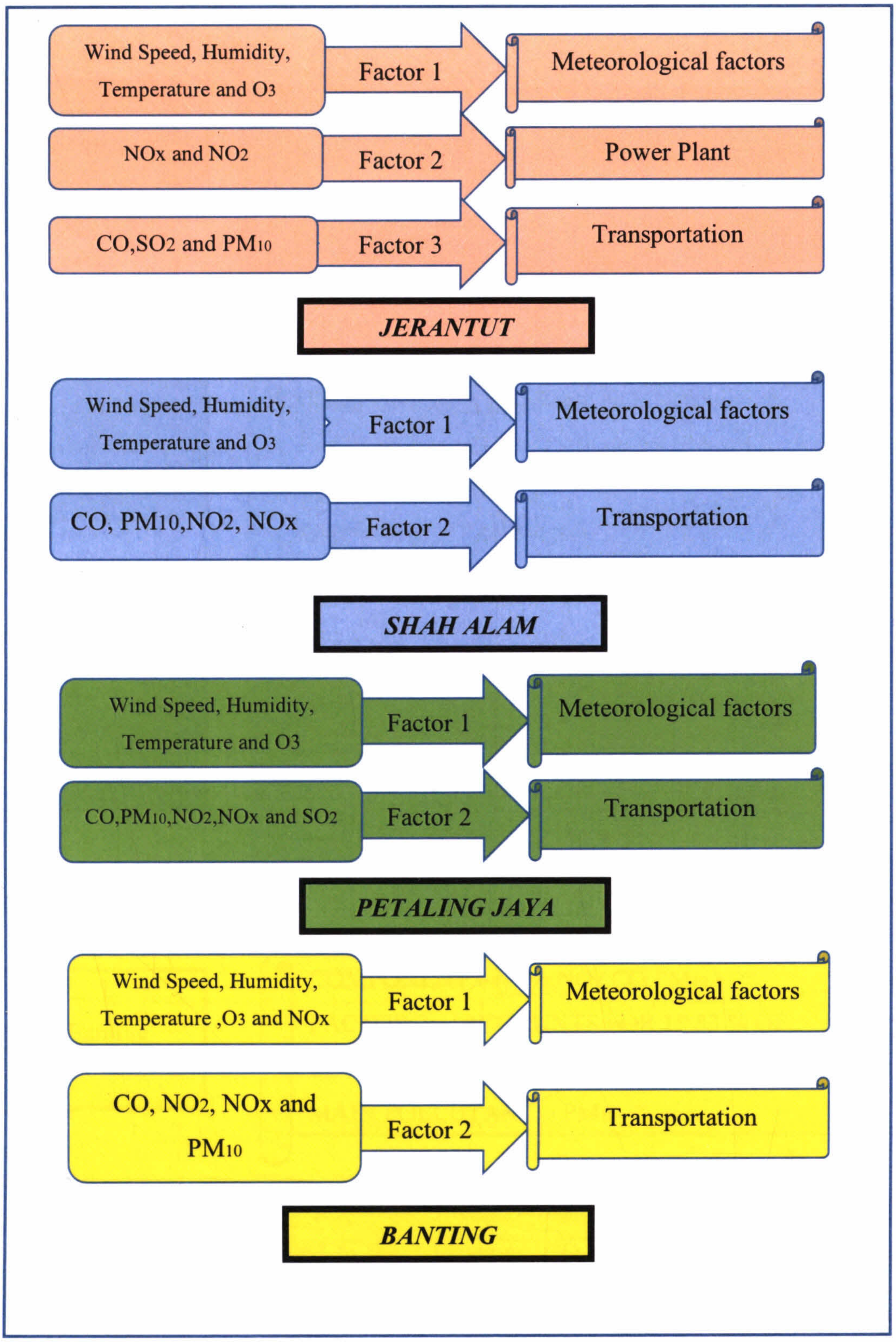


Figure 4.6 Every factor identified for each location

Jerantut

COMPONENT – (CO, SO_2, PM_{10})
FACTOR 3 - (ACCOUNTS FOR 11.93 % OF TOTAL VARIANCE)
MAIN POLLUTANT - CO

Shah Alam

COMPONENT – (CO, PM_{10}, NO_2, NO_x)
FACTOR 2 - (ACCONTS FOR 19.61% OF TOTAL VARIANCE)
MAIN POLLUTANT – NO_2

Petaling Jaya

COMPONENT – ($CO, PM_{10}, NO_2, NO_x, SO_2$)
FACTOR 2 - (ACCOUNTS FOR 21.10% OF TOTAL VARIANCE)
MAIN POLLUTANT – NO_2

Banting

COMPONENT – ($NO_2, NO_x, CO, PM_{10},$)
FACTOR 2 - (ACCOUNTS FOR 18.83 % OF TOTAL VARIANCE)
MAIN POLLUTANT – PM_{10}

Petaling Jaya has the most pollutants of traffic emission. It also has the highest of total variance compared to the 3 locations. CO were included as traffic emission in all categories of location too.

Figure 4.7 TRAP pollution for each location

4.6 Spearman Correlation Analysis.

Spearman correlation was used to identify the relationship between CO and other pollutants for each factor within different location. To simplify the result, Table 4.17 shown below along with the interpretation. If the sig value is less than 0,05, it means that the variable is significant. The highest correlation between CO and other pollutants for each location are discussed below

Table 4.17
Relationship between CO and other pollutants for each factor

Location	Pollutants	Sig value	Correlation coefficient	Interpretation
Jerantut	Wind Speed	0.00	-0.04	Significant and weak negative relationship
	Humidity	0.00	-0.144	Significant and weak negative relationship
	Temperature	0.00	0.154	Significant and weak positive relationship
	O_3	0.00	0.041	Significant and weak positive relationship
	NO_2	0.00	0.274	Significant and weak positive relationship
	NO_x	0.00	0.265	Significant and weak positive relationship
	SO_2	0.00	0.244	Significant and weak positive relationship
	PM_{10}	0.00	0.404	Significant and weak positive relationship
Shah Alam	Wind Speed	0.00	-0.448	Significant and weak negative relationship
	Humidity	0.00	0.399	Significant and weak positive relationship
	Temperature	0.00	-0.360	Significant and weak negative relationship
	O_3	0.00	-0.544	Significant and moderate negative relationship
	NO_2	0.00	0.688	Significant and moderate positive relationship
	NO_x	0.00	0.650	Significant and moderate positive relationship
	PM_{10}	0.00	0.509	Significant and moderate positive relationship

Table 4.17 (Cont'd)

Petaling Jaya	Wind Speed	0.00	-0.336	Significant and weak negative relationship
	Humidity	0.00	0.430	Significant and weak positive relationship
	Temperature	0.00	-0.374	Significant and weak negative relationship
	O_3	0.00	-0.547	Significant and moderate negative relationship
	NO_2	0.00	0.596	Significant and moderate positive relationship
	NO_x	0.00	0.714	Significant and strong positive relationship
	SO_2	0.00	0.184	Significant and weak positive relationship
	PM_{10}	0.00	0.385	Significant and weak positive relationship
Banting	Wind Speed	0.00	-0.415	Significant and weak negative relationship
	Humidity	0.00	0.318	Significant and weak positive relationship
	Temperature	0.00	-0.276	Significant and weak negative relationship
	O_3	0.00	-0.445	Significant and weak negative relationship
	NO_x	0.00	0.540	Significant and moderate positive relationship
	PM_{10}	0.00	0.543	Significant and moderate positive relationship
	NO_2	0.00	0.462	Significant and weak positive relationship
	SO_2	0.00	0.026	Significant and weak positive relationship

CO has a significant value compared to other pollutants in their respective factors at each location. It indicates that there is an association between CO and other pollutants from different categories of location since all of the significant value is less than 0.05. Therefore, interpretation could be made for the strength of correlation between CO and other pollutants from different location In Jerantut, the strongest CO correlation is between PM_{10} at 0.404 which indicates a moderate positive relationship. The sources of these pollutants mainly come from traffic activities in Jerantut town which demonstrate the relationship between them. The greatest CO correlation in Shah Alam is between NO_2 which is 0.688, showing a moderately positive association. It was because of traffic emissions. The level of traffic during working hours was projected to be low as was the levels of NO_2 and CO in the atmosphere during working hours compared to after working hours. Both of these measurements rise as the evening working hour ends and individuals begin to return to work in the morning (Azmi et al.,2010). Regardless of the activities that are taking place in this city , the government and local authorities have taken aggressive actions to control the air quality condition. It is accomplished through effective rules and enforcement directed at the general public and industries in order to maintain good air quality status. Putrajaya's conclusion contrasts from that of Shah Alam, where NOx is the highest related with CO at 0.714 indicating a strong positive association. Since Petaling Jaya is an industrial area, there are many high good vehicles such as trucks and lorries. As a result, according to Shafie et al., (2020), CO and NOx are gaseous pollutants emitted by these types of motor vehicles during the fuel combustion process as petrol and diesel directly affect the quantities of CO and NOx released. Finally, PM_{10} has the highest correlation of CO in Banting with a value of 0.543 indicating a moderately positive relationship. Although PM_{10} is an industrial air particle, the combination of CO and PM_{10} were mostly caused by all classes of motor vehicles as proven by air pollution from tail-pipe exhaust (Shafie et al.,2020).Table 4.18 shows the highest and lowest correlation of CO among other factors.

Table 4.18
Highest and lowest correlation of CO among other factors

Location	**Factors**	**Highest**	**Lowest**
Jerantut	Meteorological parameters	Temperature	Wind Speed
	Power Plant	NO_2	NO_x
	Traffic emissions	PM_{10}	SO_2
Shah Alam	Meteorological parameters	O_3	Temperature
	Traffic emissions	NO_2	PM_{10}
Petaling Jaya	Meteorological parameters	O_3	Wind Speed
	Traffic emissions	NO_x	SO_2
Banting	Meteorological parameters	O_3	Temperature
	Traffic emissions	PM_{10}	NO_2

4.7 Clustering Analysis.

4.7.1 Silhouette Measure (K-mean Clustering)

Figure 4.8 shows the difference value of silhouette measure for cluster 2 until cluster 5 to determine the optimum number of cluster.

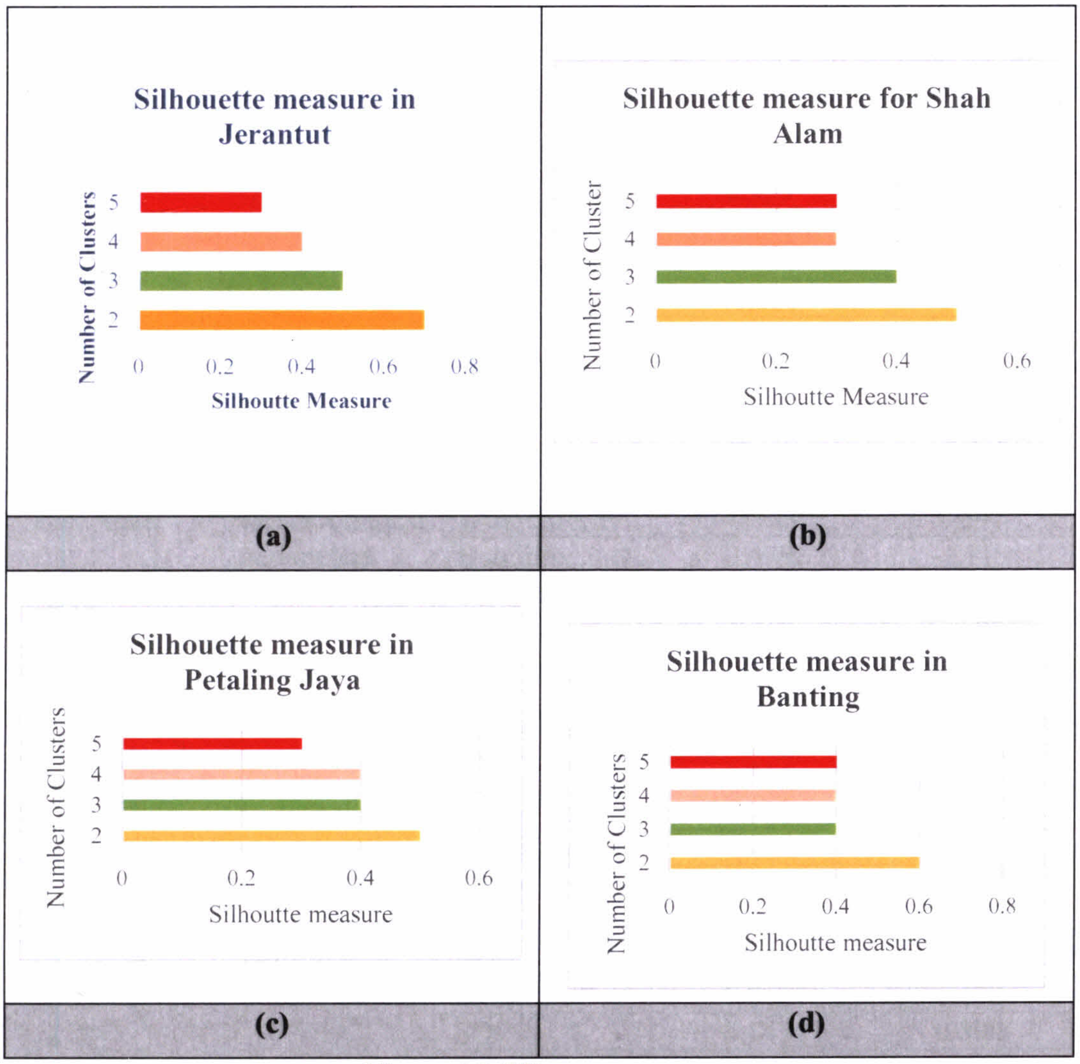


Figure 4.8 Silhouette measure

From the result , every location should have an optimum or specified number cluster of 2 since it has the highest value of Silhouette measure for each location compared to other number of clusters.

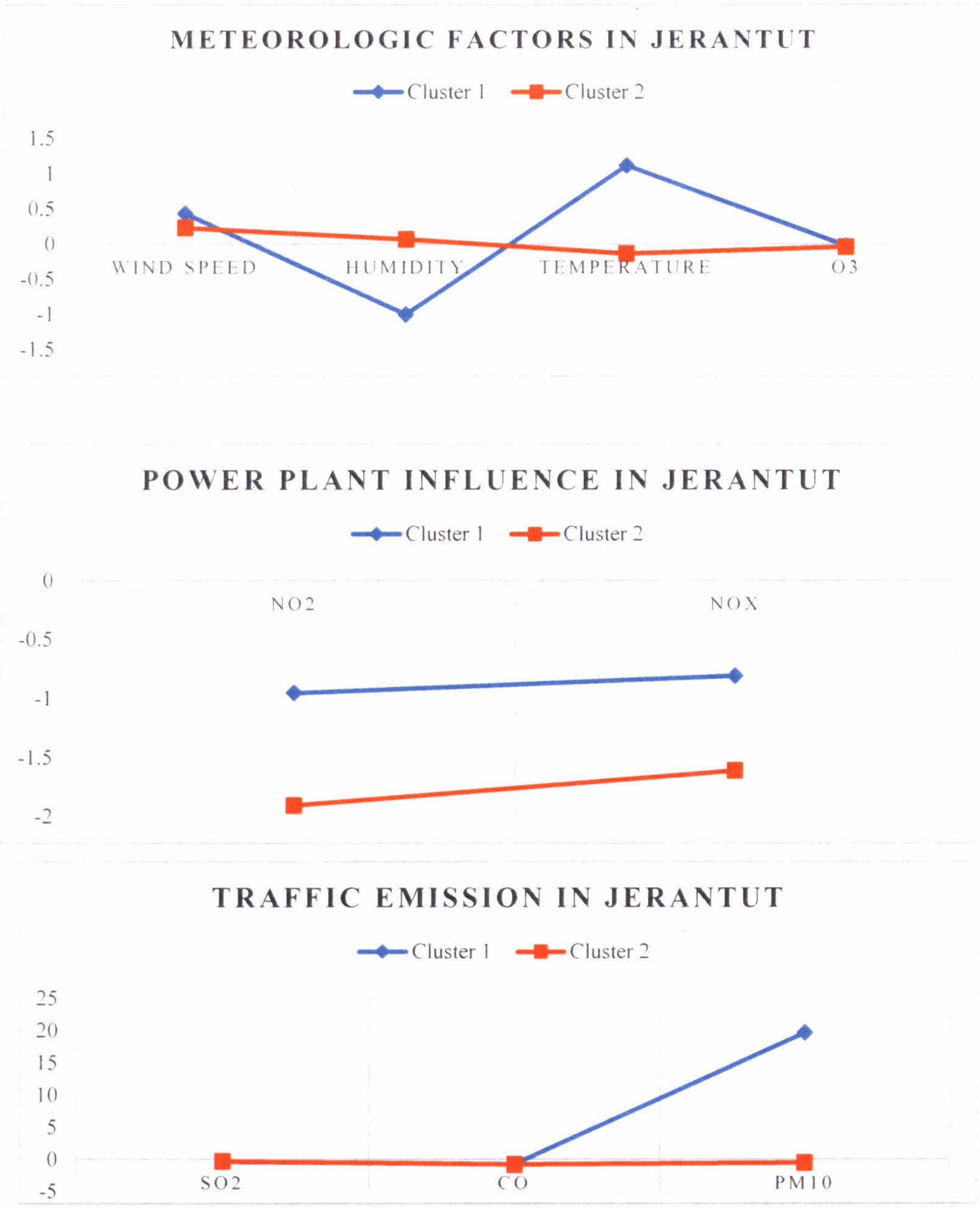
4.7.2 Final Cluster centres (K-means Clustering)

The final cluster centres are computed as the mean for each variable within each final cluster. Each factor for each location is taken into consideration during this phase so that the researcher could highlight and compare important details from objective 1 and 2. The final cluster centres results are as follows in Table 4.19.

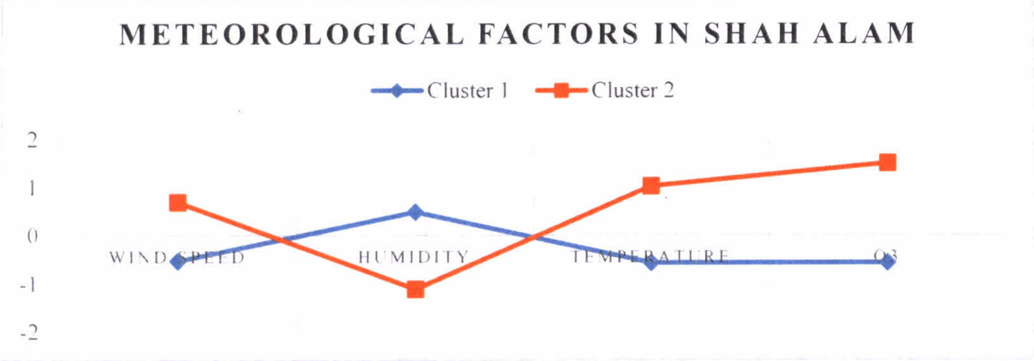
Table 4.19
Final Cluster centres

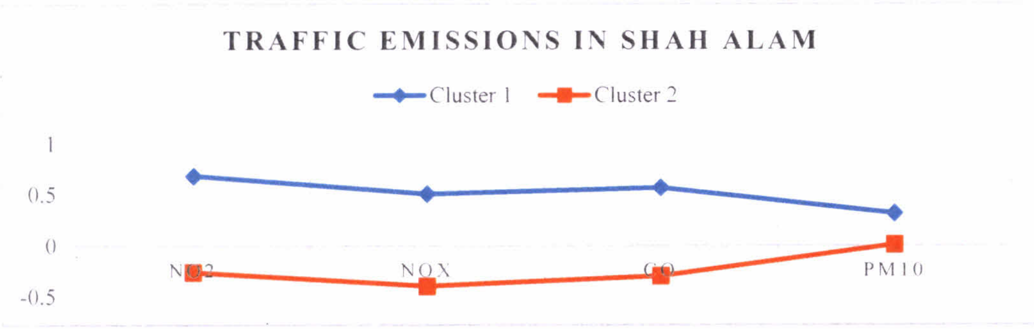
Location	Factor	Variables	Cluster 1	Cluster 2
Jerantut	Meteorological parameters	Wind Speed	0.433	0.231
		Humidity	-1.003	0.068
		Temperature	1.114	-0.138
		O_3	-0.028	-0.040
	Power Plant	NO_2	-0.953	-0.951
		NO_x	-0.810	-0.799
	Traffic Emissions	SO_2	-0.395	-0.290
		CO	-0.770	-0.775
		PM_{10}	19.604	-0.514
Shah Alam	Meteorological parameters	Wind Speed	-0.533	0.694
		Humidity	0.489	-1.115
		Temperature	-0.570	1.028
		O_3	-0.571	1.502
	Traffic Emissions	NO_2	0.685	-0.259
		NO_x	0.507	-0.394
		CO	0.567	-0.294
		PM_{10}	0.316	0.012
Petaling Jaya	Meteorological parameters	Wind Speed	0.130	-0.515
		Humidity	-1.032	0.422
		Temperature	1.043	-0.383
		O_3	0.571	-0.736
	Traffic Emissions	NO_2	0.304	1.243
		NO_x	0.068	1.490
		SO_2	0.143	0.390
		CO	-0.059	1.370
		PM_{10}	0.060	0.299
Banting	Meteorological parameters	Wind Speed	-0.406	0.962
		Humidity	0.817	-0.824
		Temperature	-0.668	0.912
		O_3	-0.562	1.092
	Traffic Emission	CO	-0.076	-0.563
		PM_{10}	0.241	-0.115
		NO_2	-0.032	-0.603
		NO_x	-0.101	-0.587

To further simplify the result of final cluster centres, interpretation along with the representation are shown in Figure 4.9

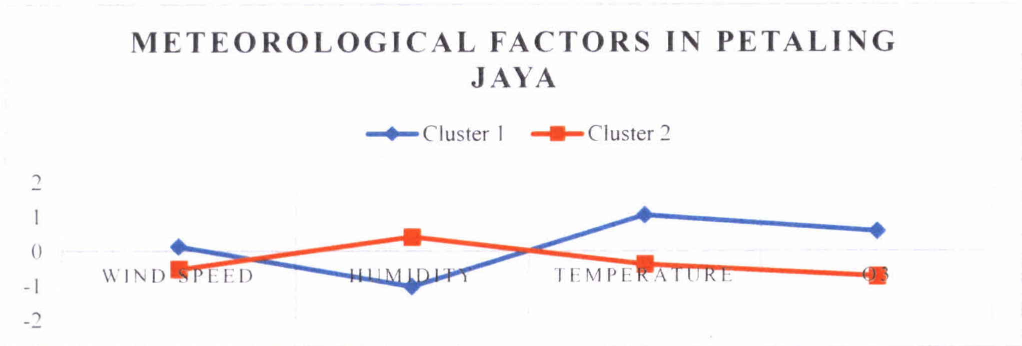


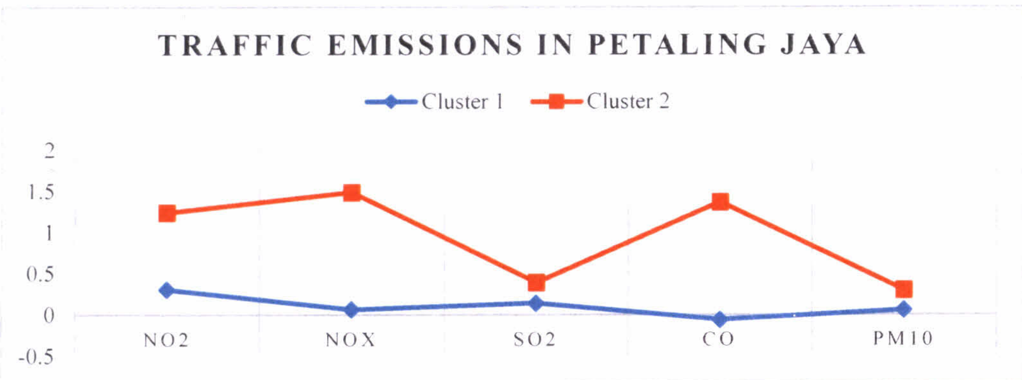
Jerantut





Shah Alam





Petaling Jaya

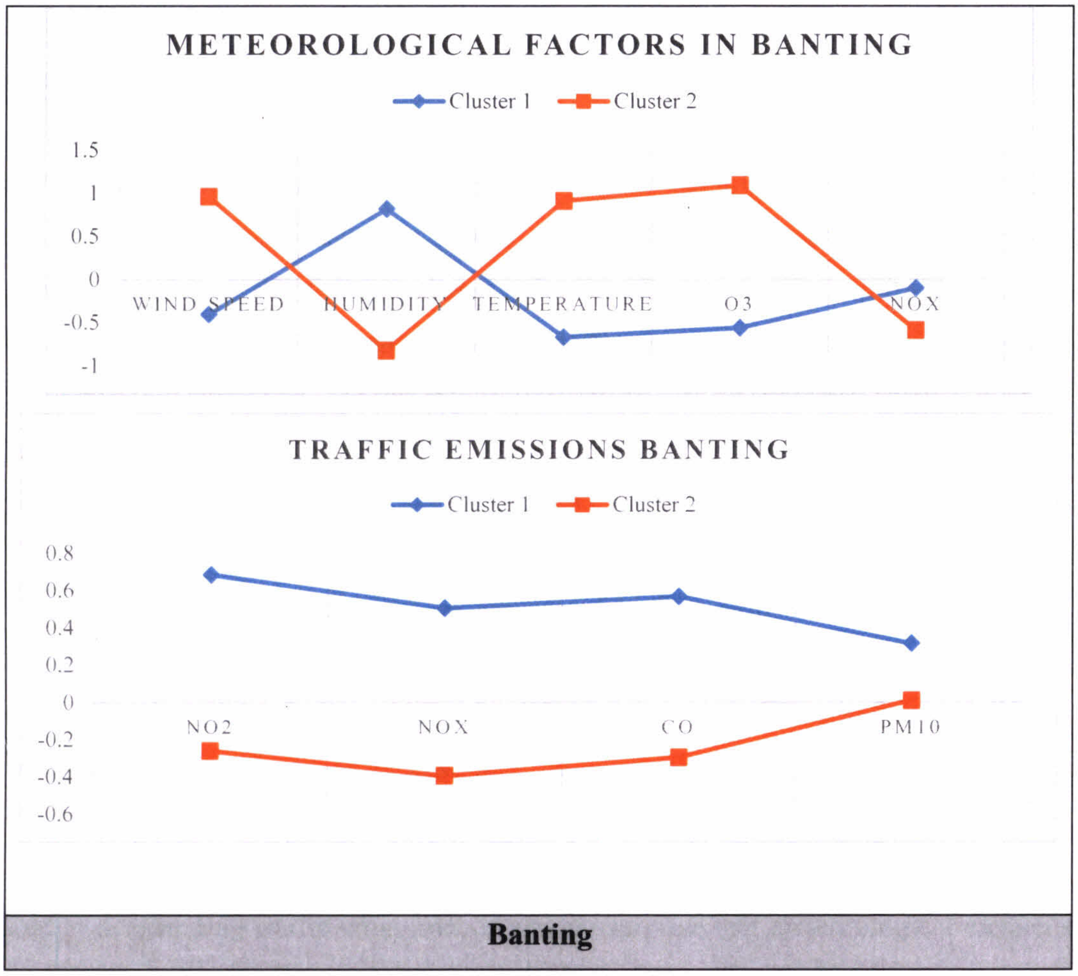


Figure 4.9 Final Cluster Centres in Graphical Form

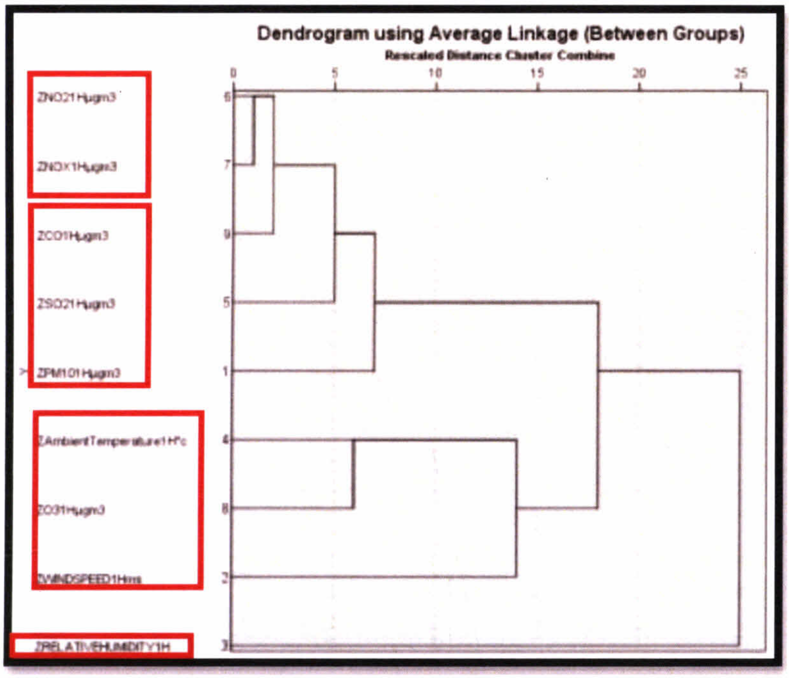
According to Figure 4.9, the researcher will only be concerned with the interpretation of traffic emissions. Jerantut has two clusters where clusters 1 is slightly higher in traffic emissions than cluster 2. When the two clusters were compared for each factor, additional information about traffic emissions could be highlighted. Temperature and O_3 influence the rise in traffic emission concentrations because cluster 1 of meteorological factors demonstrates that temperature and O_3 are much higher to cluster 2. Furthermore, air particles from power plants have the same effect with temperature and O_3. This also be demonstrated by spearman analysis which shows a positive link between CO, temperature, O_3, NO_2 and NOx.

Apart from that, Shah Alam has two clusters. Cluster 1 demonstrates that traffic emissions are significantly greater than in Cluster 2. It is greater in all particles associated with traffic emissions including NO_2, NOx, CO and PM_{10}. When all components were compared, it was discovered that cluster 1 only dominated in humidity if compared to other elements with cluster 2. This suggests that when humidity levels rise, so will the concentration of transportation emissions. High humidity accelerates the release of hazardous or dangerous compounds into the atmosphere (Korhale et al., (2022). It also breeds dust mites in the house which could lower the air quality. Petaling Jaya achieves the same results as Shah Alam. The sole difference between Shah Alam and Petaling Jaya is that SO_2 is included in Petaling Jaya when comes to traffic emission but not in Shah Alam. Humidity has a significant impact on the concentration of transportation emissions. The interpretation is valid based on spearman analysis because both Petaling Jaya and Shah Alam exhibit a positive relationship between CO and Humidity but a negative relationship between CO and other meteorological parameters.

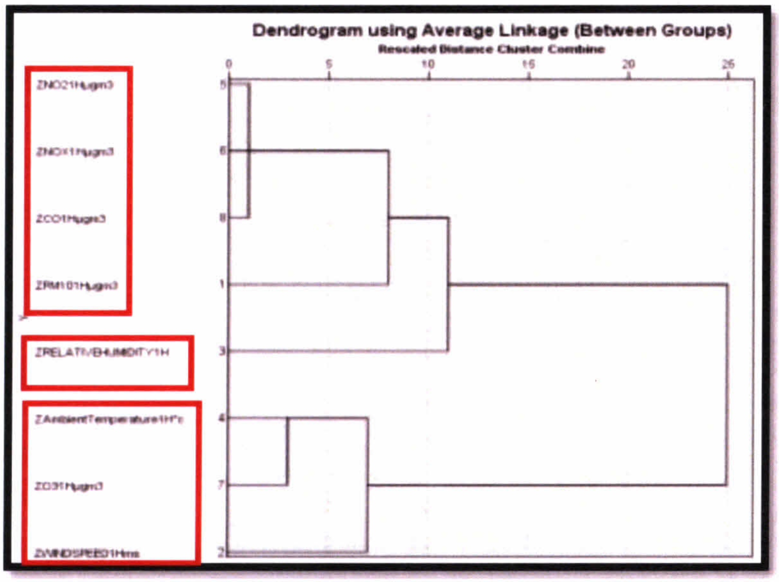
Finally, Banting comprises two clusters. Cluster 1 emits somewhat more traffic emission than Cluster 2. Humidity, like Shah Alam and Petaling Jaya, plays an essential role in increasing the concentration of traffic pollutants. Based on spearman analysis, the interpretation is justified because Banting indicates a positive association between CO and humidity but a negative relationship between CO and other meteorological parameters. In conclusion, it was clear that meteorological elements play a substantial role in determining traffic emissions. It was no surprise that meteorological elements were the most important determinants in each place. Humidity has a positive impact on traffic emissions in Banting, Shah Alam, and Petaling Jaya whereas temperature and O_3 have a positive impact on traffic emissions in Jerantut.

4.7.3 Dendrogram (Hierarchical Clustering)

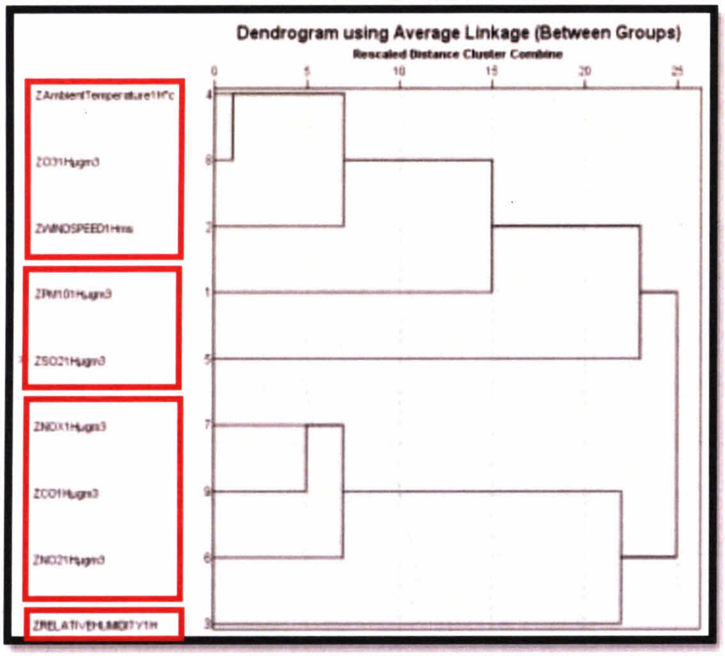
Figure 4.10 shows the dendrogram of air particles for each location. The result obtain will be compared to objective 1 in comparing the details for each air particles in each location to see the difference between both results in the context of traffic emissions. The purpose of this comparison is not to identify which method is better in grouping the component of traffic emission since both of it does not has any performance indicating stating which method is better. The comparison act as a reference which could enhance for the further future in other studies.



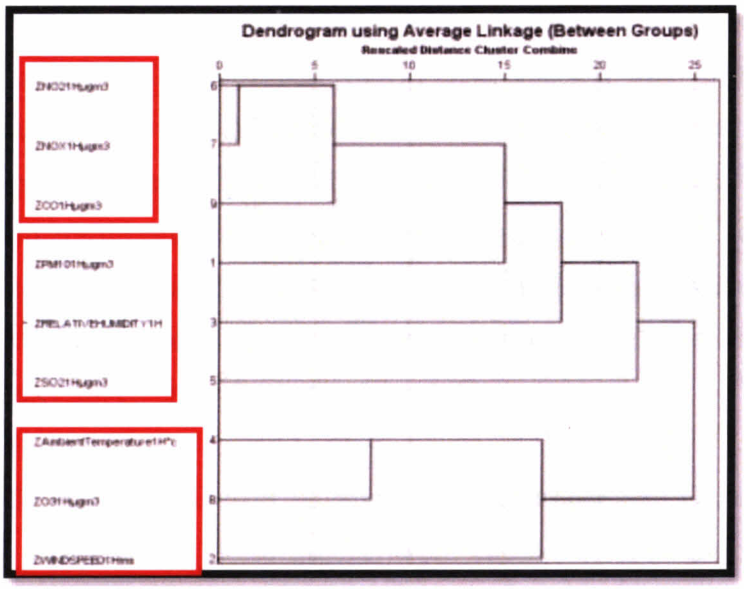
Jerantut



Shah Alam



Petaling Jaya



Banting

Figure 4.10 Dendrogram for each location

Figure 4.10 shows that NO_2 and NO_x were grouped together in Jerantut. Under factor analysis, NO_2 and NO_x were produced as a result of power plant air influence. Temperature, O_3, and wind speed were also grouped together. This is most likely due to a meteorological factor as opposed to a factor analysis. The main difference between factor analysis and cluster analysis results is that CO, PM_{10}, SO_2, and humidity were grouped together. If the prior analysis were followed, humidity should have been included in the meteorological factors. However, CO, PM_{10}, and SO_2 were emissions from transportation in the particular area.

Shah Alam has three clusters. The first group includes temperature, wind speed, and O_3 elements. These items were identified as meteorological factors by factor analysis results. NO_2,NO_x,CO, and PM_{10} were found in Cluster 2. According to factor analysis, this cluster is most likely caused by traffic emissions. The main difference from the comparison of both cluster and factor analysis is that humidity is acted upon by Cluster 3. Humidity should be included in meteorological factors. However, the components of traffic emissions such as NO_2, NOx, CO, and PM_{10} remain the same.

In Petaling Jaya, wind speed, temperature, and O_3 were all grouped together. As noted in the factor analysis, this is most likely related to meteorological factors. Furthermore, PM_{10} and SO_2 were grouped together even though both of these particles were also grouped with CO, NO_2, and NOx in factor analysis. This is most likely due to the separation of industrial and traffic emissions. For cluster analysis, PM_{10} and SO_2 were most likely generated by industrial emissions while NOx, CO and NO_2 were most likely caused by traffic emissions. Thus, cluster analysis identifies traffic emissions as NO_x, CO, and NO_2 whereas factor analysis identifies traffic emissions as PM_{10}, SO_2, NO_x, CO, and NO_2. Humidity acts as its own cluster even though humidity was one of the parts of meteorological factors in factor analysis.

Finally, Banting is divided into three clusters. Temperature, O_3, and wind speed comprise the first cluster. These aspects were caused by meteorological factors, according to factor analysis. Aside from that, transportation emissions were analysed using cluster analysis and factor analysis. Traffic emissions in cluster analysis include CO, NO_2, and NOx but traffic emissions in factor analysis include NO_2,NOx, PM_{10} and CO. Banting's final cluster is most likely influenced by industrial emissions from agriculture which include PM_{10}, humidity, and SO_2. To summarize the result, Figure 4.11 shows a better representation between both analysis in the context of traffic emission.

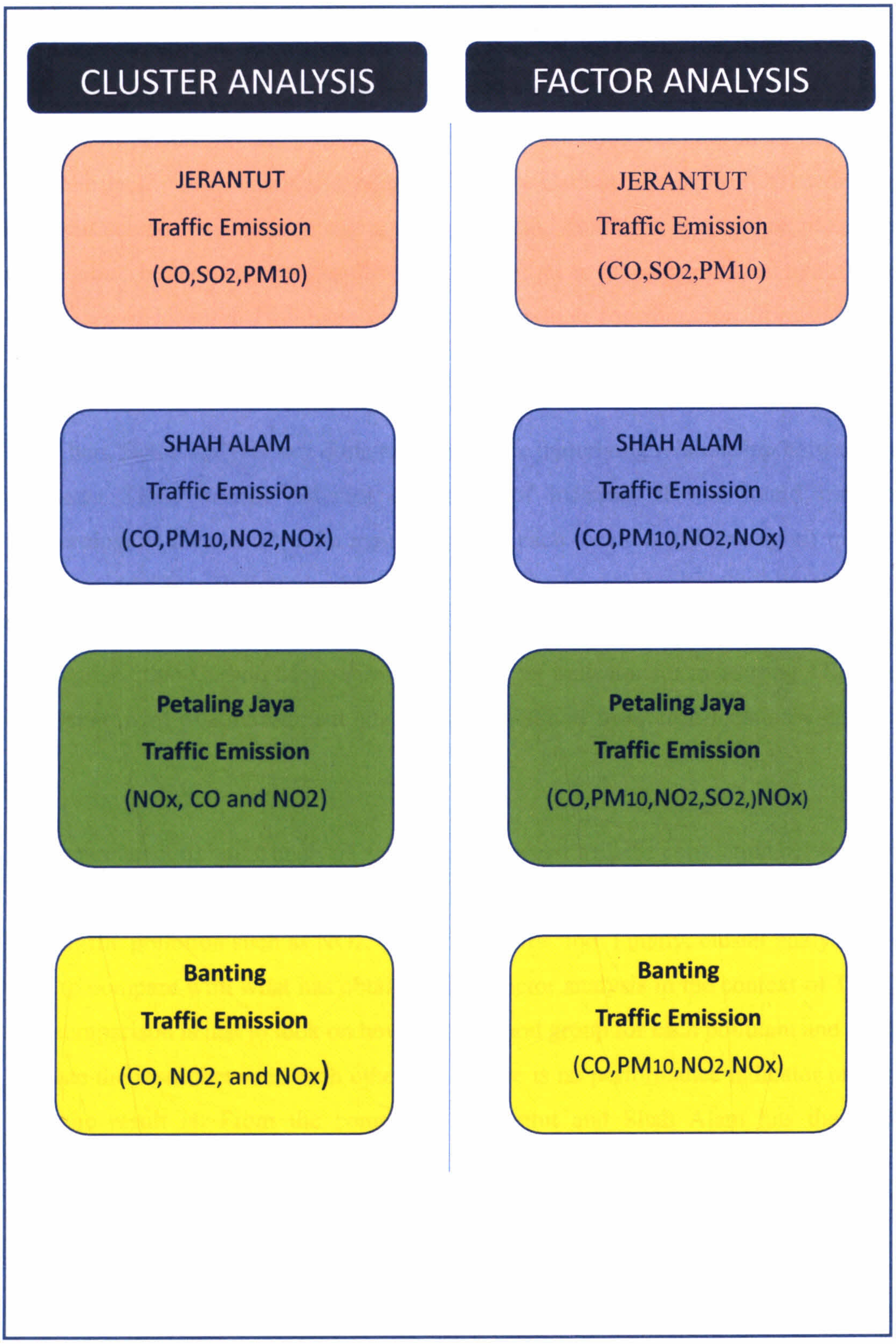


Figure 4.11 Comparison between factor and cluster analysis

4.8 Concluding Remark

This chapter discusses about the result obtain from chapter 4. After data pre-processing has been done, descriptive analysis was done on the overall dataset and dependent variable. By using visualization tools, information was gain on the dependent variable itself. For example, it helps to identify Carbon Monoxide (CO) works on different categories such as location, time period and months. Not only that, researcher could gain a better insight on the different value of mean for each pollutant by different categories of location. Pollutants from traffic has high concentration from industrial area compared to rural area.

Then, factor analysis was done to examine the underlying relationship between the dominant TRAP within different categories of location. It was found out that meteorological factors were the main factor for each location influencing air quality. However, traffic pollution also plays a vital part in air quality mainly industrial area which has a high value of total variance compared to other location. Furthermore, it is no surprised that Carbon Monoxide were put as an indicator for measuring TRAP by the Department of Environment since TRAP pollution from each location consist of CO.

Spearman correlation analysis was used to proceed with the correlation between CO and the factors obtained by each location previously. CO has high positive correlation with traffic pollution such as NO_2, NOx , and PM_{10} too. Finally, cluster analysis was used to compare with what has obtained from factor analysis in the context of TRAP. The comparison is just to look on how each method group for each pollutant and not to validate the result between each other since there is no performance indicator on how well the result is. From the comparison, Jerantut and Shah Alam has the same components of traffic emission compared between both method meanwhile Petaling Jaya and Banting differs from on e another.

CHAPTER FIVE
CONCLUSION AND RECOMMENDATION

5.0 Chapter Overview

This chapter concludes the overall of the study in accessing potential dominant factors of traffic related air pollution (TRAP) by using both cluster and factor analysis Furthermore, a few suggestions also will be discussed on the topic or issue for future purposes. The future researchers might gain insight and make further improvements based on this study.

5.1 Conclusion

Air quality is always an issue since vehicles nowadays have become a must have thing for people in going from one place to another place. TRAP will soon become a major concern globally if there is no policies or strategy to be made soon. However, before the strategies were being made, it was advised to learn on how TRAP pollution works in different location so that we could have a better insight in making the decisions. There is no denying that air quality has been focusses a lot among researchers, but this thesis focusses on specifically on the TRAP itself. Therefore, multivariate techniques were used such as factor and cluster analysis on the particular topic to obtain an insight what is currently happening on TRAP by different location of categories.

Differences in categories of location brings different outcome towards air pollution. Since the study focusses on transportation related are pollution, the researcher could identify the differences of TRAP pollution between each category of location. Petaling Jaya which is an industrial area has the highest impact when comes to traffic emission meanwhile Jerantut has the lowest impact respectively. Traffic emission in each category of location behave differently among another. Even the components inside the traffic emission differs among another. Azid et al. (2015) also stated that highest traffic emission were mainly from industrial area since those area could bring working opportunity towards the society hence it will indirectly increase traffic emission.

Aside from that, determine the strength of association between CO and other pollutants for each components within different categories of location also should have been taken into consideration since CO is the main variable for traffic emission. It was identified that there is a significant association between CO and other pollutants for each factor within different location groups. However, it was found out that CO were mainly highly correlate with another pollutant which comes from traffic emission too. For example, CO in Shah Alam were highly correlated with NO_2 due to the facts that it was probably due to motorcycles. Even study by Kamaruzzaman et al. (2017) and Sofwan et al. (2021) shows that CO were highly correlated with their respective factor that relates with traffic emission such as NO_2 and NO_x.

Finally, cluster analysis was used to compare the groups of air pollutants within each location along with what was produced by factor analysis. The method here is just purely for comparison since it has no performance indicator stated which one is better. However, researcher could take it as a future reference to improve the study for the further future.

5.2 Recommendation

Similar to other research studies, this study also recommends further analysis. The same process of comparing traffic emission can be implemented in any other areas besides than those location since not all stations for each category areas are considered in the study due to the availability in finishing the study on time. This is because that each location having its own characteristic such as the size of city, density of population and other more. Since not all pollutants were available in the study, extra pollutant such as carbon Monoxide, methane and organic gases could be added for the future study in looking at the bigger picture in accessing traffic emission. Besides that, regarding to the topic in accessing traffic emission, researcher might be interested in accessing traffic emission by using different method and software. Despite of using SPSS, other software can also be explored by use MATLAB, Stata or SAS in analysing traffic emission. Modern technique such as machine learning and other methods also could be applied in this topic too.

REFERENCES

Abdullah, S., Ismail, M., Fong, S. Y., & Ahmed, N. (2016). Evaluation for Long Term PM 10 Concentration Forecasting using Multi Linear Regression (MLR) and Principal Component Regression (PCR) Models. *EnvironmentAsia*, *9*(2).

Afroz, R., Hassan, M. N., & Ibrahim, N. A. (2003). Review of air pollution and health impacts in Malaysia. *Environmental research*, *92*(2), 71-77.

Ahamad, F., Latif, M. T., Dominick, D., & Tang, R. (2012). Daily maximum ozone variability at selected urban and semi-urban areas in Peninsular Malaysia. In *UMT 11th International Annual Symposium on Sustainability Science and Management. Terengganu, Malaysia*.

Ahmat, H., Musa, N. S., Nazamid, N., & Zaharin, N. A. (2019). Classification of high and low level of PM10 concentrations in Klang and Shah Alam, Malaysia. *Malaysian Journal of Computing*, *4*(2), 325-334.

Ang, L. W., Koh, B. K., Chan, K. P., Chua, L. T., James, L., & Goh, K. T. (2009). Epidemiology and control of hand, foot and mouth disease in Singapore. *Ann Acad Med Singapore*, *38*(2), 106-12.

Asmarini, M. (2015). Factors contributing to air pollution levels using factor analysis *Doctoral dissertation, Universiti Malaysia Perlis (UniMAP).*

Awang, N. R., Elbayoumi, M., Ramli, N. A., & Yahaya, A. S. (2016). The influence of spatial variability of critical conversion point (CCP) in production of ground level ozone in the context of tropical climate. *Aerosol and Air Quality Research*, *16*(1), 153-165.

Azmi, S. Z., Latif, M. T., Ismail, A. S., Juneng, L., & Jemain, A. A. (2010). Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. *Air Quality, Atmosphere & Health*, *3*(1), 53-64.

Azid, A., Juahir, H., Toriman, M. E., Endut, A., Kamarudin, M. K. A., Rahman, M. N. A., ... & Yunus, K. (2015). Source apportionment of air pollution: a case study in Malaysia. *Jurnal Teknologi*, *72*(1).

Barmpadimos, I., Hueglin, C., Keller, J., Henne, S., & Prévôt, A. S. H. (2011). Influence of meteorology on PM 10 trends and variability in Switzerland from 1991 to 2008. *Atmospheric Chemistry and Physics*, *11*(4), 1813-1835.

Boschi, N. (1999). Defining an educational framework for indoor air sciences education. In *Education and Training in Indoor Air Sciences* (pp. 3-6). Springer, Dordrecht.

Browne, M. W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate behavioral research*, *36*(1), 111-150.

Carslaw, D. C., & Beevers, S. D. (2013). Characterising and understanding emission sources using bivariate polar plots and k-means clustering. *Environmental modelling & software*, *40*, 325-329.

Chatfield, C., & Collins, A. J. (2018). *Introduction to multivariate analysis*. Routledge.

Chen, P., Bi, X., Zhang, J., Wu, J., & Feng, Y. (2015). Assessment of heavy metal pollution characteristics and human health risk of exposure to ambient PM2. 5 in Tianjin, China. *Particuology*, *20*, 104-109.

Chuang, H. C., Shie, R. H., Chio, C. P., Yuan, T. H., Lee, J. H., & Chan, C. C. (2018). Cluster analysis of fine particulate matter (PM2.5) emissions and its bioreactivity in the vicinity of a petrochemical complex. *Environmental Pollution*, *236*, 591-597.

Costello, A. B., & Osborne, J. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical assessment, research, and evaluation*, *10*(1), 7.

Cox, L. (1999). *Nitrogen oxides (NOx) why and how they are controlled*. Diane Publishing.

Cuhadaroglu, B. U. R. H. A. N., & Demirci, E. M. R. U. L. L. A. H. (1997). Influence of some meteorological factors on air pollution in Trabzon city. *Energy and buildings*, *25*(3), 179-184.

Dahlin, S., Lantto, C., Englund, J., Westerberg, B., Regali, F., Skoglundh, M., & Pettersson, L. J. (2019). Chemical aging of Cu-SSZ-13 SCR catalysts for heavy-duty vehicles–Influence of sulfur dioxide. *Catalysis Today*, *320*, 72-83.

Das, R., Khezri, B., Srivastava, B., Datta, S., Sikdar, P. K., Webster, R. D., & Wang, X. (2015). Trace element composition of PM2. 5 and PM10 from Kolkata–a heavily polluted Indian metropolis. *Atmospheric Pollution Research*, *6*(5), 742-750.

Ding, Y., & Li, F. (2017). Examining the effects of urbanization and industrialization on carbon dioxide emission: evidence from China's provincial regions. *Energy*, *125*, 533-542.

Dominick, D., Juahir, H., Latif, M. T., Zain, S. M., & Aris, A. Z. (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric environment*, *60*, 172-181.

Farda, M., & Balijepalli, C. (2018). Exploring the effectiveness of demand management policy in reducing traffic congestion and environmental pollution: Car-free day and odd-even plate measures for Bandung city in Indonesia. *Case Studies on Transport Policy*, *6*(4), 577-590.

Fang, C., & Ren, Y. (2017). Analysis of emergy-based metabolic efficiency and environmental pressure on the local coupling and telecoupling between urbanization and the eco-environment in the Beijing-Tianjin-Hebei urban agglomeration. *Science China Earth Sciences*, *60*(6), 1083-1097.

Fang, Y., Cheng, H., Li, X., Xu, Y., Xu, H., Chen, Z., ... & Cao, J. (2021). Short-term exposure to ambient air pollution and atrial fibrillation hospitalization: A time-series study in Yancheng, China. *Ecotoxicology and Environmental Safety*, *228*, 112961.

Fruchter, B. (1954). Introduction to factor analysis.

Field, A. (2000). *Discovering statistics using IBM SPSS statistics*. sage.

Geddes, J. A., Murphy, J. G., & Wang, D. K. (2009). Long term changes in nitrogen oxides and volatile organic compounds in Toronto and the challenges facing local ozone control. *Atmospheric Environment*, *43*(21), 3407-3415.

Gwilliam, K. M., Kojima, M., & Johnson, T. (2004). *Reducing air pollution from urban transport*. Washington, DC: World Bank.

Guo, J. X., Zeng, Y., Zhu, K., & Tan, X. (2021). Vehicle mix evaluation in Beijing's passenger-car sector: From air pollution control perspective. *Science of the Total Environment*, *785*, 147264.

Hadeed, S. J., O'Rourke, M. K., Burgess, J. L., Harris, R. B., & Canales, R. A. (2020). Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of The Total Environment*, *730*, 139140.

Halim, N. D. A., Latif, M. T., Ahamad, F., Dominick, D., Chung, J. X., Juneng, L., & Khan, M. F. (2018). The long-term assessment of air quality on an island in Malaysia. *Heliyon*, *4*(12), e01054.

Halim, N. D. A., Latif, M. T., Mohamed, A. F., Maulud, K. N. A., Idrus, S., Azhari, A., ... & Sofwan, N. M. (2020). Spatial assessment of land use impact on air quality in mega urban regions, Malaysia. *Sustainable Cities and Society*, *63*, 102436.

Hsu, C. H., & Cheng, F. Y. (2019). Synoptic weather patterns and associated air pollution in Taiwan. *Aerosol and Air Quality Research*, *19*(5), 1139-1151.

Hua, A. K. (2018). Applied chemometric approach in identification sources of air quality pattern in Selangor, Malaysia. *Sains Malaysiana*, *47*(3), 471-479.

Isiyaka, H. A., & Azid, A. (2015). Air quality pattern assessment in Malaysia using multivariate techniques. *Malaysian Journal of Analytical Sciences*, *19*(5), 966-978.

Isiyaka, H., Juahir, H., Toriman, M. E., Gasim, B. M., Azid, A., Amri, M. K., ... & Garba, M. A. (2014). Spatial assessment of air pollution index using environ metric modeling techniques. *Advances in Environmental Biology*, *8*(24), 244-256.

Jamil, M. S., Japeri, A. Z. U. S. M., Bakar, A. A. A., Ali, K. A. M., & Ahmat, H. (2019). Identification of source contributions to air pollution in penang using factor analysis. *International Journal of Integrated Engineering*, *11*(8), 221-228.

Jones, T. A. (1969). Skewness and kurtosis as criteria of normality in observed frequency distributions. *Journal of Sedimentary Research, 39*(4), 1622-1627.

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

Kamaruzzaman, A., Saudi, A. S. M., Azid, A., Balakrishnan, A., Abu, I. F., Amin, N. A., & Rizman, Z. I. (2017). Assessment on air quality pattern: A case study in Putrajaya, Malaysia. *Journal of Fundamental and Applied Sciences*, *9*(4S), 789-800.

Kappos, A. D., Bruckmann, P., Eikmann, T., Englert, N., Heinrich, U., Höppe, P., ... & Wichmann, H. E. (2004). Health effects of particles in ambient air. *International journal of hygiene and environmental health*, *207*(4), 399-407.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis. *John Wiley & Sons.*

Keresztes, R., & Rapo, E. (2017). Statistical analysis of air pollution with specific regard to factor analysis in the Ciuc basin, Romania. *Studia Universitatis Babes-Bolyai, Chemia*, *62*(3).

Kettenring, J. R. (2006). The practice of cluster analysis. *Journal of classification, 23*(1), 3-30.

Kline, P. (2014). *An easy guide to factor analysis*. Routledge.

Korhale, N., Anand, V., Panicker, A., & Beig, G. (2022). Measurements of surface ozone and its precursors in different microenvironments of coastal Indian metropolis of Mumbai. *International Journal of Environmental Science and Technology*, 1-18.

Lam, J. C., Wan, K. K. W., Wong, S. L., & Lam, T. N. T. (2012). Principal component analysis and longterm building energy simulation correlation. *Energy Conversion and Management, 51*(1), 135–139.

Landrigan, P. J. (2017). Air pollution and health. *The Lancet Public Health*, *2*(1), e4-e5.

Latif, M. T., Dominick, D., Ahamad, F., Khan, M. F., Juneng, L., Hamzah, F. M., & Nadzir, M. S. M. (2014). Long term assessment of air quality from a background station on the Malaysian Peninsula. *Science of the total environment*, *482*, 336-348.

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of experimental social psychology*, *74*, 150-156.

Liu, Y., Wu, J., Yu, D., & Ma, Q. (2018). The relationship between urban form and air pollution depends on seasonality and city size. *Environmental Science and Pollution Research, 25*(16), 15554-15567.

Liu, X., Hadiatullah, H., Tai, P., Xu, Y., Zhang, X., Schnelle-Kreis, J., ... & Zimmermann, R. (2021). Air pollution in Germany: Spatio-temporal variations and their driving factors based on continuous data from 2008 to 2018. *Environmental Pollution*, *276*, 116732.

Li, C., Wang, Z., Li, B., Peng, Z. R., & Fu, Q. (2019). Investigating the relationship between air pollution variation and urban form. *Building and Environment*, 147, 559-568

Li, M., Wang, W. L., Wang, Z. Y., & Xue, Y. (2018). Prediction of PM2.5 concentration based on the similarity in air quality monitoring network. *Building and Environment*, 137, 11-17.

Li, X., Hussain, S. A., Sobri, S., & Said, M. S. M. (2021). Overviewing the air quality models on air pollution in Sichuan Basin, China. *Chemosphere*, *271*, 129502.

Liu, T. H., & Yoon, Y. (2019). Effects of meteorological factors to reduce large-scale PM10 emission estimation errors on unpaved roads. *Atmospheric Environment,* 217, 116956.

Makmom Abdullah, A., Armi Abu Samah, M., & Yee Jun, T. (2012). An overview of the air pollution trend in Klang Valley, Malaysia. *Open Environmental Sciences*, 6(1).

Metia, S., Ha, Q. P., Duc, H. N., & Azzi, M. (2018). Estimation of power plant emissions with unscented Kalman filter. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11*(8), 2763-2772.

Mukaka, M. M. (2012). Statistics corner: a guide to appropriate use of correlation in medical research. *Malawi Medical Journal, 24*(3), 69-71.

Mohamad, N. D., Ash'aari, Z. H., & Othman, M. (2015). Preliminary assessment of air pollutant sources identification at selected monitoring stations in Klang Valley, Malaysia. *Procedia Environmental Sciences*, 30, 121-126.

Mohamad, N. S., Deni, S. M., & Ul-Saufie, A. Z. (2018). Application of the First Order of Markov Chain Model in Describing the PM10 Occurrences in Shah Alam and Jerantut, Malaysia. *Pertanika Journal of Science & Technology*, 26(1).

Nayebere, S. R., Aburizaiza, O. S., Siddique, A., Carpenter, D. O., Hussain, M. M., Zeb, J., ... & Khwaja, H. A. (2018). Ambient air quality in the holy city of Makkah: A source apportionment with elemental enrichment factors (EFs) and factor analysis (PMF). *Environmental Pollution*, 243, 1791-1801

Oh, W., & Kato, S. (2018). The effect of airspeed and wind direction on human's thermal conditions and air distribution around the body. *Building and Environment*, 141, 103-116.

Othman, M., & Latif, M. T. (2021). Air pollution impacts from COVID-19 pandemic control strategies in Malaysia. *Journal of cleaner production*, 291, 125992.

Parveen, N., Siddiqui, L., Sarif, M. N., Islam, M. S., Khanam, N., & Mohibul, S. (2021). Industries in Delhi: Air pollution versus respiratory morbidities. *Process Safety and Environmental Protection, 152*, 495-512.

Paul, R. K. (2010). Multicollinearity: Causes, effects and remedies. *IASRI, New Delhi*, *1*(1), 58-65.

Payus, C., Abdullah, N., & Sulaiman, N. (2013). Airborne particulate matter and meteorological interactions during the haze period in Malaysia. International *Journal of Environmental Science and Development, 4*(4), 398.

Payus, C. M., Thevan, A. V., & Sentian, J. (2019). Impact of school traffic on outdoor carbon monoxide levels. *City and Environment Interactions*, 4, 100032.

Rahman, S. R. A., Ismail, S. N. S., Raml, M. F., Latif, M. T., Abidin, E. Z., & Praveena, S. M. (2015). The assessment of ambient air pollution trend in Klang Valley, Malaysia. *World Environment, 5*(1), 1-11.

Rahman, S. A., Ismail, S. N. S., Sahani, M., Ramli, M., & Latif, M. (2017). A case crossover analysis of primary air pollutants association on acute respiratory infection (ARI) among children in urban region of Klang valley, Malaysia. *Annals of Tropical Medicine and Public Health, 10*(1).

Rummel, R. J. (1967). Understanding factor analysis. *Journal of conflict resolution*, *11*(4), 444-480.

Rossi, R., Ceccato, R., & Gastaldi, M. (2020). Effect of Road Traffic on Air Pollution. Experimental Evidence from COVID-19 *Lockdown. Sustainability, 12*(21), 8984.

Sahrir, S., Abdullah, A. M., Ponrahono, Z., & Sharaai, A. H. (2019). Environmetric study on air quality pattern for assessment in Klang Valley, Malaysia. *International Journal of Recent Technology and Engineering*, *8*, 17-24.

Schauer, J. J., Kleeman, M. J., Cass, G. R., & Simoneit, B. R. (2001). Measurement of emissions from air pollution sources. 3. C1− C29 organic compounds from fireplace combustion of wood. *Environmental science & technology, 35*(9), 1716-1728.

Schauer, R. (Ed.). (2012). *Sialic acids: chemistry, metabolism, and function* (Vol. 10). Springer Science & Business Media.

Shafii, N. Z., Saudi, A. S. M., Mahmud, M., & Rizman, Z. I. (2017). Spatial assessment on ambient air quality status: A case study in Klang, Selangor. *Journal of Fundamental and Applied Sciences*, *9*(4S), 964-977.

Shafie, S. H. M., & Mahmud, M. (2020). Urban air pollutant from motor vehicle emissions in Kuala Lumpur, Malaysia. *Aerosol and Air Quality Research*, 20.

Shaharuddin, M., Zaharim, A., Nor, M. J. M., Karim, O. A., & Sopian, K. A. M. A. R. U. Z. Z. A. M. A. N. (2008). Application of wavelet transform on airborne suspended particulate matter and meteorological temporal variations. WSEAS *Transactions on Environment and Development, 4*(2), 89-98

Shekarrizfard, M., Faghih-Imani, A., Crouse, D. L., Goldberg, M., Ross, N., Eluru, N., & Hatzopoulou, M. (2016). Individual exposure to traffic related air pollution across land-use clusters. *Transportation Research Part D: Transport and Environment*, *46*, 339-350.

Shuhaili, A., Ihsan, S. I., & Faris, W. F. (2013). Air pollution study of vehicles emission in high volume traffic: Selangor, Malaysia as a case study. *WSEAS Transactions on Systems*, *12*(2), 67-84.

Sofwan, N. M., Mahiyuddin, W. R. W., Latif, M. T., Ayub, N. A., Yatim, A. N. M., Mohtar, A. A. A., ... & Sahani, M. (2021). Risks of exposure to ambient air pollutants on the admission of respiratory and cardiovascular diseases in Kuala Lumpur. *Sustainable Cities and Society*, *75*, 103390.

Tranmer, M., & Elliot, M. (2008). Multiple linear regression. *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, *5*(5), 1-5.

Verma & Desai, B., (2008), Effect of Meteorological Conditions on Air Pollutions of Surat City, *Journal of International Environmental Application Science,* 3, 358-367.

Wang, Y., Yao, L., Xu, Y., Sun, S., & Li, T. (2021). Potential heterogeneity in the relationship between urbanization and air pollution, from the perspective of urban agglomeration. *Journal of Cleaner Production*, 298, 126822.

Wu, C., Hu, W., Zhou, M., Li, S., & Jia, Y. (2019). Data-driven regionalization for analyzing the spatiotemporal characteristics of air quality in China. *Atmospheric Environment*, 203, 172-182.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, *9*(2), 79-94.

Zaini, N. A., Ean, L. W., & Ahmed, A. N. (2021, November). Forecasting of Carbon Monoxide Concentration Based on Sequence-to-Sequence Deep Learning Approach. In *International Visual Informatics Conference* (pp. 518-529). Springer, Cham.

Zakaria, U. A., Saudi, A. S. M., Abu, I. F., Azid, A., Balakrishnan, A., Amin, N. A., & Rizman, Z. I. (2017). The assessment of ambient air pollution pattern in Shah Alam, Selangor, Malaysia. *Journal of Fundamental and Applied Sciences,* *9*(4S), 772-788.

Zhang, Y. (2019). Dynamic effect analysis of meteorological conditions on air pollution: A case study from Beijing. *Science of the Total Environment*, *684*, 178-185.

Zhang, J., Zhang, L. Y., Du, M., Zhang, W., Huang, X., Zhang, Y. Q., ... & Xiao, H. (2016). Identifying the major air pollutants base on factor and cluster analysis, a case study in 74 Chinese cities. *Atmospheric Environmen*t, 144, 37-46

Zhao, Y., Nguyen, N. T., Presto, A. A., Hennigan, C. J., May, A. A., & Robinson, A. L. (2016). Intermediate volatility organic compound emissions from on-road gasoline vehicles and small off-road gasoline engines. *Environmental science & technology, 50*(8), 4554-4563.

Zhao, Y., Saleh, R., Saliba, G., Presto, A. A., Gordon, T. D., Drozd, G. T., ... & Robinson, A. L. (2017). Reducing secondary organic aerosol formation from gasoline vehicle exhaust. *Proceedings of the National Academy of Sciences, 114*(27), 6984-6989.

APPENDIX

1. Summarize on Factor Analysis on air pollution (Main Theme)

No	Author	Study Location	Objective	Findings
1	Jamil et al.(2019)	Penang	To determine their characteristics and contributions to air pollution in Penang	Factors – 3 Factors are identified : 1.(Meteorological factors) 2.(Traffic Emission) . 3 (Industrial activities) Factor components of traffic emission : (CO,NO2 and PM10)
2	Rahman et al (2015)	Klang, Shah Alam, Cheras, Putrajaya, Banting, Kuala Selangor,Petaling Jaya	To explore the trend of ambient air pollution (i.e., PM10, CO, NO2, O3) within the eight selected Malaysian air monitoring stations in Klang valley	Factors – 3 Factors are identified:1.(Transportation) 2.(Meteorological parameters) . 3 (The influence of 03) Factor components of traffic emission : (CO and NO2)
3	Azid et al (2015)	Pasir Gudang, Kuching, Bukit Rambai, Nilai,Klang, Balok Baru, Pengkalan Chepa,Paka, Labuan	Identification of potential sources of variations in air quality around the study area based	Factors – 2 factors were identified : 1.(Gas-pollutants) 2.(Non-Gas Pollutant) Factor's component – Factor 1(CH4, NmHC, and THC) Factor 2 (NOx, CO, and VOCs)
4	Mohamad et al (2015)	Port Klang, Petaling Jaya,Shah Alam, Kuala Lumpur	Aims to carry out a preliminary assessment for pollutant sources identification	Factors – 2 factors were identified : 1.(Traffic Emission) 2.(Meteorological Parameters) Factor components of traffic emission : (CO,NO2 and PM10)
5	Asmarini (2015)	Perai, Shah Alam, Seberang Jaya, Jerantut	Aims of this study is to determine the major factors and causes of air pollution	Factors – 3 factors were identified : 1.(Organic Pollution Factor) 2.(Meteorological Factor) 3. (Fuel Factor) Factor components of traffic emission : (CO, NO2, SO2, and PM10)
6	Isiyaka and Azid (2015)	Pasir Gudang ,Kemaman ,Perai,Sibu,Sarawak	Investigate the spatial characteristics in the pattern of air quality monitoring sites	Factors – 2 factors were identified : 1.(Heavy Industrial chemical) 2.(Photochemical pollutions)
7	Abdullah et al., (2015)	Kuala Terengganu, Jerantut and Pahang	Study intended to determine the air pollution potential sources at different land use of urban, sub-urban and rural areas	Factors - Factor 1 (Local traffic sources) , Factor 2 (meteorological parameters) Factor's components – Factor 1 (PM10, CO and NO2) , Factor 2 (wind speed and humidity)
8	Keresztes, R., & Rapo, E. (2017).	Ciuc Basin,Romania	Aims to examine the changes of the concentration of air pollutant over an interval of two years using factor analysis	Factors – 2 factors were identified : 1.(Transportation) 2.(03 and meteorological conditions)
9	Zhang et al. (2016)	74 Chinese cities	Investigated the major air pollutants and its spatial and seasonal distribution in 74 Chinese cities.	Factors – 2 factors were identified : 1.(Traffic emission in Spring) 2.(Traffic emission during winter)

2. Cluster analysis on air pollution (main theme)

No	Author	Study Location	Objective	Findings
1	Halim et al.(2018)	Pulau Langkawi	Aims to evaluate the air quality on Langkawi Island	Cluster – 3 cluster were identified Cluster components – Cluster 1 correspond to traffic emission which has high concentration of PM10, CO, NO, NOx, NO2, and SO2
2	Isiyaka and Azid (2015)	Pasir Gudang ,Kemaman ,Perai,Sibu,	Investigate the spatial characteristics in the pattern of air quality monitoring sites	Cluster – 2 cluster were identified ; 1 (Moderately populated area) , 2.(Less Populated area) Cluster component – Cluster 1(associated with commercial and industrial operations as well as severe traffic congestion) Cluster 2 (residential neighbourhood, tourist destination and has less industrial activity)
3	Hua (2018)	Klang, Petaling Jaya, Shah Alam, Kuala Selangor, Putrajaya,Cheras,Kuala Lumpur ,Banting	Aimed to assess the air quality data and identify the pattern of air pollution sources	Cluster – 3 cluster were identified Cluster components – Most of these cluster were affected by traffic and industrial emission.
4	Awang et al .(2016)	Pasir Gudang, Perai,Kemaman,Kuching, Kota Bharu,Klang,Gombak,Jerantut	To determine the major sources of the air pollutants that influence ozone Critical Conversion Point (CCP).	Cluster – 3 Cluster were identified Cluster components – Cluster 1 designated as the primary city centre, and it has a high concentration of carbon monoxide since it is surrounded by residential areas, commercial areas, and other amenities
5	Sahrir et al. (2019)	Nilai, Petaling Jaya, Shah Alam Putraiaya	Determine the significant pollutant parameters contributing to air quality in Klang Valley	Cluster – 2 cluster were identified Cluster components - Both clusters were mainly affected by its location of area, concentration of Carbon Dioxide and PM10 also plays a major factor
6	Wu et al. (2019)	JPCAP region, China	Identified the key JPCAP regions and corresponding pollution control grades.	Cluster – 4 cluster were identified Cluster components - These clusters were influence by the contaminants pollute Chinese air to differing degrees based on each respective city.
7	Shekarrizfard et al (2016)	Montreal,Canada	To test the relationship between emissions and exposures in different Montreal areas.	Cluster – 2 Cluster were identified Cluster components - Found out that people who reside in the suburb's locations emit more NOx but are exposed to lower NO2 concentrations at home from the cluster itself
8	Zhang et al. (2016)	74 Chinese cities	Investigate the principal air contaminants, as well as their geographical and seasonal distribution in 74 Chinese cities	Cluster- 6 clusters were identified Cluster components- Cluster 5 are the most severely polluted in the area by having high concentration of SO2, CO, PM10, and PM2.5
9	Li et al (2019)	Shanghai,China	Cluster analyses are utilized to reveal the similar behaviour in Shanghai	Cluster – 3 Cluster were identified Cluster components- Each cluster differed depending on the pollutant as did the urban form PM2.5 concentrations

3. Spearman Correlation Analysis on air pollution (main theme)

No	Author	Study Location	Objective	Findings
1	Zakaria et al (2017)	Shah Alam	Establish the association between air pollution trends with the industrial activities in Shah Alam, Selangor	CO – Significant with SO2, NO2, O3, PM10 and Humidity CO – Highly associated with NO2)
2	Sofwan et al., (2021)	Kuala Lumpur	Assess the risks of exposure to air pollutants (PM10, CO, NO2 and SO2).	CO – Significant with NO2,SO2, humidity and temperature. CO – Highly associated with NO2
3	Shafii et al. (2017)	Klang	To evaluate the spatial variation pattern of air quality status in Klang, Selangor	CO - CO levels were shown to been significant with wind speed, humidity, temperature, SO2, NO2, PM10 and O3 CO – Highly associated with NO2
4	Kamaruzzaman et al., (2017)	Putrajaya	To evaluate the air pollution in Putrajaya	CO - CO were significant with SO2, NO2, O3 and PM10. CO – Highly associated with NO2
5	Parveen et al. (2021)	New Delhi , India	Examine the association between meteorological variables and air pollutant.	CO - CO were significant with meteorological factors such as with humidity, temperature, wind speed and rainfall CO - Highly associated with NO2
6	Liu et al., (2021)	California , USA	To further analyse between PM2.5, PM10 and gaseous pollutants	CO- CO were significant with PM10 and PM2.5 CO- Shows a positive relationship between CO and both PM10 and PM2.5