UNIVERSITI TEKNOLOGI MARA

A STUDY ON AIR POLLUTION INDEX IN SABAH AND SARAWAK USING PRINCIPAL COMPONENT ANALYSIS AND ARTIFICIAL NEURAL NETWORK

NUR ELISSA SYAZRINA BINTI ZULKEPLI

BACHELOR OF SCIENCE (Hons.)
MANAGEMENT MATHEMATICS

JULY 2020

Universiti Teknologi MARA

A Study on Air Pollution Index in Sabah and Sarawak Using Principal Component Analysis and Artificial Neural Network

NUR ELISSA SYAZRINA BINTI ZULKEPLI

**Report submitted in fulfillment of the requirements for
Bachelor of Science (Hons.) Management Mathematics
Faculty of Computer and Mathematical Sciences**

July 2020

SUPERVISOR'S APPROVAL

A STUDY ON AIR POLLUTION INDEX IN SABAH AND SARAWAK USING PRINCIPAL COMPONENT ANALYSIS AND ARTIFICIAL NEURAL NETWORK

By

NUR ELISSA SYAZRINA BINTI ZULKEPLI
2017136539

This report was prepared under the direction of supervisor, Norwaziah binti Mahmud. It was submitted to the Faculty of Computer and Mathematical Sciences and was accepted in partial fulfilment of the requirements for the degree of Bachelor of Science (Hons.) Management Mathematics.

Approved by:

……………………………………
Norwaziah binti Mahmud
Supervisor

AUGUST 5, 2020

STUDENT'S DECLARATION

I certify that this report and the research to which it refers are the product of my own work and that any ideas or quotation from the work of other people, published or otherwise are fully acknowledged in accordance with the standard referring practices of the discipline.

…………………………………

NUR ELISSA SYAZRINA BINTI ZULKEPLI

2017136539

AUGUST 5, 2020

ACKNOWLEDGEMENTS

Alhamdulillah, praise and thank to Allah because of His Almighty and His utmost blessings, I was able to finish this research within the time duration given.

Firstly, my special thanks and gratitude go to my dear supervisor and lecturer, Madam Norwaziah binti Mahmud and Madam Diana Sirmayunie binti Mohd Nasir, for their continuous encouragement and support for helping me. I am also grateful for their patience in helping me to understand the need to complete this study.

Special appreciation also goes to my beloved parents. My final year project could not be accomplished without the support of my parents and family members. During the hard period their encouragement and support is much appreciated and noted daily. Knowing they are always there with me through ups and downs has been a great relief. Thanks again to my parents and family members for being my pillar and helping me to do better.

I would also like to record my sincere thanks to all my friends, classmates for their help and sharing information. Also, special thanks to my best friends who are very noteworthy and trusting individuals who give all their support and do not hope for anything in return. Thank you very much to all my friends.

Finally, I have recorded my sense of gratitude for all those who have given their hands directly or indirectly to listen and help me solve my problems.

ABSTRACT

Nowadays, the analysis of the level of air pollutants is very necessary for environmental science research. The increased air pollution in Sabah and Sarawak attracted the attention of all Malaysians. Extended exposure to air pollution would lead to health problems. This study focused on identifying trends in air quality in Sabah and Sarawak based on data from the Department of Environment (DOE). There are five selected Malaysian monitoring stations in Sabah and Sarawak based on five air pollutants over four years (2015-2018). The aim of this study is to classify the indicator of variable predictors using the principal component analysis (PCA) method and to compare the best model for predicting air pollution index (API) in Sabah and Sarawak using the artificial neural network (ANN) model. The PCA environmental approach is used to identify sources of air pollution. ANN is used to compare the best model for predicting the API in Sabah and Sarawak. After the varimax rotation, only two pollutants (PM_{10} and NO_2) were the most significant pollutants out of the other five pollutants. These two pollutants used as input layers in Model B and the five pollutants used as input layers in Model A. These two models were used to compare the best model in the ANN method. The output of the ANN models evaluated by the coefficient of determination (R^2) and the root mean square error (RMSE). To identify the best model, declare the highest value of R^2 and the smallest value of RMSE. The findings indicate that the ANN technique has been successfully implemented as a decision-making tool as well as problem-solving for proper management of the atmosphere.

Keywords: Air pollution index (API), Principle component analysis (PCA), Artificial neural network (ANN), Varimax rotation, Eigenvalues

TABLE OF CONTENTS

CONTENTS **PAGE**

CHAPTER ONE: INTRODUCTION

CHAPTER TWO: LITERATURE REVIEW

CHAPTER THREE: RESEARCH METHODOLOGY

CHAPTER FOUR: RESULTS AND DISCUSSIONS

CHAPTER FIVE: CONCLUSIONS AND RECOMMENDATIONS

APPENDICES

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Information on air quality in Sabah and Sarawak is provided in this chapter. A problem arising from the quality of the air and the reason why this research is being conducted is also explained.

1.1 Background of the Study

Air pollution is one of the most important environmental issues in the world today. Pollution can take many forms, such as air pollution, water pollution, ground pollution and noise pollution. Air pollution can be defined as the pollution of the atmosphere by another contaminant that is life-threatening. It is also a form of pollution that affects the environment, which is typically caused by smoke and other harmful gases, namely oxides of carbon, sulfur and nitrogen. Air pollution does not only endanger human health, but it would not be appropriate to live in the future if people continue to increase air pollution.

Based on previous research, Ku Yusof et al. (2019) found that three of Malaysia's major sources of air pollution are heavily supplied by industry, motor vehicles and open combustion. In the last few years, Indonesia has been blamed for the burning of the forest. Smoke from burning has spread to Malaysia, causing the worst polluted area to be in Sabah and Sarawak. Sabah and Sarawak are facing the highest level of air pollution readings. The highest air pollution index (API) was found in Sri Aman, Sarawak, which is approximately 395. This is the first area in Malaysia to have a hazardous level since the transboundary haze caused by burning agricultural practices began to choke the country. There are three areas in Sarawak that continue to hoover at very unhealthful levels. While in Sabah, there are also two areas with unhealthy API levels. The regions are Kuching (236), Samarahan (202), Sri Aman (225), Sibu (190) and Sarikei (183) (The Star, 2019).

However, forest burning in Indonesia cannot be blamed solely on the fact that the local industry in Malaysia has also caused air pollution. The smoke emitted by factories is not sufficiently filtered to cause the smoke to become dangerous and to affect human health.

Kingsy et al. (2016) explain that air pollution has caused the greatest death in the world. The World Health Organisation (WHO) concludes that 2,400,000 people die every year as a result of air pollution (WHO Department of Public Health, 2019). Abd Rani, Azid, Khalit, Juahir, and Samsudin (2018) claim that statistics from the Malaysian Ministry of Health show that 0.1036 of the deaths are from respiratory diseases. The Star (2019) explained that due to poor air quality, the number of patients receiving conjunctivitis, asthma and skin rashes in Labuan increased by 40%. There is a different impact on human health for each pollutant. For example, particulate matter below 10 microns (PM_{10}) can lead to lung cancer, while ozone (O_3) can reduce lung function and cause coughing. In addition, the presence of carbon monoxide (CO) can affect fetal growth, while nitrogen monoxide (NO) can cause respiratory distress with symptoms such as cough, nasal congestion and sore throat, and also sulfur dioxide ($S0_2$) can affect people with asthma as the respiratory tract is limited.

In order to be aware of the pollution, citizens have been provided with easy to understand knowledge of air pollution, which has been established in Malaysia since 1989. Five requirements are used to measure API for air pollutants such as ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and particulate matter under 10 microns (PM_{10}) used to calculate API (Azid, Juahir, Latif, Zain, & Osman, 2013). Abd Rahman, Lee, Suhartono and Latif (2015) explain that the Malaysian API was developed on the basis of the API introduced by the United States Environmental Protection Agency (USEPA). The index value set, as shown in Table 1.1, can be categorized as good, moderate, unhealthy, very unhealthy, and hazardous, which may represent air quality status.

Table 1.1: Air Pollutant Index (API) of Malaysia

API	Air Pollution Level
0 to 50	Good
51 to 100	Moderate
101 to 200	Unhealthy
201 to 300	Very Unhealthy
301 and above	Hazardous

(Source: DOE, 2019)

There are many studies to evaluate the air quality index, but not many focuses on Sabah and Sarawak. The objective of this study is therefore to evaluate the air quality index in Sabah and Sarawak using the principal component analysis (PCA) and the artificial neural network (ANN).

1.2 Problem Statement

Good air quality is a key element of human life. People need to take care of the air to maintain good air quality. Air pollution is a phenomenon that Malaysia is faced with every year. Long-term exposure to air pollution would lead to health problems, especially for children and senior citizens. In the short term, air pollution would result in an increased number of cases of asthma attacks, bronchitis and even heart failure. Although children who were exposed to air pollution, even in the womb, were at risk of developing asthma, coughing and lung cancer when they were older. Moreover, when reading the API is getting worse and the school does not take any action, it may have an impact on the health of students. Recently, Sabah and Sarawak are facing the highest air pollution record. In addition, there is only a small amount of air pollution research that has been done in Sabah and Sarawak. Therefore, a study must be conducted to predict future air quality.

1.3 Objective of the Study

The aim of this study is

1. To classify the indicator of the variables predictors by using the PCA method.
2. To compare the best model to predict API in Sabah and Sarawak by using ANN method.

1.4 Scope of the Study

PCA will be used to find the important air pollution index predictors while ANN will be used to predict future air quality in Sabah and Sarawak. Data were provided by the Department of Environment (DOE) of Malaysia in terms of substance measurement, namely ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and particulate matter less than 10 microns (PM_{10}). It was acquired on a daily basis from 2015 to 2018.

1.5 Significance of the Study

The study of the application of forecasting to the air quality pattern can help to find the most important pollutant parameters in Sabah and Sarawak. In addition to providing a clear identification of significant air pollutant parameters, the prediction formulation will also provide a better understanding of air quality patterns. This will benefit the Ministry of Education in Malaysia from taking a step forward to allow all schools to close when reading the API is very unhealthful. In addition, this study will also help the Ministry of Tourism to be aware of what might happen in the future to ensure that all tourists are protected. Moreover, the use of a forecasting model can solve a complex problem that involves a lot of information or vast data.

CHAPTER 2
LITERATURE REVIEW

This chapter explains the previous study on the principal component analysis and the artificial neural network used in the Sabah and Sarawak air pollution index. This chapter focuses on the introduction of the principal component analysis and the artificial neural network and the application of this method to real-world problems.

2.1 Principal Component Analysis (PCA)

The PCA is a powerful tool for data analysis because high-dimensional data structures can be difficult to find, where the luxury of graphical representation is not in order. Mohamad, Ash'aari, and Othman (2015) explain that this technique is used for defining linear combinations of the initial variables which are useful for accounting for variations in those variables. PCA provides the most important and relevant variables indicating the source of the variance, since less important variables are excluded from its entire data set during the study, with very limited loss in the original information.

PCA can be used to modify the size of a large data set. According to Azid et al. (2014), PCA is used to describe the variability of the wide number of interconnected factors by turning these towards a different, small set of uncorrelated (independent) factors, known as principal components (PCs). PCs are the eigenvector of a matrix of covariance or a matrix of correlation, and each PC derives a maximum proportion of the total variance. A PC that contains eigenvalue larger or equal to 1 is used and considered significant to obtain the new variables (Alonso, Arribas, Manzoor, & Caceres, 2018).

Dominick, Juahir, Latif, Zain, and Aris (2012) explains that after rotation, the loading factor is important as it represents how much the variable contributes to that particular PC and to what degree one variable is identical to another. The value of factor loading which is larger than 0.75 are known as strong, the values between 0.50 to 0.75 are known as moderate and the value between 0.30 to 0.49 is known as weak factor loading (Azid et al., 2014). Besides, in order to measure the sampling adequacy, the Kaiser-Meyer-Olkin (KMO) test will be carried out. According to Azid et al. (2014), the value of the KMO test must be larger than 0.5 if not it will consider dropping from the analysis.

2.2 Artificial Neural Network (ANN)

A multi-layer perceptron (MLP) was commonly used as an ANN. The design has been defined by three basic layers of processing units that are the input layer, hidden layer, and output layer and it is interconnected. One or more layers called the hidden layer to distinguish the input and output. The hidden layer was used to make a description of the details acquired by the input nodes in order to carry out the non-linear mapping between input and output (Abd Rahman et al., 2015). However, the amounts of neurons in the three different layers rely on the problems.

According to Singh, Basant, Malik, and Jain (2009), the network does not have enough degrees of freedom to study the procedure properly when the value of hidden neurons is small. When the value of hidden neurons is big, the training process may take a long time. The arrows connecting one layer to another layer shows the strength of each connection, which can be determined by weight.

Abd Rahman, Lee, Latif, and Suhartono (2013) explains that forward and backward are the two main phases that are included in the network. According to Azid et al. (2014), the training data is spread by the hidden layer during the forward process and the resulting value contrast to the actual values for calculating the error among them. After that, the error that has been calculated is disseminated to the hidden layer. ANN’s study is divided into three parts which are data training (60%), testing (20%), and validation (20%). In the training phase, the data point is used to predict and study the shape of the parameters. In the testing phase, it is chosen to analyze the generalization capabilities of the networks that are said to be trained while for the validating phase, it is liable for carrying out the last inspection on the network trained (Azid et al., 2014).

2.3 Application of Artificial Neural Network

In real-world issues, there are some applications of ANN. For example, the application of ANN in the stock market index. Currency predicting is a major financial issue that needs more attention. ANN is among the effective ways of modelling market value, which can simply adjust to the changes in the market and does not accommodate usual formulas (Guresen, Kayakutlu, & Daim, 2011). This study appraises the efficacy of models of neural networks that considered to be complex and efficient in the forecasts stock-market. ANN is also used in rainfall forecasting in Queensland, Australia (Abbot & Marohasy, 2012). Researcher usually use climate indices to predict rainfall in Queensland, but the models have so far restricted them to considering combinations of linear correlations individually. By using ANN, it has the capacity to review big values of climate indices and another input together in order to search the settlement independently of the relationship is considered. Many rainfall prediction applications use a feed-forward neural network using the generic MLP equipped with the back-propagation algorithm.

Wind speed forecasting also uses ANN method. The wind speed forecasting proposed ANN to make a one-step prediction in advance and it will do it well when the wind data does not swing viciously. However, a study indicts the forecaster of ANN was 10% preferable than the persistence model for forecasting (Li & Shi, 2010). In addition, predicting students' performance also one of the applications that use ANN. Livieris, Drakopoulou, and Pintelas (2012) use ANN as a tool for predicting student's performance in order to identify weak students with learning problems in time. It is a very interesting and difficult task to establish an exact prediction model based on a classifier to determine weak students. It is also the best algorithm for learning to create a precise prediction model.

Other than that, forecasting the hardness of wood through heat treatment is one of the applications used in the ANN method. It is typically used in the wood sector to finish the problem of storing products, decrease the number of experiments and optimize the operations. It is also used to predict the hardness change during heat treatment. According to Van Nguyen, Nguyen, Ji, Do, and Guo (2018), the data are separated into three-layer to investigate the impacts of time and temperature towards the hardness of the wood. ANN also can be used in network traffic anomaly prediction. It is used to predict future number of network traffic anomalies so that the network traffic anomalies might be avoided. Ciptaningtyas, Fatichah, and Sabila (2017) stated that ANN is the best network traffic anomaly prediction tool because of the study outcomes shows that MSE is 0.003856.

2.4 Other Methods use to Evaluate Air Pollution Index

Many methods have been used to study the API. For example, multiple linear regression models (MLR). Based on Ku Yusof et al. (2019), MLR models have been used to estimate the value of PM_{10} during haze and non-haze seasons. Based on the results, shows that carbon monoxide clearly has a strong relationship with PM_{10} at 66.32%. Besides, the method that is used to evaluate API is the fuzzy time series (FTS). The fuzzy set has been used as membership, which means that the concentrations of air pollution quality are characterized in terms of good, moderate, and poor. According to Abd Rahman et al. (2013), FTS has been chosen to apply in the management of air quality as it will help to plan and improve better air quality in the future. It is essential in FTS to recognize the best inputs to attain the need for the fuzzy logical group (FLG) for a better outcome of forecasting.

The logistic regression model is among the approaches used in the index of air pollution or in haze estimation. The prediction of the seasonal element coefficient, the occurrence percentage of haze weather in winter was the greatest, which is 44.67% while the ratio in summer was the least with 22.80%. The cumulative logistic model was used to examine but also to evaluate the possibility of every kind of haze weather (Zhu, Zhang, & Chen, 2017). The other method that is used in the air pollution index is fuzzy logic. According to Nejadkoorki (2011), fuzzy logic is used to recognize air pollution and areas at risk by numerical value to assist the Air Pollution Control District (APCD). The fuzzy logic approach is possible to increase the utilization of decision making. This fuzzy also use for ranking the air pollution risk, dangerous, and high risk are gratifying value from 0 and 1.

Other than that, the *K*-Means clustering algorithm is one of the methods used to evaluate API. It is used to evaluate air pollution in order to increase the precision and competence of the *K*-Means clustering algorithm. In this method, the data points are categorized and separated into the same parts of *K* and the early centroids are taken at the centre of each part. Kingsy et al. (2016) state that the *K*-Means clustering algorithm provides high values of air quality index and low implementation time which is different from other techniques. Descriptive statistics is also one of the methods that are used for API. The analysis of this method is carried out by using the resource of air pollution, effects of air pollution, and others by using the frequency tables, cross-tabulation, descriptive statistics, and chi-square test of independence (Ismail & Ahmed, 2018). Ismail and Ahmed (2018) state that, by using this method, the biggest sources of air pollution have been found.

2.5 Summary

As a conclusion, several experiments are carried out using different techniques and the ANN. It is also usually been used in other studies. Most of the previous studies have stated that ANN is the most effective technique to use, even though the API can be measured using a variety of algorithms (Azid et al., 2014). This method is ideal for analyzing the index of air pollution to forecast future conditions. Besides, this technique is also the best technique compared to other techniques that had been used for predicting the API. There is also least study being carried out in Sabah and Sarawak. Therefore, the ANN method is an appropriate approach to be used in this study.

CHAPTER 3
RESEARCH METHODOLOGY

This chapter discusses the technique used by XLSTAT software and JMP software to analyze principal components analysis and artificial neural network methods. This chapter also gives a brief explanation of how the principal component analysis and the neural network were used. XLSTAT software and JMP software is an appropriate tool which can be used to predict the air pollution index in Sabah and Sarawak.

3.1 Method of Data Collection

The data used in this study are secondary data collected from the Department of Environment (DOE) of Malaysia containing the Malaysian air pollution index. The year used for this research is from 2015 to 2018, which is based on certain areas in Sabah and Sarawak on a daily basis. The data are categorized into several pollutants: ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and particulate matter of less than 10 microns (PM_{10}). Five continuous air monitoring stations were selected. The stations are located at Kuching (CA0004), Kota Kinabalu (CA0030), Sibu (CA0026), Labuan (CA042) and Tawau (CA0039).

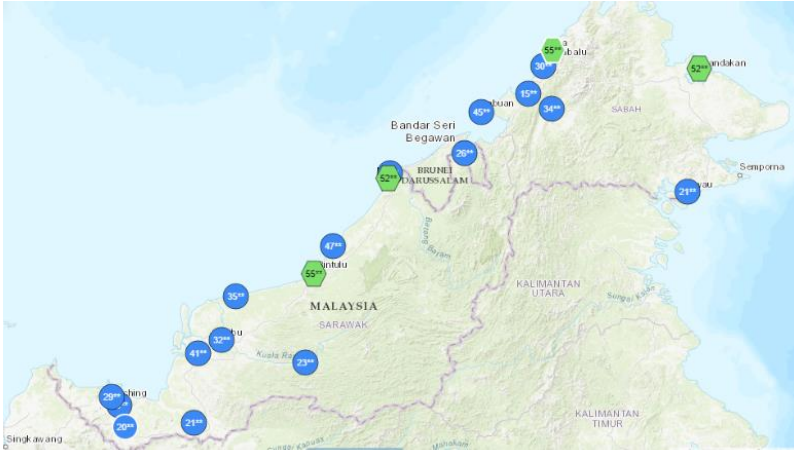


Figure 3.1: Location Map for Air Monitoring Stations in Sabah and Sarawak

3.2 Method of Data Analysis

The analysis of data using PCA and ANN is explained in Figure 3.2. There are several steps that need to be taken to achieve the objectives.

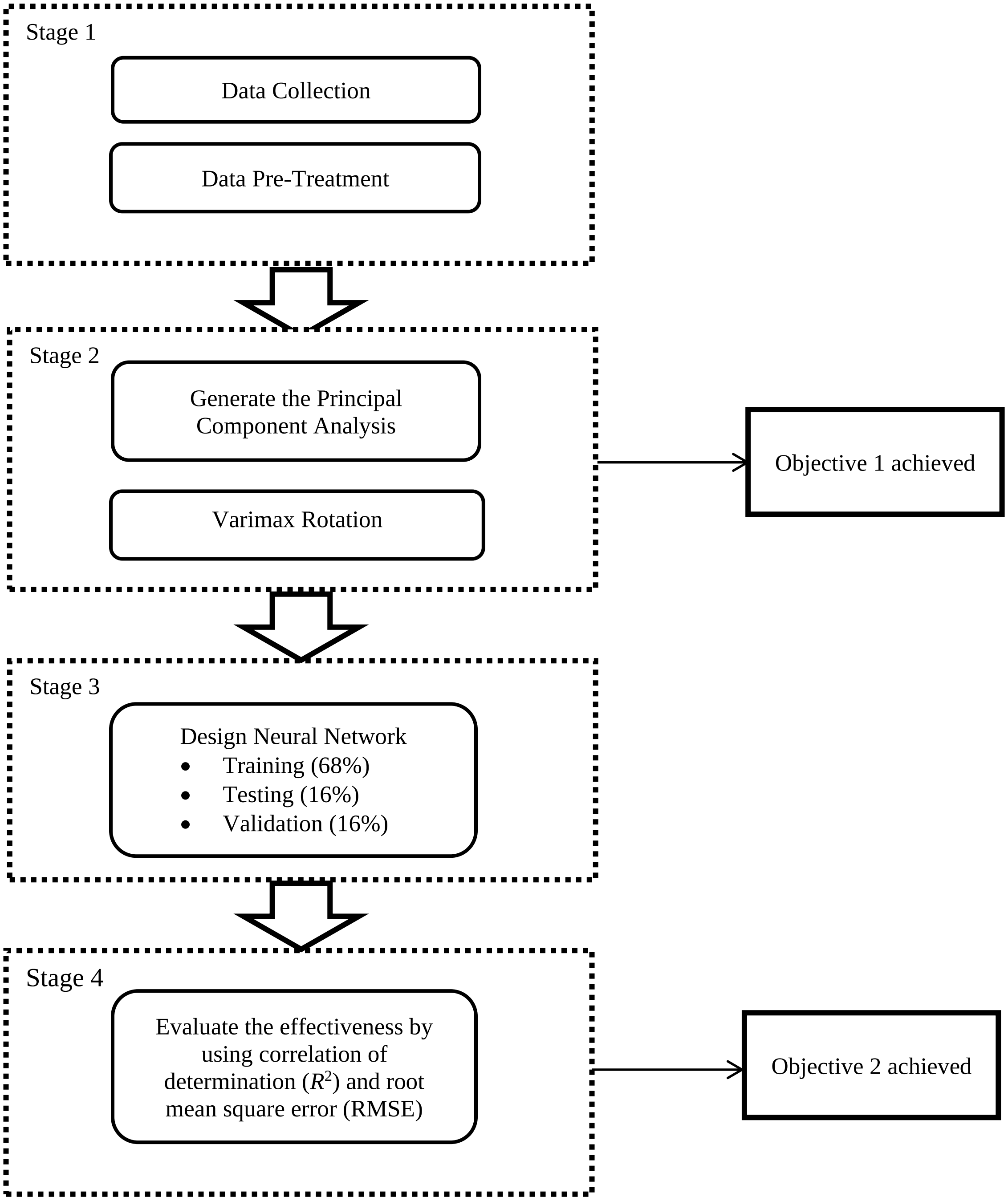


Figure 3.2: Procedures for the PCA and ANN model

3.2.1 Data Collection

The prediction model was developed in this study using 36,525 (5 variables × 7,305 observations). The amount of data missing from the data collection was very small (~3%) compared to the general data collection. The data are collected by the Division of Air Quality, DOE.

3.2.2 Data Pre-Treatment

In the present study, a number of steps for the PCA and ANN models are presented in Figure 3.2. If a data set is missing, the nearest neighbour method will be used to treat cases using the XLSTAT 2019 add-in tools. This method looks at the gap for each point and at the closet point. As a result, the nearest neighbour method was used to estimate the missing value based on the gap endpoints used by Eqn. (3.1) and Eqn. (3.2) respectively.

$$y = y_1 \ if \ x \leq x_1 + \frac{x_2 - x_1}{2}, \tag{3.1}$$

or

$$y = y_2 \ if \ x > x_1 + \frac{x_2 - x_1}{2}, \tag{3.2}$$

where *y* is the interpolant, *x* is the interpolant’s time point, the gap starting point coordinates are y_1 and x_1, and the gap endpoints are y_2 and x_2.

3.2.3 Principal Component Analysis (PCA)

The function of a large data set can be minimized by using the PCA, which is considered to be one of the most common and effective statistical methods for determining the possible structure of a set of variables. PCA appears to be able to demonstrate the most important factors that may indicate the source of the contaminants, while the less significant variables are excluded from the collection of data with minimal loss of original information. This approach is used by converting them to a new one to describe the variability of a large number of interconnected variables, smaller non-correlated (independent) variables, called principal components (PCs). In order to produce PCs, the PCA must be performed. The PCs are used as input variables for the API prediction model using the ANN approach. The PCs can be expressed as the Eqn. (3.3):

$$z_{ij} = a_{i1}x_{1j} + a_{i2}x_{2j} + ... + a_{im}x_{mj} \tag{3.3}$$

Where,

z	=	Component score
a	=	Component loading
x	=	Measured value of the variable
i	=	Component number
j	=	Sample number
m	=	Total number of variables

To calculate the PCA, the researcher must first click on the XLSTAT software data analysis menu and then select the “Principal Component Analysis” option. Once the dialog box is in place, select the variables needed for the account in the classification. Pearson type PCA has been chosen. When the variable name has been included, the option “variable labels” must be selected. The result will be shown in three choices, the range, the sheet and the workbook. Upon leaving the general tab the options must be picked. There is no need to select any options and the additional data tab. However, for the missing data tab, the option “do not accept missing data” is chosen, which was set as default. For the outputs tab, all results are selected to be displayed except for the Bartlett’s Sphericity Test. Colored labels and filters will not be available on the charts tab. The form of biplot is the correlation and the coefficient are set at n /p.

3.2.4 Varimax Rotation

It is advisable to rotate PCs generated by the PCA using varimax rotation because they are not easily interpreted. The use of varimax rotation was intended to reduce the complexity of the components by increasing large loads within the component and reducing small loads within the component. The varimax rotation method is used as this approach facilitates the structure of the variable and makes its analysis simpler and more accurate. In the varimax rotation process, only PCs with their eigenvalues greater than 1 will be used and considered important for obtaining new variables, i.e. varimax factors (VFs) or factor loads. This approach is referred to as the Kaiser Criterion. This criterion is used to solve the problem of the number of components to be preserved. The number of VFs to be used for varimax rotations is based on the number of variables that are consistent with the standard characteristics and may include non-observable, theoretical and latent variables. VFs are values used to calculate the correlation between variables. The values of the VFs are divided into three categories as shown in Table 3.2. For this analysis, the selection threshold for VFs with absolute values above 0.75 was set. The PCA analysis is implemented using XLSTAT 2019 add-in software (Azid et al., 2014).

Table 3.2: Category of Varimax Factors Values

Varimax Factors Values	Category
Greater than 0.75 (>0.75)	Strong
From 0.50 to 0.75 (0.50≥VF≥0.75)	Moderate
From 0.30 to 0.49 (0.30≥VF≥0.49)	Weak

(Source: Azid et al., 2014)

3.2.5 ANN – API Prediction Model

In this study, feed-forward ANN is used for prediction purposes and for the determination of the most important parameters affecting API values. This model is divided into three layers known as the input layer, the hidden layer and the output layer. The total input numbers (independent test set) and the hidden layer are calculated by the design of the research problem and based on the expected horizon, while the output layer (independent test set) has a single node. The value of the API is the output layers. With different input variables, two different feed-forward ANN models were developed as shown in Table 3.3.

Table 3.3: Feed-Forward ANN Models

Models	Descriptions
Model A	This model was built with five variables as input layers based on the original data.
Model B	This model was created by factor scores of rotated PCs with an input value of eigenvalues greater than 1

The structure of the ANN model is shown in Figure 3.3 and Figure 3.4. The trial-and-error procedure between one to ten hidden nodes in the network structure is tested to approximate any level of accuracy and the best model for predictive values is sought.

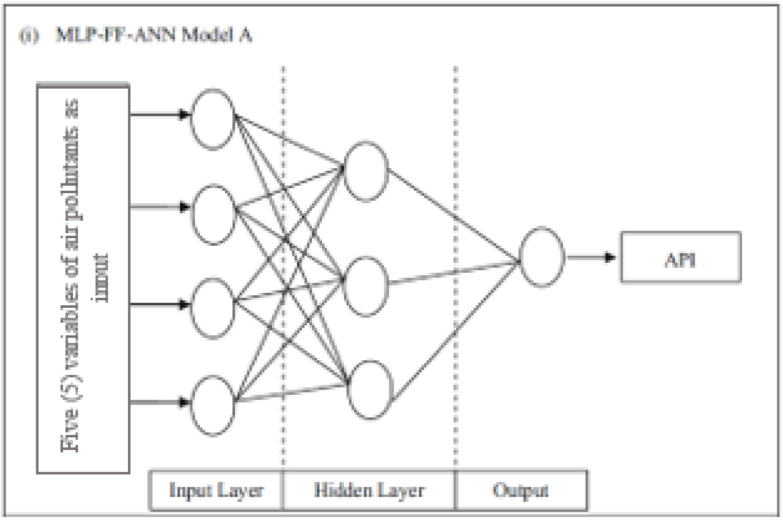


Figure 3.3: ANN model network structure for five variable of air pollutants

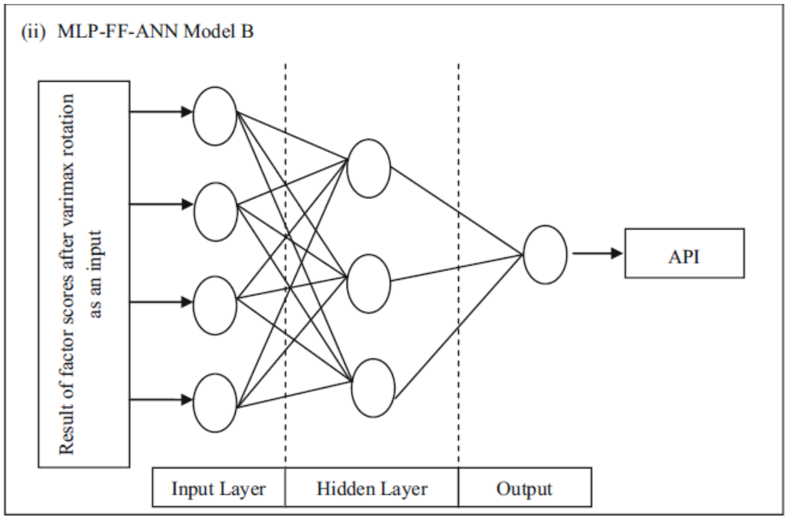


Figure 3.4: ANN model network structure after varimax rotation

The data set is divided into three classes which are training, testing and validation. The analysis of ANN undergoes three phases which the percentage value for each phase is determined by using Alyuda Neurointelligence software. For the training phase, the data set is used to estimate and understand the trends of the parameters. For the testing phase, it can be used to assess the potential for generalization of the allegedly trained network, while validation of the network is necessary to carry out the final test.

In order to calculate the ANN using JMP15, the researcher must select the analytical option and choose the predictive and neural option. Once the dialog box is in place, assign the output variable to the Y, Response role, and then assign the other variable to the X, Factor. For Holdback, the proportion enters 0.2 and for random seed enters 1234. Once there is a hidden node box, enter the number of nodes from one to ten for each model. Check the option transform covariates and click the go button.

3.2.6 Evaluate the Effectiveness

In order to determine the outcome of the ANN models, two performance features are reflected on the model evaluation, which is known as the coefficient of determination (R^2) and the root mean square error (RMSE). As best linear model, the highest value of R^2 and the lowest value of RMSE are stated. The R^2 criterion for efficiency is defined as Eqn. (3.4):

$$R^2 = 1 - \frac{\sum (x_i - y_i)^2}{\sum y_i{}^2 - \frac{\sum y_i{}^2}{n}} \tag{3.4}$$

Here, the numerator is the explained variation while the denominator is the total variation. The value of R^2 is between 0 and 1, which is explained in Table 3.4.

Table 3.4: Description of the Value R^2

Values	Description
Closer to 0	Extremely poor fit
Closer to 1	Perfect fit

(Source: Investopedia, 2019)

While in Eqn. (3.5), RMSE is counted where x_i refers to the data observed, y_i the expected data, and n is the amount data and percentage of initial uncertainty that the models explain.

$$RMSE = \sqrt{\frac{1}{n} \sum \frac{n}{i=1} \left(x_i - y_i \right)}^2 \tag{3.5}$$

After that, the expected number of ANN models will be contrasted with others in order to acquire a model that relies on a few variables for estimating APIs. These

ANN models are implemented using the JMP15 software, which is a flexible and easy-to-use tool.

CHAPTER 4
RESULTS AND DISCUSSIONS

The results of this study are briefly discussed in this chapter. The step by which the principal component analysis and artificial neural network are used to analyze and predict air quality patterns in Sabah and Sarawak will be discussed.

4.1 Result Analysis

4.1.1 Data Pre-Treatment

By using the steps in Chapter 3, the nearest neighbor method is used to estimate the missing data set. The endpoints of the gaps are used as the estimation value for all the missing data sets.

4.1.2 Result for PCA

After running the data using XLSTAT 2019, the result showed which pollutant is important for the determination of the air quality index. Table 4.1 presents the results of the Kaiser-Meyer-Olkin (KMO) test. The KMO test was used to assess the adequacy of sampling across all variables. This shows that all pollutants are greater than 0.5. This evaluates that all pollutants were sufficient and could be used for further study. To determine the associations between air pollutants and derived factors, the calculation of the varimax factor was carried out.

Table 4.1: Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy

Pollutant	**Result**
PM_{10}	0.558
SO_2	0.603
NO_2	0.599
O_3	0.677
CO	0.620

4.1.3 Identification of Air Pollution Source based on Varimax Rotation

Table 4.2 shows the result after the varimax rotation. As a result, only two of the main components (PCs) out of the five PCs were selected. Both PCs were chosen because of their eigenvalues exceeding 1 which constitute 61.528% of the total variance. However, the eigenvalues that were less than 1 were ignored due to their similarity with more significant factors. This shows that multicollinearity was present between the original variables. Although the total amount is below 70%, the deduction point has been calculated using the scree plot as shown in Figure 4.1.

Table 4.2: Descriptive statistics of selected original PCs

Variable	PC1	PC2
Eigenvalue	1.987	1.090
Variability (%)	39.733	21.795
Cumulative (%)	39.733	61.528

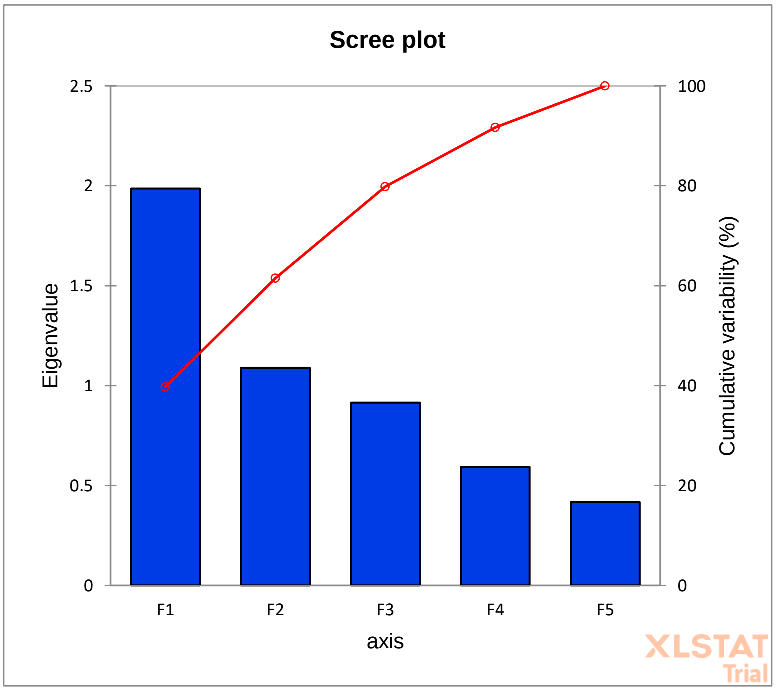


Figure 4.1: Scree plot for PCA

PC1 and PC2 were chosen as the eigenvalues are greater than 1 and are considered to be significant in the varimax rotation analysis. Thus, PC1 and PC2 are considered important for obtaining new variables known as the varimax factor (VF). Absolute values larger than 0.75 of the VF are also set as the threshold of choice throughout this analysis, since the number was strong, which shows moderate to heavy loads on the removed factors. Two out of five air pollutants meet the 0.75 VF threshold as shown in Table 4.3 and Figure 4.2. The variables with values greater than 0.75 shall be PM_{10} and NO_2. These pollutants have been the major contributors to pollutants at the selected monitoring stations in Sabah and Sarawak.

Table 4.3: Rotated factor loadings using two PCs

Variable	VF1	VF2
PM_{10}	0.415	**0.758**
SO_2	0.725	-0.449
NO_2	**0.822**	-0.244
O_3	0.521	-0.139
CO	0.585	0.485
Eigenvalue	1.987	1.090
Variability (%)	39.733	21.795
Cumulative (%)	39.733	61.528

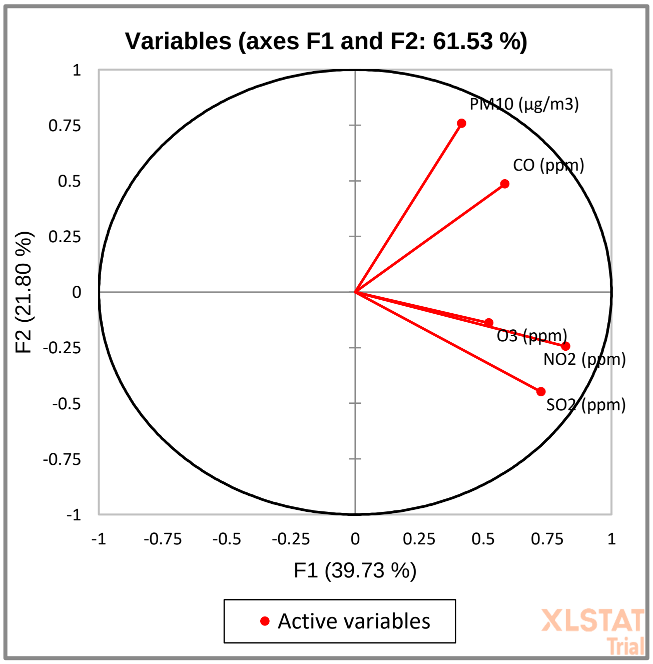


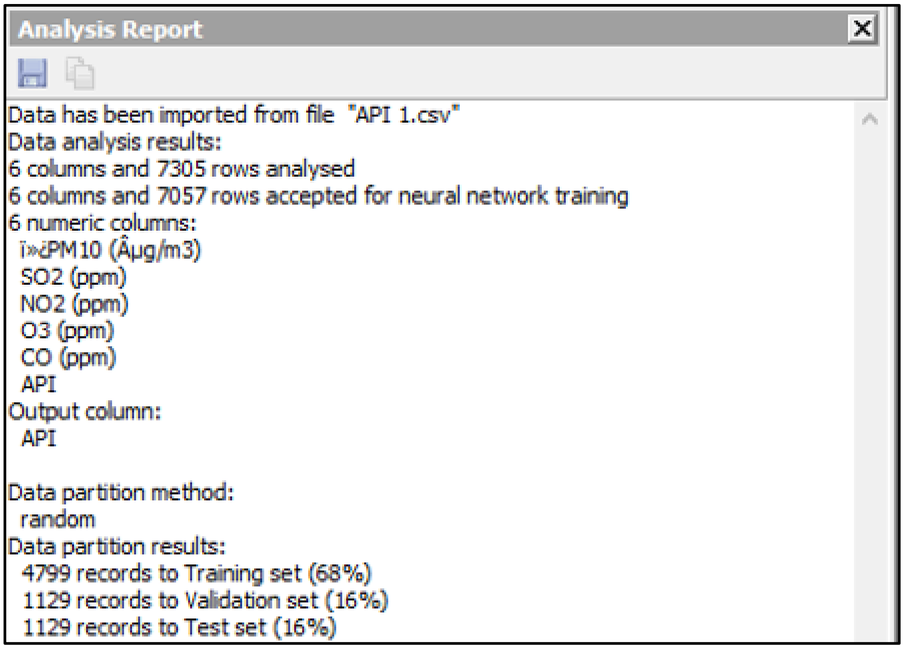
Figure 4.2: Varimax Factor plot after varimax rotation

The VF1 contributes approximately 39.733% to the quality of air data variability. Only one pollutant with absolute values exceeding 0.75, which is NO_2 with 0.822. Based on the category of VF values set out in Table 3.2, NO_2 has a strong VF value. The presence of NO_2 in urban outdoor air is large due to traffic. However, nitric oxide (NO) produced by automobiles or by any other combustion process interacts with atmospheric oxygen to produce NO_2. Indoor NO_2 was primarily released from non-fired heaters and gas stoves (Ministry of the Environment, 2018). While the VF2 contributes about 21.795% of the variability in air quality data. There is one pollutant that contains absolute values that exceed 0.75, which is PM_{10} at 0.758. PM_{10} also has strong VF values based on the category of VF values. PM_{10} was the main aspect of dust produced by construction operations and building sites (Azid et al., 2014). The key component of PM_{10} also includes harmful exhaust emissions, soil dust and open-burning operations. According to Azid et al. (2014), the number of most recent registered Malaysian motor vehicles increased by 4.42% between 2004 and 2010. This shows that the automobiles in Sabah and Sarawak are among the most important factors in terms of air quality.

4.2 Model evaluation of Air Pollution Index

4.2.1 Design the Neural Network

In this analysis, ANN models have been used for predictive purposes. The model contains three layers that are the data input layer, the hidden layer, and the output layer. In this study, two ANN models were developed. Models are Model A and Model B, where Model A includes all air pollutants as input layers, while Model B includes the results of the varimax factor resulting from the varimax rotation as input layers. The values of the API were used as output layers for both models. The trial-and-error process for the number of hidden nodes in a structured network has been tested. The data generated by the use of pre-treatment is divided into three classes for both models. The three classes are training, validation and testing of the data. In order to set the percentage for each class, Alyuda Neurointelligence software has been used by insert all the data. Figure 4.3 shows the data was partitioned into three different sets with different percentage. The percentage for each class is training (68%), validation (16%) and testing (16%).



```
Data has been imported from file  "API 1.csv"
Data analysis results:
6 columns and 7305 rows analysed
6 columns and 7057 rows accepted for neural network training
6 numeric columns:
  ï»¿PM10 (Âµg/m3)
  SO2 (ppm)
  NO2 (ppm)
  O3 (ppm)
  CO (ppm)
  API
Output column:
  API

Data partition method:
  random
Data partition results:
  4799 records to Training set (68%)
  1129 records to Validation set (16%)
  1129 records to Test set (16%)
```

Figure 4.3: Analysis Report

For the predictive purpose JMP15 software has been used. Figure 4.4 and Figure 4.5 shows the scatter plot of the output of the JMP15 software. The scatter plot for both models using the original raw data and the VF as input variable data is between the real and the expected API. The point is all along the line, meaning that the expected values were close to the actual values.

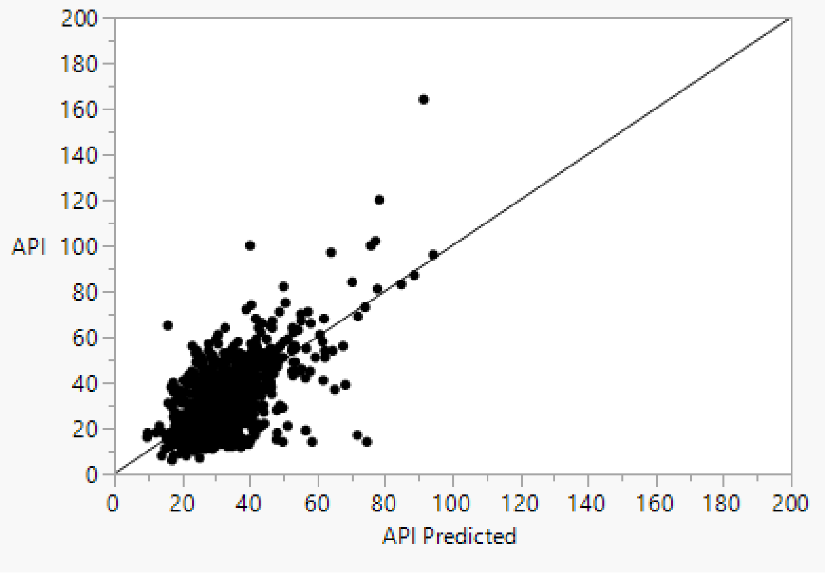


Figure 4.4: Scatter plot API prediction performance for original raw data

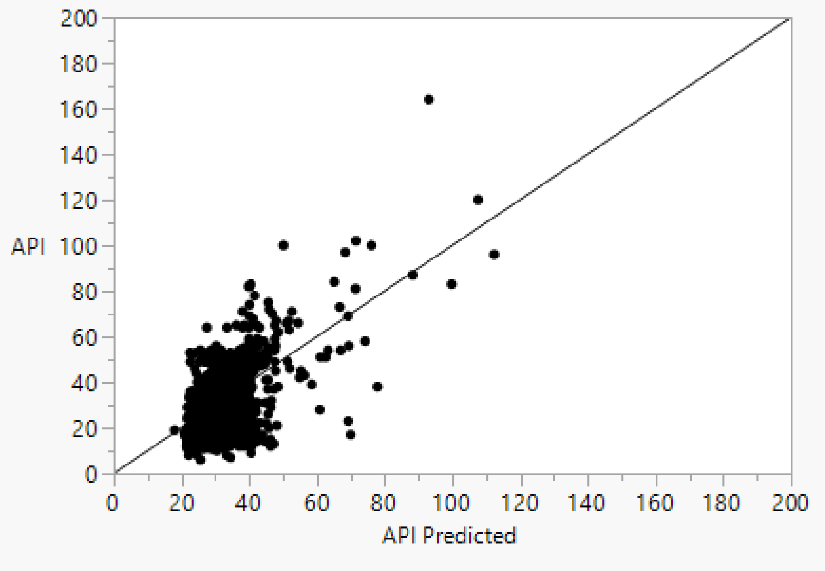


Figure 4.5: Scatter plot API prediction performance for rotated PCA scores

4.2.2 Result of the Effectiveness

Approximately 20 networks structured have been tested to develop the ANN models. R^2 and RMSE are the predictive performance results of both models, as shown in Table 4.4. Results were provided for both the training sets and the validation sets. Validation set results are used as a description of the predictive value of the model for future findings.

Table 4.4: Prediction Performance for Model A and Model B by using ANN

Model	Hidden node	R^2	RMSE
Model A	1	0.3081	11.6624
	2	0.3802	11.0378
	3	0.3689	11.1385
	4	0.3997	10.8625
	5	0.4322	10.5647
	6	0.4171	10.7044
	7	0.4278	10.6057
	8	0.4012	10.8496
	9	0.3951	10.9045
	10	0.4197	10.6808
Model B	1	0.2476	12.3443
	2	0.3107	11.8156
	3	0.3043	11.8704
	4	0.3115	11.8083
	5	0.3208	11.7286
	6	0.2924	11.9712
	7	0.3214	11.7229
	8	0.3157	11.7727
	9	0.3125	11.8002
	10	0.3256	11.6865

Five variables have been used as input variables in Model A. The R^2 and RMSE values for Model A are 0.4322 and 10.5647, respectively. This shows that only five hidden nodes that can fit for Model A. The value of R^2 is extremely poor because it is closer to 0. Besides, in Model B, only two variables were used as input variables, PM_{10} and NO_2. In Model B, the value of R^2 is 0.3256 and 11.6865 for RMSE. There were ten hidden nodes that can fit for Model B. The value of R^2 shows that it is also extremely poor fit because the value is closer to 0. However, Model B compares better results than Model A. Although, the R^2 value of the Model B is lower than Model A, but the models can estimate the API within acceptable accuracy. This means that the use of rotated PCs using varimax rotation is more efficient and effective due to reduction of predictor variables without losing important information. According to Azid et al. (2013), Model B is selected as the best model since most of the expected data are not significantly different from the actual data. Although the R^2 values are more accurate in Model A than in Model B, Model B uses less and less complex variables than Model A, which is the strength of this model. Model B therefore not only saves time, but also saves the cost of monitoring purposes. It is proved that Model B that use rotated PCs using varimax rotation as input variables are absolutely very useful tools in helping decision making and problem solving for better atmospheric management. It is therefore also proven that ANN models can be used to predict API values with negligible accuracy from all available inputs.

4.3 Summary

This chapter discussed that the result of PCA solution using XLSTAT 2019 shows the selected air pollutant for Sabah and Sarawak. In addition, the best model to predict APIs in Sabah and Sarawak can be solved by using ANN in JMP15 software.

CHAPTER 5
CONCLUSIONS AND RECOMMENDATIONS

This chapter discusses the conclusion of this study and the recommendation to improve the study in the future.

5.1 Conclusions

In conclusion, the main component analysis (PCA) can identify the most significant air pollutants in Sabah and Sarawak. The Kaiser-Meyer-Olkin (KMO) test indicates that sampling adequacy is greater than 0.5 and that all pollutants considered to have been eligible for further analysis. The two PCs produced by the rotated PCA show that two out of five pollutants are the most important pollutants that have caused air pollution in Sabah and Sarawak. Two pollutants selected for PM_{10} and NO_2. In addition, the artificial neural network (ANN) was divided into two models, Model A and Model B. Model A contains five pollutants as input variables, PM_{10}, SO_2, NO_2, O_3, and CO, while Model B uses only the selected pollutant based on the PCA result. The finding shows that Model A gives greater value to R^2 than Model B. However, Model B provides better predictions in terms of R^2 than Model A. Although the Model B forecast output is poorer than Model A, the models can predict the API to an appropriate degree of accuracy. This shows that only rotated PCs appear to be more efficient and effective in reducing air pollutants without missing any essential details. This can also prove that varimax rotational PCs and ANN models were the most important models to help make a decision and solve the problem of best environmental control.

5.2 Recommendations

There are a number of recommendations for future research. First, future research may use another predictive method that is not used in this study. For example, the fuzzy time series can be used instead of the ANN models. In addition, this study may also be continued with another method of mathematical analysis, such as multivariate techniques, spatial analysis, and other methods that are appropriate for data. The future researcher also can continue the methodology of this study in another field such as marketing, insurance, and education. In addition, the Malaysia Ministry of Environment can use this study as a reference to analyze and predict the air quality pattern in Peninsular Malaysia. Lastly, the next researcher can use other air pollutants such as methane (CH_4), non-methane hydrocarbon (NmHC), total hydrocarbon (THC), wind direction, wind speed, and humidity.

REFERENCES

Abbot, J., & Marohasy, J. (2012). Application of artificial neural networks to rainfall forecasting in Queensland, Australia. *Advances in Atmospheric Sciences*, *29*(4), 717–730. https://doi.org/10.1007/s00376-012-1259-9

Abd Rahman, N. H., Lee, M. H., Latif, M. T., & Suhartono. (2013). Forecasting of air pollution index with artificial neural network. *Jurnal Teknologi*, *63*(2), 59–64. https://doi.org/10.11113/jt.v63.1913

Abd Rahman, N. H., Lee, M. H., Suhartono, & Latif, M. T. (2015). Artificial neural networks and fuzzy time series forecasting: an application to air quality. *Quality and Quantity*, *49*(6), 2633–2647. https://doi.org/10.1007/s11135-014-0132-6

Abd Rani, N. L., Azid, A., Khalit, S. I., Juahir, H., & Samsudin, M. S. (2018). Air pollution index trend analysis in Malaysia, 2010-15. *Polish Journal of Environmental Studies*, *27*(2), 801–808. https://doi.org/10.15244/pjoes/75964

Alonso, D. N., Arribas, L. V. P., Manzoor, S., & Cáceres, J. O. (2019). Statistical tools for air pollution assessment: multivariate and spatial analysis studies in the Madrid region. *Journal of analytical methods in chemistry*, *2019*(6), 1-9. https://doi.org/10.1155/2019/9753927

Azid, A., Juahir, H., Latif, M. T., Zain, S. M., & Osman, M. R. (2013). Feed-forward artificial neural network model for air pollutant index prediction in the southern region of Peninsular Malaysia. *Journal of Environmental Protection*, *04*(12), 1–10. https://doi.org/10.4236/jep.2013.412a1001

Azid, A., Juahir, H., Toriman, M. E., Kamarudin, M. K. A., Mohd Saudi, A. S., Che Hasnam, C. N., et al. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: a case study in Malaysia. *Water, Air, and Soil Pollution*, *225*(8). https://doi.org/10.1007/s11270-014-2063-1

Ciptaningtyas, H. T., Fatichah, C., & Sabila, A. (2017, March). *Network traffic anomaly prediction using artificial neural network*. Paper presented at AIP Conference Proceedings. *1818*(1), p. 020010. AIP Publishing. https://doi.org/10.1063/1.4976874

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. http://dx.doi.org/10.1016/j.atmosenv.2012.06.021

Guresen, E., Kayakutlu, G., & Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, *38*(8), 10389–10397. https://doi.org/10.1016/j.eswa.2011.02.068

Investopedia (2019). *Coefficient of Determination*. Retrieved December 2, 2019, from https://www.investopedia.com/terms/c/coefficient-of-determination.asp

Ismail, S., & Ahmed, S. (2018). Air pollution, its sources and health effects : a case study of Delhi. *The Research Journal of Social Sciences, 9*(4), 62–74.

Kingsy, G. R., Manimegalai, R., Geetha, D. M. S., Rajathi, S., Usha, K., & Raabiathul, B. N. (2016, November). *Air pollution analysis using enhanced K-Means clustering algorithm for real time sensor data*. Paper presented at IEEE Region 10 Annual International Conference (TENCON) 2016. https://doi.org/10.1109/TENCON.2016.7848362

Ku Yusof, K. M. K., Azid, A., Abdullah Sani, M. S., Samsudin, M. S., Muhammad Amin, S. N. S., Abd Rani, N. L., & Jamalani, M. A. (2019). The evaluation on artificial neural networks (ANN) and multiple linear regressions (MLR) models over particulate matter (PM10) variability during haze and non-haze episodes: A decade case study. *Malaysian Journal of Fundamental and Applied Sciences, 15*(2), 164–172. https://doi.org/10.11113/mjfas.v15n2.1004

Li, G., & Shi, J. (2010). On comparing three artificial neural networks for wind speed forecasting. *Applied Energy*, *87*(7), 2313–2320. https://doi.org/10.1016/j.apenergy.2009.12.013

Livieris, I. E., Drakopoulou, K., & Pintelas, P. (2012, September). *Predicting students' performance using artificial neural networks*. Paper presented at the 8th Pan-Hellenic Conference with Information and Communication Technologies in Education (pp. 321-328).

Ministry for the Environment (2020). *Nitrogen dioxide*. Retrieved April 8, 2020, from https://www.mfe.govt.nz/air/specific-air-pollutants/nitrogen-dioxide

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. https://doi.org/10.1016/j.proenv.2015.10.021

Nejadkoorki, F. (2011). Air Pollution Monitoring Using Fuzzy Logic in Industries. In Vahdat, S. E., Nakhaee, F. M, *Advanced Air Pollution* (pp. 21-30). Iran. https://doi.org/10.5772/16947

Singh, K. P., Basant, A., Malik, A., & Jain, G. (2009). Artificial neural network modeling of the river water quality-a case study. *Ecological Modelling*, *220*(6), 888–895. https://doi.org/10.1016/j.ecolmodel.2009.01.004

The Star Online (2019, September 09). *More efforts needed to 'clear the air'*. Retrieved November 24, 2019, from https://www.thestar.com.my/news/nation/2019/09/09/more-efforts-needed-to-clear-the-air

The Star Online (2019, September 22). *Haze-related diseases rise by 40%, says Health DG*. Retrieved November 24, 2019, from https://www.thestar.com.my/news/nation/2019/09/22/haze-related-diseases-rise-by-40-says-health-dg

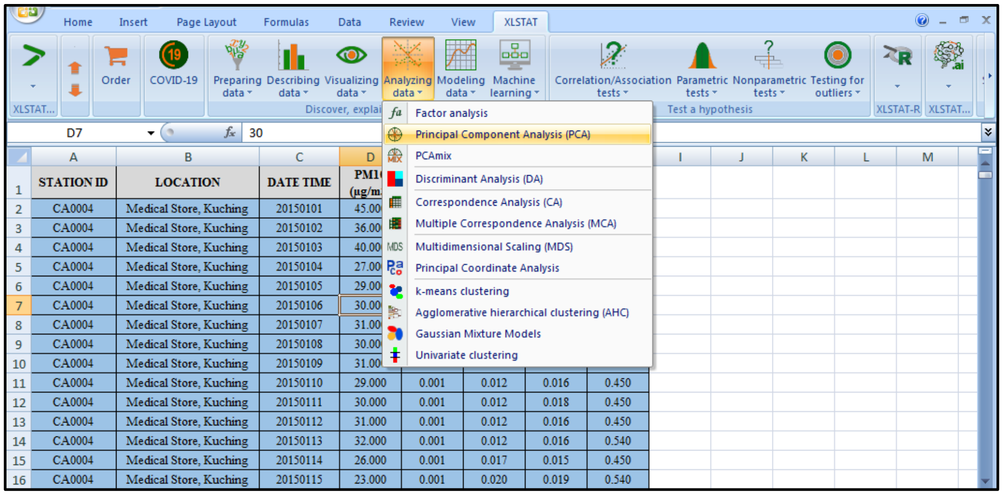
Van Nguyen, T. H., Nguyen, T. T., Ji, X., Do, K. T. L., & Guo, M. (2018, July). *Using artificial neural networks (ANN) for modeling predicting hardness change of wood during heat treatment*. Paper presented at IOP Conference Series: Materials Science and Engineering. https://doi.org/10.1088/1757-899X/394/3/032044

World Health Organization (2019). *Air Pollution*. Retrieved December 2, 2019, from https://www.who.int/

Zhu, Y., Zhang, T., & Chen, C. (2017). Study on probability estimation of haze in Beijing based logistic regression model. *Journal of Geoscience and Environment Protection*, *5*(06), 37. https://doi.org/10.4236/gep.2017.56005

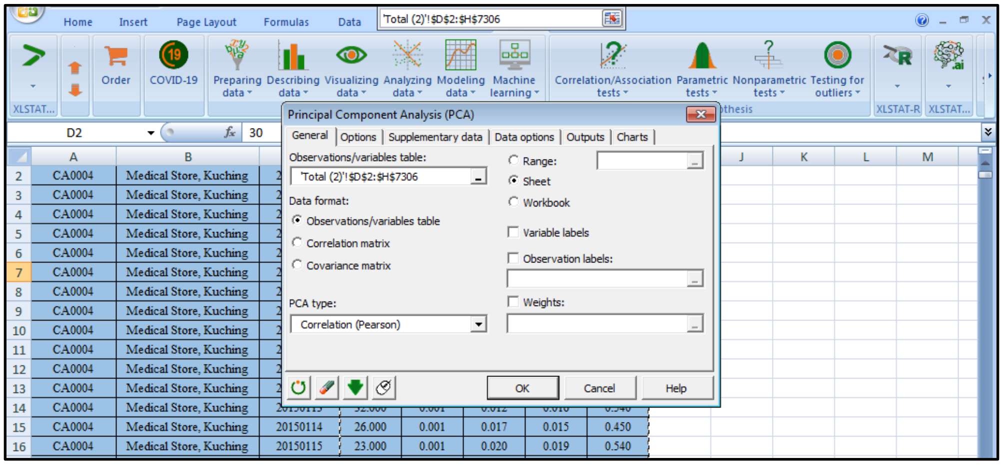
APPENDICES

APPENDIX A: OUTPUT SOFTWARE FOR XLSTAT 2019



D7 | 30

	A	B	C	D				
1	STATION ID	LOCATION	DATE TIME	PM1((µg/m				
2	CA0004	Medical Store, Kuching	20150101	45.00				
3	CA0004	Medical Store, Kuching	20150102	36.00				
4	CA0004	Medical Store, Kuching	20150103	40.00				
5	CA0004	Medical Store, Kuching	20150104	27.00				
6	CA0004	Medical Store, Kuching	20150105	29.00				
7	CA0004	Medical Store, Kuching	20150106	30.00				
8	CA0004	Medical Store, Kuching	20150107	31.00				
9	CA0004	Medical Store, Kuching	20150108	30.00				
10	CA0004	Medical Store, Kuching	20150109	31.00				
11	CA0004	Medical Store, Kuching	20150110	29.000	0.001	0.012	0.016	0.450
12	CA0004	Medical Store, Kuching	20150111	30.000	0.001	0.012	0.018	0.450
13	CA0004	Medical Store, Kuching	20150112	31.000	0.001	0.012	0.016	0.450
14	CA0004	Medical Store, Kuching	20150113	32.000	0.001	0.012	0.016	0.540
15	CA0004	Medical Store, Kuching	20150114	26.000	0.001	0.017	0.015	0.450
16	CA0004	Medical Store, Kuching	20150115	23.000	0.001	0.020	0.019	0.540



'Total (2)'!$D$2:$H$7306

D2 | 30

Principal Component Analysis (PCA) — General

Observations/variables table: 'Total (2)'!$D$2:$H$7306

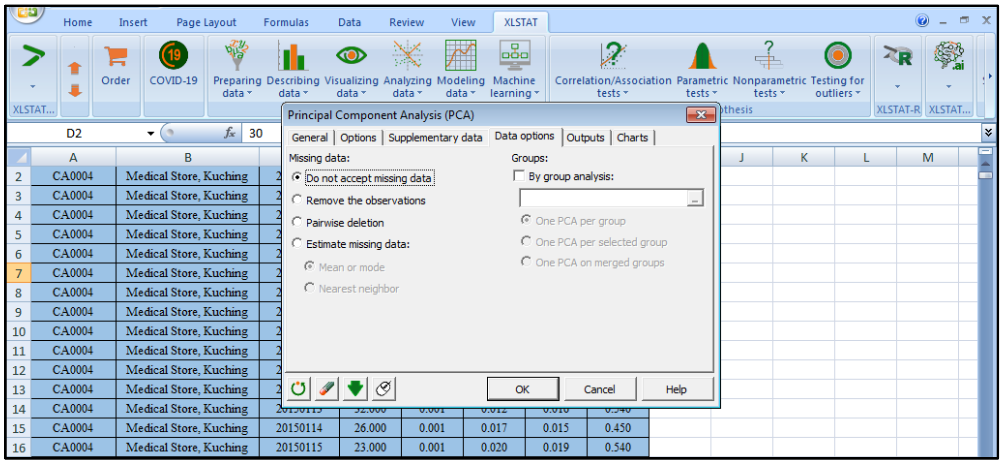
Data format:
● Observations/variables table
○ Correlation matrix
○ Covariance matrix

PCA type: Correlation (Pearson)

○ Range:
● Sheet
○ Workbook

□ Variable labels
□ Observation labels:
□ Weights:

	A	B	C	D				
2	CA0004	Medical Store, Kuching	2					
3	CA0004	Medical Store, Kuching	2					
4	CA0004	Medical Store, Kuching	2					
5	CA0004	Medical Store, Kuching	2					
6	CA0004	Medical Store, Kuching	2					
7	CA0004	Medical Store, Kuching	2					
8	CA0004	Medical Store, Kuching	2					
9	CA0004	Medical Store, Kuching	2					
10	CA0004	Medical Store, Kuching	2					
11	CA0004	Medical Store, Kuching	2					
12	CA0004	Medical Store, Kuching	2					
13	CA0004	Medical Store, Kuching	2					
14	CA0004	Medical Store, Kuching	20150113	32.000	0.001	0.012	0.016	0.540
15	CA0004	Medical Store, Kuching	20150114	26.000	0.001	0.017	0.015	0.450
16	CA0004	Medical Store, Kuching	20150115	23.000	0.001	0.020	0.019	0.540



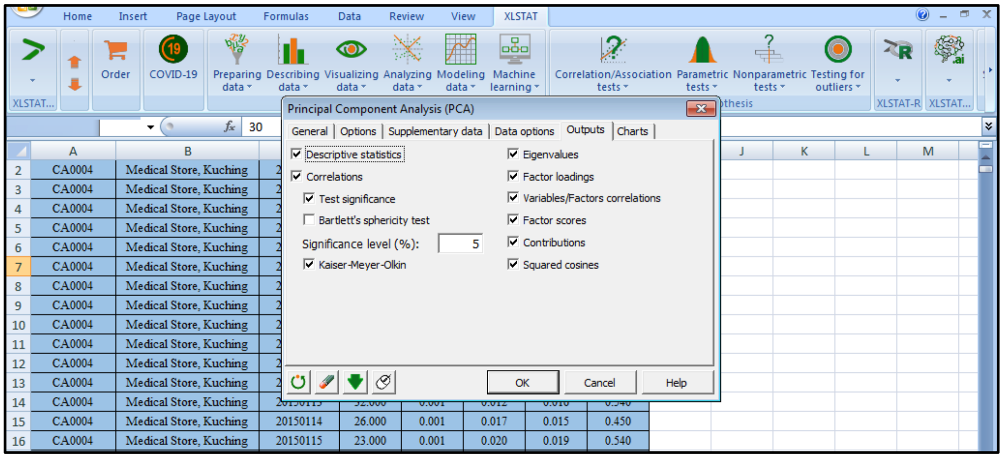
D2 | 30

Principal Component Analysis (PCA) — Data options

Missing data:
● Do not accept missing data
○ Remove the observations
○ Pairwise deletion
○ Estimate missing data:
● Mean or mode
○ Nearest neighbor

Groups:
□ By group analysis:
● One PCA per group
○ One PCA per selected group
○ One PCA on merged groups

	A	B	C	D				
2	CA0004	Medical Store, Kuching	2					
3	CA0004	Medical Store, Kuching	2					
4	CA0004	Medical Store, Kuching	2					
5	CA0004	Medical Store, Kuching	2					
6	CA0004	Medical Store, Kuching	2					
7	CA0004	Medical Store, Kuching	2					
8	CA0004	Medical Store, Kuching	2					
9	CA0004	Medical Store, Kuching	2					
10	CA0004	Medical Store, Kuching	2					
11	CA0004	Medical Store, Kuching	2					
12	CA0004	Medical Store, Kuching	2					
13	CA0004	Medical Store, Kuching	2					
14	CA0004	Medical Store, Kuching	20150113	32.000	0.001	0.012	0.016	0.540
15	CA0004	Medical Store, Kuching	20150114	26.000	0.001	0.017	0.015	0.450
16	CA0004	Medical Store, Kuching	20150115	23.000	0.001	0.020	0.019	0.540



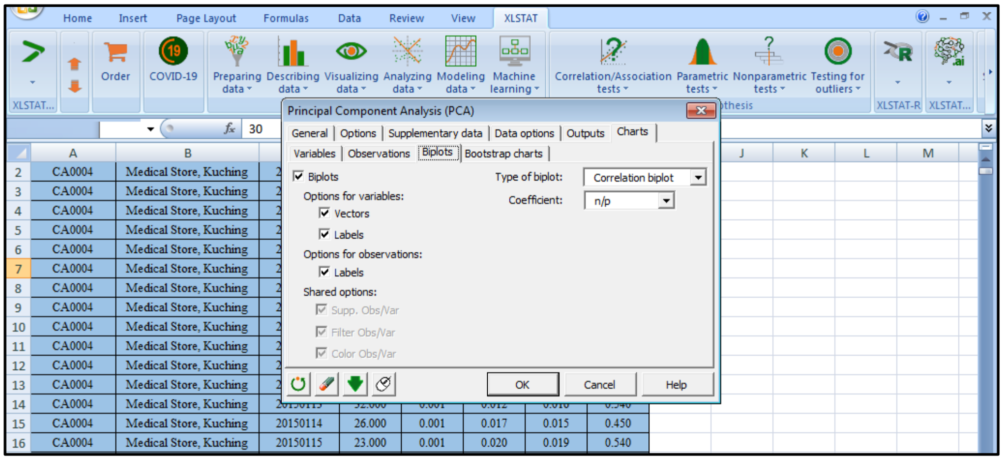
Principal Component Analysis (PCA)

General | Options | Supplementary data | Data options | Outputs | Charts

☑ Descriptive statistics
☑ Correlations
☑ Test significance
□ Bartlett's sphericity test
Significance level (%): 5
☑ Kaiser-Meyer-Olkin

☑ Eigenvalues
☑ Factor loadings
☑ Variables/Factors correlations
☑ Factor scores
☑ Contributions
☑ Squared cosines

	A	B	C	D	E	F	G	H
2	CA0004	Medical Store, Kuching						
3	CA0004	Medical Store, Kuching						
4	CA0004	Medical Store, Kuching						
5	CA0004	Medical Store, Kuching						
6	CA0004	Medical Store, Kuching						
7	CA0004	Medical Store, Kuching						
8	CA0004	Medical Store, Kuching						
9	CA0004	Medical Store, Kuching						
10	CA0004	Medical Store, Kuching						
11	CA0004	Medical Store, Kuching						
12	CA0004	Medical Store, Kuching						
13	CA0004	Medical Store, Kuching						
14	CA0004	Medical Store, Kuching						
15	CA0004	Medical Store, Kuching	20150114	26.000	0.001	0.017	0.015	0.450
16	CA0004	Medical Store, Kuching	20150115	23.000	0.001	0.020	0.019	0.540



Principal Component Analysis (PCA)

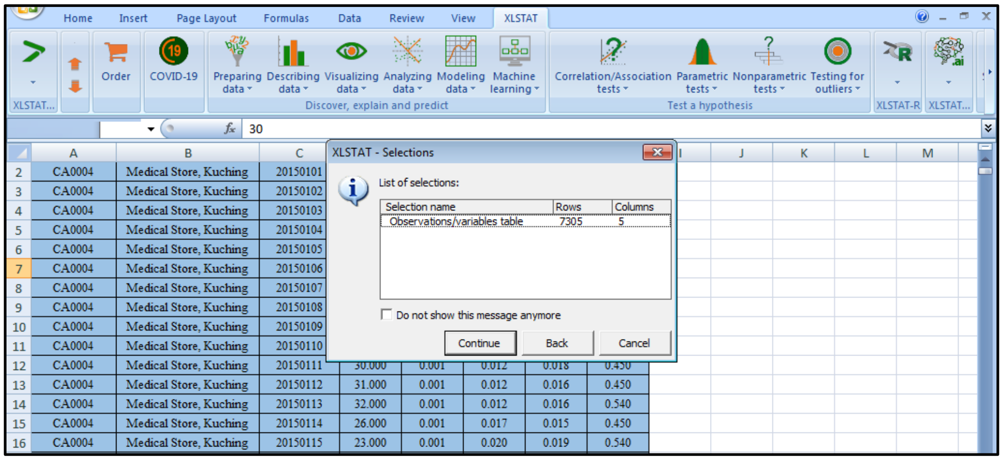
General | Options | Supplementary data | Data options | Outputs | Charts

Variables | Observations | Biplots | Bootstrap charts

☑ Biplots
Options for variables:
☑ Vectors
☑ Labels
Options for observations:
☑ Labels
Shared options:
☑ Supp. Obs/Var
☑ Filter Obs/Var
☑ Color Obs/Var

Type of biplot: Correlation biplot
Coefficient: n/p

	A	B	C	D	E	F	G	H
2	CA0004	Medical Store, Kuching						
3	CA0004	Medical Store, Kuching						
4	CA0004	Medical Store, Kuching						
5	CA0004	Medical Store, Kuching						
6	CA0004	Medical Store, Kuching						
7	CA0004	Medical Store, Kuching						
8	CA0004	Medical Store, Kuching						
9	CA0004	Medical Store, Kuching						
10	CA0004	Medical Store, Kuching						
11	CA0004	Medical Store, Kuching						
12	CA0004	Medical Store, Kuching						
13	CA0004	Medical Store, Kuching						
14	CA0004	Medical Store, Kuching						
15	CA0004	Medical Store, Kuching	20150114	26.000	0.001	0.017	0.015	0.450
16	CA0004	Medical Store, Kuching	20150115	23.000	0.001	0.020	0.019	0.540



XLSTAT - Selections

List of selections:

Selection name	Rows	Columns
Observations/variables table	7305	5

□ Do not show this message anymore

	A	B	C	D	E	F	G	H
2	CA0004	Medical Store, Kuching	20150101					
3	CA0004	Medical Store, Kuching	20150102					
4	CA0004	Medical Store, Kuching	20150103					
5	CA0004	Medical Store, Kuching	20150104					
6	CA0004	Medical Store, Kuching	20150105					
7	CA0004	Medical Store, Kuching	20150106					
8	CA0004	Medical Store, Kuching	20150107					
9	CA0004	Medical Store, Kuching	20150108					
10	CA0004	Medical Store, Kuching	20150109					
11	CA0004	Medical Store, Kuching	20150110					
12	CA0004	Medical Store, Kuching	20150111	30.000	0.001	0.012	0.018	0.450
13	CA0004	Medical Store, Kuching	20150112	31.000	0.001	0.012	0.016	0.450
14	CA0004	Medical Store, Kuching	20150113	32.000	0.001	0.012	0.016	0.540
15	CA0004	Medical Store, Kuching	20150114	26.000	0.001	0.017	0.015	0.450
16	CA0004	Medical Store, Kuching	20150115	23.000	0.001	0.020	0.019	0.540

Summary statistics:

Variable	Observa tions	Obs. with	Obs. without	Minimum	Maximum	Mean	Std. deviation
PM10 (µg/m3)	7305	0	7305	0.000	718.930	41.414	27.441
SO2 (ppm)	7305	0	7305	0.000	0.065	0.002	0.006
NO2 (ppm)	7305	0	7305	0.000	0.053	0.012	0.006
O3 (ppm)	7305	0	7305	0.000	0.075	0.027	0.010
CO (ppm)	7305	0	7305	0.000	3.150	0.747	0.343

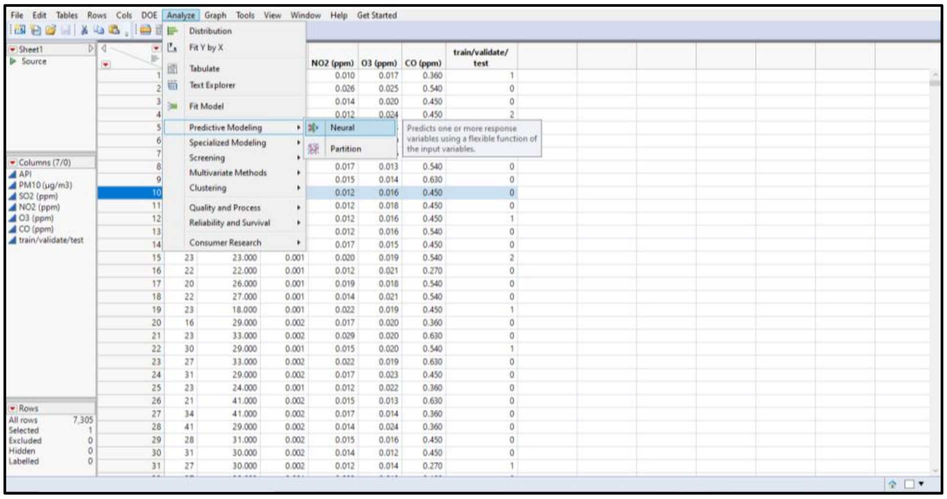
Correlation matrix (Pearson (n)):

Variables	PM10 (µg/m3)	SO2 (ppm)	NO2 (ppm)	O3 (ppm)	CO (ppm)
PM10 (µg/m3)	1	0.038	0.143	0.175	0.297
SO2 (ppm)	0.038	1	0.559	0.231	0.205
NO2 (ppm)	0.143	0.559	1	0.291	0.332
O3 (ppm)	0.175	0.231	0.291	1	0.060
CO (ppm)	0.297	0.205	0.332	0.060	1

Contribution of the variables (%):

	F1	F2	F3	F4	F5
PM10 (µg/m3)	8.689	52.674	7.580	31.049	0.008
SO2 (ppm)	26.461	18.461	3.826	13.992	37.259
NO2 (ppm)	33.972	5.483	2.382	0.730	57.433
O3 (ppm)	13.670	1.774	65.351	17.925	1.280
CO (ppm)	17.208	21.607	20.862	36.304	4.019

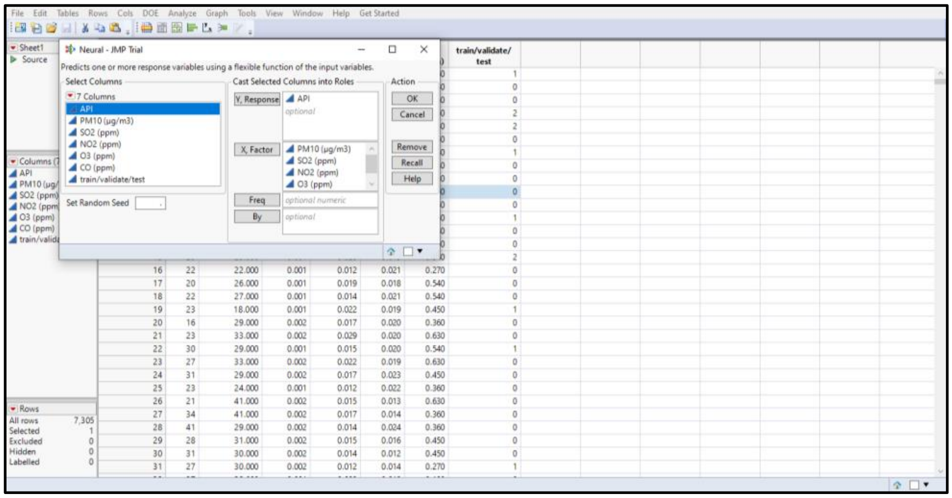
APPENDIX B: OUTPUT SOFTWARE FROM JMP15 FOR MODEL A AND MODEL B



Predicts one or more response variables using a flexible function of the input variables.

All rows	7,305
Selected	1
Excluded	0
Hidden	0
Labelled	0

				NO2 (ppm)	O3 (ppm)	CO (ppm)	train/validate/ test
1				0.010	0.017	0.360	1
2				0.026	0.025	0.540	0
3				0.014	0.020	0.450	0
4				0.012	0.024	0.450	2
5							
6							
7							
8				0.017	0.013	0.540	0
9				0.015	0.014	0.630	0
10				0.012	0.016	0.450	0
11				0.012	0.018	0.450	0
12				0.012	0.016	0.450	1
13				0.012	0.016	0.540	0
14				0.017	0.015	0.450	0
15	23	23.000	0.001	0.020	0.019	0.540	2
16	22	22.000	0.001	0.012	0.021	0.270	0
17	20	26.000	0.001	0.019	0.018	0.540	0
18	22	27.000	0.001	0.014	0.021	0.540	0
19	23	18.000	0.001	0.022	0.019	0.450	1
20	16	29.000	0.002	0.017	0.020	0.360	0
21	23	33.000	0.002	0.029	0.020	0.630	0
22	30	29.000	0.001	0.015	0.020	0.540	1
23	27	33.000	0.002	0.022	0.019	0.630	0
24	31	29.000	0.002	0.017	0.023	0.450	0
25	23	24.000	0.001	0.012	0.022	0.360	0
26	21	41.000	0.002	0.015	0.013	0.630	0
27	34	41.000	0.002	0.017	0.014	0.360	0
28	41	29.000	0.002	0.014	0.024	0.360	0
29	28	31.000	0.002	0.015	0.016	0.450	0
30	31	30.000	0.002	0.014	0.012	0.450	0
31	27	30.000	0.002	0.012	0.014	0.270	1

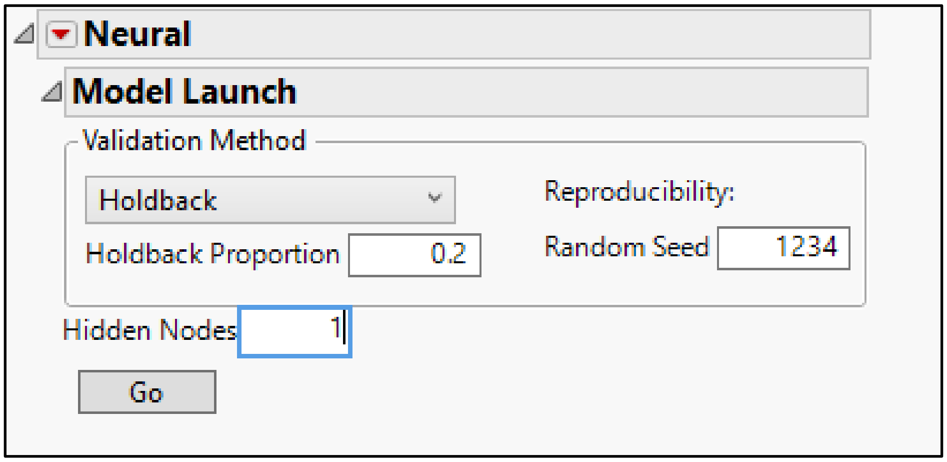


Predicts one or more response variables using a flexible function of the input variables.

Y, Response: API

X, Factor: PM10 (µg/m3), SO2 (ppm), NO2 (ppm), O3 (ppm)

Set Random Seed



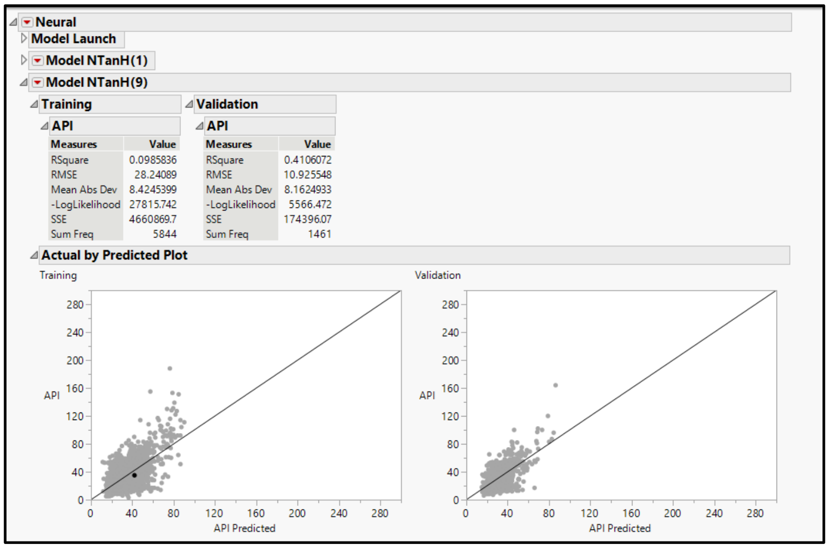
Holdback

Holdback Proportion 0.2

Reproducibility:

Random Seed 1234

Hidden Nodes 1



Neural

Model Launch

Model NTanH(1)

Model NTanH(9)

Training

API

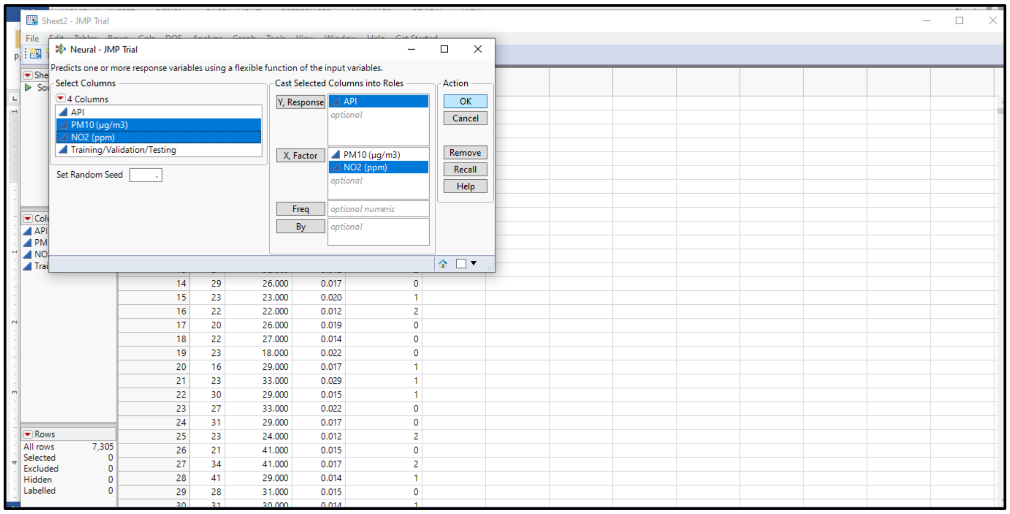
Measures	Value
RSquare	0.0985836
RMSE	28.24089
Mean Abs Dev	8.4245399
-LogLikelihood	27815.742
SSE	4660869.7
Sum Freq	5844

Validation

API

Measures	Value
RSquare	0.4106072
RMSE	10.925548
Mean Abs Dev	8.1624933
-LogLikelihood	5566.472
SSE	174396.07
Sum Freq	1461

Actual by Predicted Plot



Sheet2 - JMP Trial

Neural - JMP Trial

Predicts one or more response variables using a flexible function of the input variables.

Select Columns

4 Columns

- API
- PM10 (µg/m3)
- NO2 (ppm)
- Training/Validation/Testing

Set Random Seed

Cast Selected Columns into Roles

Y, Response: API

X, Factor: PM10 (µg/m3), NO2 (ppm)

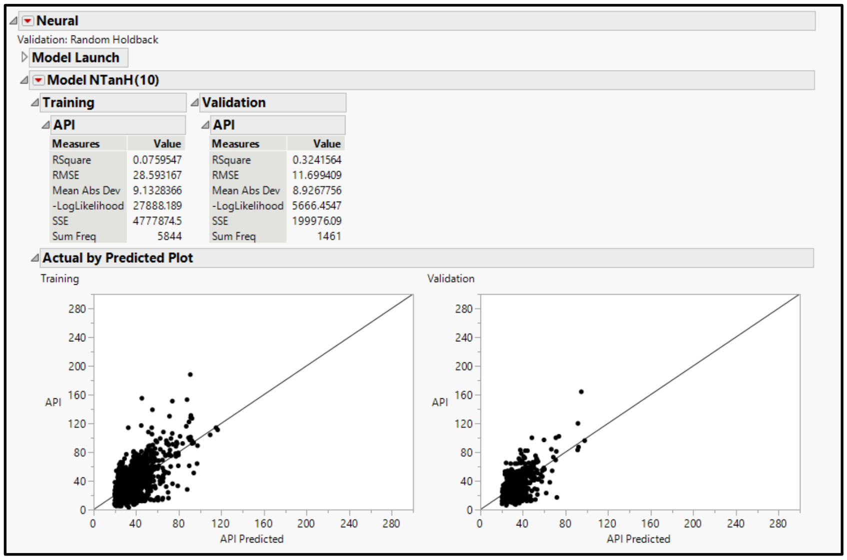
Freq: optional numeric

By: optional

Rows

All rows	7,305
Selected	0
Excluded	0
Hidden	0
Labelled	0

14	29	26.000	0.017	0
15	23	23.000	0.020	1
16	22	22.000	0.012	2
17	20	26.000	0.019	0
18	22	27.000	0.014	0
19	23	18.000	0.022	0
20	16	29.000	0.017	1
21	23	33.000	0.029	1
22	30	29.000	0.015	1
23	27	33.000	0.022	0
24	31	29.000	0.017	0
25	23	24.000	0.012	2
26	21	41.000	0.015	0
27	34	41.000	0.017	2
28	41	29.000	0.014	1
29	28	31.000	0.015	0
30	31	30.000	0.014	1



Neural

Validation: Random Holdback

Model Launch

Model NTanH(10)

Training

API

Measures	Value
RSquare	0.0759547
RMSE	28.593167
Mean Abs Dev	9.1328366
-LogLikelihood	27888.189
SSE	4777874.5
Sum Freq	5844

Validation

API

Measures	Value
RSquare	0.3241564
RMSE	11.699409
Mean Abs Dev	8.9267756
-LogLikelihood	5666.4547
SSE	199976.09
Sum Freq	1461

Actual by Predicted Plot