PREDICTION OF AIR POLLUTION INDEX (API) AND WATER QUALITY INDEX (WQI) USING SUPPORT VECTOR MACHINE (SVM)

LEONG WEI CONG

UNIVERSITI SAINS MALAYSIA

2017

PREDICTION OF AIR POLLUTION INDEX (API) AND WATER QUALITY INDEX (WQI) USING SUPPORT VECTOR MACHINE (SVM)

By

LEONG WEI CONG

Thesis submitted in partial fulfilment of the requirement

for the degree of Bachelor of Chemical Engineering

JUNE 2017

ACKNOWLEDGEMENT

First and foremost, I would like to convey my sincere gratitude to my supervisor, Associate Professor IR. Dr. Zainal Ahmad for his precious encouragement, guidance and generous support throughout this work.

I would also extend my gratitude towards all my colleagues for their kindness cooperation and helping hands in guiding me carrying out my work. They are willing to sacrifice their time in guiding and helping me throughout the work besides sharing their valuable knowledge.

Apart from that, I would also like to thank all SCE staffs for their kindness cooperation and helping hands. Indeed their willingness in sharing ideas, knowledge and skills are deeply appreciated.

Once again, I would like to thank all the people, including those whom I might have missed out and my friends who have helped me directly or indirectly. Their contributions are very much appreciated. Thank you very much.

Leong Wei Cong

June 2017

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS

SYMBOL		UNIT
C	Penalty parameter in SVM	-
ε	Intensive loss function in SVM	-
γ	Regularization parameter in LS-SVM	-
σ^2	Kernel parameter corresponding to kernel type in LS-SVM	-

LIST OF ABBREVIATIONS

ANN	Artificial Neural Networks
API	Air Pollution Index
BOD	Biochemical Oxygen Demand
CO	Carbon Monoxide
COD	Chemical Oxygen Demand
DO	Dissolved Oxygen
DOE	Department Of Environment
INWQS	Interim National Water Quality Standards For Malaysia
IQR	Interquartile Range Rule
LS-SVM	Least Square Support Vector Machine
MAQI	Malaysian Air Quality Index
MSSE	Mean Of Sum Squares Error
NH_3N	Ammoniacal Nitrogen
NO_2	Nitrogen Dioxide
NO_x	Nitrogen Oxides
O_3	Ozone
PM	Particulate Matter
PSI	Pollutant Standard Index
R^2	Coefficient Of Determination
RBF	Radial Basis Function
RMG	Recommended Malaysian Air Quality Guidelines
SO_2	Sulphur Dioxide
SRM	Structural Risk Minimization

SS	Suspended Solids
SSE	Sum Squares Error
SVM	Support Vector Machine
US EPA	United States Environmental Protection Agency
VOC	Volatile Organic Compounds
WQI	Water Quality Index

RAMALAN INDEKS PENCEMARAN UDARA (API) DAN INDEKS KUALITI AIR (WQI) MENGGUNAKAN MESIN SOKONGAN VECTOR (SVM)

ABSTRAK

Kajian ini adalah untuk menjana model mesin sokongan vector (SVM) yang sesuai untuk meramalkan indeks pencemaran udara (API) dan indeks kualiti air (WQI). Pengiraan semasa API dan WQI adalah rumit dan memakan masa. Dengan model SVM ini, API dan WQI dapat diramal dengan serta-merta dengan menggunakan peramal yang sama yang digunakan dalam pengiraan. Terdapat tiga parameter utama yang mengawal prestasi model SVM, ia adalah parameter C, ε dan jenis fungsi kernel yang digunakan. Dalam kajian ini, hanya fungsi kernel sahaja yang akan disiasat, mereka adalah lelurus, fungsi radial asas (RBF) dan fungsi kernel polinomial. Keputusan model akan dianalisis dengan menggunakan ralat jumlah kuasa dua (SSE), min ralat jumlah kuasa dua (MSSE) dan pekali penentuan (R^2). Selepas jenis fungsi kernel yang terbaik dipilih untuk model API dan WQI SVM, jenis-jenis fungsi kernel itu akan digunakan lagi untuk melatih model LS-SVM untuk membandingkan ketepatan antara model SVM dan LS-SVM. Ia telah mendapati bahawa fungsi kernel yang terbaik untuk model API SVM adalah fungsi kernel RBF, R^2 ia adalah 0.9843 manakala bagi model WQI SVM adalah fungsi kernel polynomial dan R^2 ia adalah 0.8796. Selain itu, dalam kajian in, didapati bahawa model WQI LS-SVM yang dilatih dengan peramal yang betul telah mempunyai ketepatan yang lebih tinggi dan ia punya R^2 adalah 0.9227 berbanding dengan model WQI SVM yang dilatih dengan semua peramal yang sedia ada dan ia punya R^2 adalah 0.9184. Malangnya, API LS-SVM model tidak dapat dilatih kerana jumlah set data yang besar adalah sukar untuk diproses oleh komputer yang sedia ada.

PREDICTION OF AIR POLLUTION INDEX (API) AND WATER QUALITY INDEX (WQI) USING SUPPORT VECTOR MACHINE (SVM)

ABSTRACT

This study was about to generate a suitable support vector machine (SVM) model to predict the air pollution index (API) and water quality index (WQI). The current calculations of API and WQI were complex and time consuming. With the SVM model, the API and WQI can be predicted immediately by using the same predictors used in the calculation. There were three main parameters that control the performance of the SVM model, they were parameter C, ε and the type of kernel function used. In this study, only the type kernel function was investigated, they were linear, radial basis function (RBF) and polynomial kernel function. The results of the model were then analysed by using sum squares error (SSE), mean of sum squares error (MSSE) and coefficient of determination (R^2). After the best type of kernel function was chosen for API and WQI SVM models, the types of kernel function were further utilised to train the least square support vector machine (LS-SVM) models to compare the accuracy between SVM and LS-SVM models. It was found that the best kernel function for API SVM model was RBF kernel function with R^2 of 0.9843 while for WQI SVM model was polynomial kernel function with R^2 of 0.8796. Moreover, it was found that WQI LS-SVM model that trained with correct predictors was having higher accuracy with R^2 of 0.9227 compared with WQI SVM model that trained with all the predictors with R^2 of 0.9184. Unfortunately, API LS-SVM model was not be able to train since the large amount of set of data was difficult to be processed by the computer available.

CHAPTER ONE

INTRODUCTION

1.1 Research Background

In our daily life, an average adult male consumed about 13.5 kg of air and 2 kg of water each day (Nieto et al., 2013;Sánchez et al., 2011). Therefore, the cleanliness of air and water were very important to humans to have a healthy life. However, we were not be able to determine the quality of water and air by just using naked eye. Thus, in order to understand the cleanliness and quality of air and water, some tools were developed to evaluate the quality of air and water. These tools are air pollution index and water quality index.

Air pollution index (API) was initially established in response to health issues related to the deteriorating air quality. API was used to report on the state of air pollutants, which was widely accepted as important determinants of adverse health effects (Hajek and Olej, 2015). Air pollution was key factor of the environmental problems in metropolitan cities and it was clear that there are many air pollution indicators affecting human health. If the concentration levels of these indicators exceeded the air quality guidelines, short term and chronic human health problems may occur (Sánchez et al., 2011;Nieto et al., 2013).

Next, water quality index (WQI) was a single number, which used a set of physicochemical water parameters to express the quality of water at a certain place and time. Water quality can be used to assess the water properties in reference to human health and natural quality effects. The poor quality of surface water was a serious problem in the world which threatens human health, ecosystems, and plant or animal life (Mohammadpour et al., 2015).

Thus, monitoring and predicting of air and water qualities were popular and important topic today due to the health impact caused by air and water. However, accurate and easier models for prediction were needed because such models would allow forecasting compliance and non-compliance in both short-term and long-term aspects (Wang et al., 2008). Therefore, in this research paper, support vector machine (SVM) in Matlab was used to predict the API and WQI based on the data collected in Perak and Penang.

1.2 Problem Statement

Air and water were indeed the most important things in our life. Humans needed oxygen in the air every second to continue the life while 50 to 75 percent of human bodies was actually water. Besides humans, plants also needed air to carry out photosynthesis. While, water, a prime natural resource and precious national asset, formed the chief constituent of ecosystem. Besides the need of water for drinking, water played a vital role in various sectors of economy such as agriculture, livestock production, forestry, industrial activities, hydropower generation, fisheries and other creative activities. Thus, clean air and water were very important for us, living organisms on the earth, to have healthy life and at the same time, continue our life.

API and WQI were the numbers that indicating the quality of the air and water. API needed to be kept as low as possible which also means cleaner air while WQI needed to be kept as high as possible. However, currently these indices which announced by the government were mostly calculated in a complex method which involves sub-index calculation (Department of Environment, 2005). In addition, for API, the value was represent by the highest sub-API value only which neglected the effect of others pollutants (Ibrahim, 2000).

Therefore, there was a strong need in predicting these indices in an easier and more accurate way. Many methods had been used in predicting for both API and WQI. It was found that there were lack of standardize method worldwide (Amornsamankul and Kraipeerapun, 2007;Moazami et al., 2016). Different places used different methods they found to be the best, which was higher accuracy and stability. Surprisingly, currently in Malaysia, there was no prediction method for these indices yet. All the indices shown were still coming from lengthy calculation. Thus, this study was to find a high accuracy and stability method to predict the API and WQI by using SVM.

1.3 Research Objectives

There were four objectives in this study and they were:

1) To generate a suitable predicting method for API and WQI that can be used in Perak and Penang by using SVM and least square support vector machine (LS-SVM) in Matlab.
2) To find out the best kernel function for the models for both API and WQI.
3) To study the accuracy of the prediction between SVM and LS-SVM.
4) To find the effect of using correct predictor in training the model for WQI on accuracy.

1.4 Scope of Study

In this work, the SVM and LS-SVM were used to predict the API and WQI. The SVM and LS-SVM model were built in Matlab automatically after determining the type of kernel function, which was also the target of the study. The best kernel function was found for both API and WQI models. All the other parameters that may affect the performance of the model were not investigated.

The performance of the models were analysed by using SSE, MSSE and R^2. The performance of the SVM and LS-SVM models were then compared to determine which model gave the best performance.

1.5 Organization of Thesis

The following were the contents for each chapter in this study:

Chapter 1 introduced the importance of API and WQI, problem statement, research objectives and organization of thesis.

Chapter 2 discussed the literature review of this study that included the API, WQI, SVM, prediction of API and WQI, usage of SVM in other areas.

Chapter 3 covered the case study of this work, process modelling and performance criteria.

Chapter 4 referred to the experimental results and discussions of the data obtained. Further elaboration on the accuracy on different type of kernel function, the comparison between SVM and LS-SVM were provided in this chapter.

Chapter 5 concluded all the findings obtained in this study. Recommendations were also included as well.

CHAPTER TWO

LITERATURE REVIEW

2.1 Introduction

It was very important to have a good foundation before building any houses, same theory happened here. In order to make the study later easier to be carried out, we needed to understand and some terms and methods that being used in the study.

In this chapter, we first looked into API and the parameters that contribute to this index. We understood the brief history of API and its calculation to have the API value from the raw data. Then, we also looked into the five parameters that being used in Malaysia to determine the value of API and their effects on health.

Then, after API, we continue studied on WQI which will be undergo prediction as API in this paper. More or less the same, we first looked into the brief history of WQI and its calculation. Then only we start to focus on the six parameters that contribute to the WQI value.

Finally yet importantly, we studied about the method that will be used for the prediction in the paper, which was SVM. We studied its brief advantages and the reason of its widely usage in different areas. After that, we also studied few researches that used SVM either in similar area as in this paper or in others area. At the same time, LS-SVM, which will be used to compare the result with SVM later in this paper was briefly introduce about its advantages over SVM.

2.2 Air Pollution Index

Every human beings needed the oxygen in the air to survive, and it was very important to have clean air that without or with less air pollutants. This was because air

pollutants can be dangerous to human health and thus, the understanding of air pollutant concentration was very important. In order to understand and monitor the air quality, API had been used. The purpose of API was to help people to understand the quality of air, which later provides significance guidance for public's exposure to air pollution.

The API was developed by United States Environmental Protection Agency (US EPA) for the first time in 1976 to characterize regional air quality (Lu et al., 2011). It was calculated from data of pollutants by a segmented linear function that transforms ambient concentrations onto a scale between 0 to 500. However, the API system did purpose some disadvantages and its major shortcoming was that it did not take into account the combined effects of all pollutants on human health. The value of API was only defined by one of the pollutants, the pollutant which has the highest concentration relative to its standard in a given of time (Ibrahim, 2000).

In Malaysia, the API system was only adopted in 1996 and it was the successor for Malaysian Air Quality Index (MAQI). It was adopted due to the need for regional harmonisation and for easy comparison with countries in ASEAN. Moreover, the API system of Malaysia was closely following the Pollutant Standard Index (PSI) system of the United States. Although API system in Malaysia follows closely to PSI system, it was actually working based on Recommended Malaysian Air Quality Guidelines (RMG) for air pollutants, which formulated in 1989 by Department of Environment (DOE). RMG had defined the concentration limits of selected air pollutants, which might adversely affect the health, and welfare of public.

API system normally included the major air pollutants, which might bring potential harm to human health once they reached unsafe level. While the air pollutants included in Malaysia's API system were ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2), sulphur dioxide (SO_2) and particulate matter (PM) with a diameter of less

than 10 micron (Ibrahim, 2000). Moreover, in Malaysia, air quality with API values exceeding 100 were considered likely to cause health effects to public and different API values also indicating different conditions and effects to public. The air quality in terms of human health impacts within each API categories are categorised as Table 2.1 under Malaysia's API system.

As mentioned, the API value was defined by the highest concentration pollutants and it was also the way to calculate API in Malaysia. To calculate the API for a given time period, the sub-index values or sub-API for all five air pollutants included in the API system are first calculated using sub-index equations as shown in Table 2.2 for the air quality data collected from the continuous air quality monitoring stations. Then, the highest sub-API among all the sub-APIs calculated during the particular period will be the reporting API to public. This calculation technique was shown in Figure 2.1.

Table 2.1 Category of API and its effect (Ibrahim, 2000)

API	Condition	Level of Pollution	Health Measures
0 – 50	Good	Pollution low and has no ill effects on health.	No restriction of activities for all groups of people.
51 – 100	Moderate	Moderate pollution and has no ill effects on health.	No restriction of activities for all groups of people.
101 – 200	Unhealthy	Mild aggravation of symptoms among high-risk persons, like those with heart or lung disease.	Restriction of outdoor activities for high-risk persons. General population should reduce vigorous outdoor activity.
201 – 300	Very Unhealthy	Significant aggravation of symptoms and decreased exercise tolerance in person with heart or lung disease.	Elderly and persons with known heart or lung disease should stay indoors and reduce physical activity.
More than 300	Hazardous	Severe aggravation of symptoms and endangers health.	Elderly and persons with known heart or lung disease should stay indoors and reduce physical activity. General population should reduce vigorous outdoor activity.

Table 2.2 Equations of sub-API in Malaysia (Ibrahim, 2000)

Pollutant	Value	Sub-API
CO	conc. < 9 ppm	API = conc. × 11.11111
	9 < conc. < 15	API = 100 + {[conc. – 9] × 16.66667}
	15 < conc. < 30	API = 200 + {[conc. – 15] × 6.66667}
	conc. > 30 ppm	API = 300 + {[conc. – 30] × 10}
O_3	conc. < 0.2 ppm	API = conc. × 1000
	0.2 < conc. < 0.4	API = 200 + {[conc. – 0.2] × 500}
	conc. > 0.4 ppm	API = 300 + {[conc. – 0.4] × 1000}
NO_2	conc. < 0.17 ppm	API = conc. × 588.23529
	0.17 < conc. < 0.6	API = 100 + {[conc. – 0.17] × 232.56}
	0.6 < conc. < 1.2	API = 200 + {[conc. – 0.6] × 166.667}
	conc. > 1.2 ppm	API = 300 + {[conc. – 1.2] × 250}
SO_2	conc. < 0.04 ppm	API = conc. × 2500
	0.04 < conc. < 0.3	API = 100 + {[conc. – 0.04] × 384.61}
	0.3 < conc. < 0.6	API = 200 + {[conc. – 0.3] × 333.333}
	conc. > 0.6 ppm	API = 300 + {[conc. – 0.6] × 500}
PM 10	conc. < 50 $\mu g/m^3$	API = conc.
	50 < conc. < 150	API = 50 + {[conc. – 50] × 0.5}
	150 < conc. < 350	API = 100 + {[conc. – 150] × 0.5}
	350 < conc. < 420	API = 200 + {[conc. – 350] × 1.4286}
	420 < conc. < 500	API = 300 + {[conc. – 420] × 1.25}
	conc. > 500 $\mu g/m^3$	API = 400 + [conc. – 500]

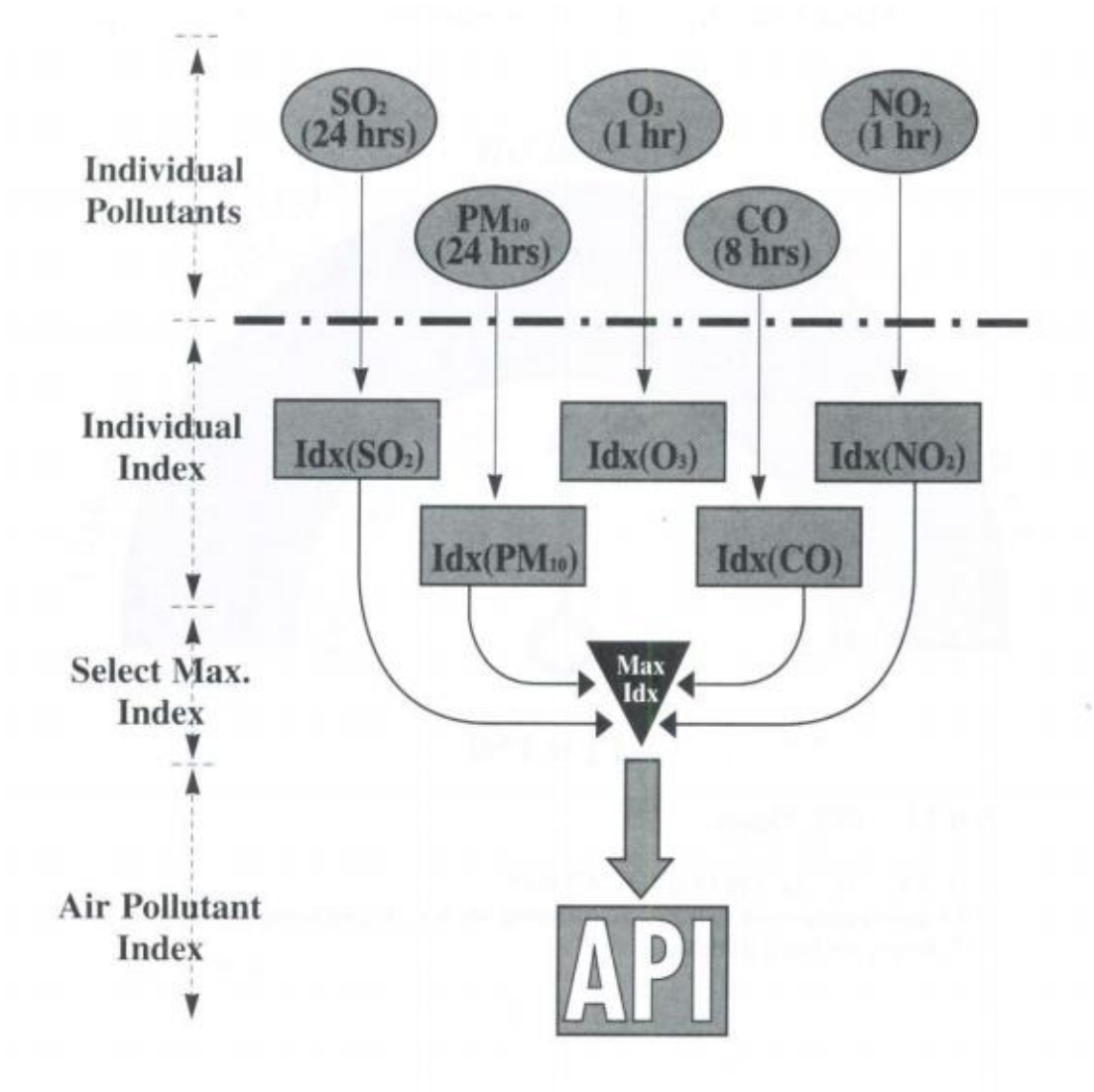

Figure 2.1 Flowchart of calculation technique for API (Ibrahim, 2000)

2.2.1 Ozone

O_3 was a gas composed of three atoms of oxygen. O_3 can occur in both Earth's upper atmosphere and at ground level. It can be good or bad, depending on where it was found. Good O_3 were called stratospheric ozone, which occurs naturally in the upper atmosphere. At there, it formed a protective layer to protect us from the sun's harmful ultraviolet rays and it had no direct effect to human or environment. However, tropospheric or so-called, ground level ozone, which was created by chemical reactions between oxides of nitrogen (NO_x) and volatile organic compounds (VOC) rather than

emitted directly into the air was called bad ozone (United States Environmental Protection Agency, 2016c). These reactions happened when pollutants emitted by human activities chemically react in the presence of sunlight. It was bad due to its harmful effect on human and environment.

Due to its photochemical origin, O_3 displays strong seasonal and diurnal patterns, with higher concentrations in summer and in the afternoon. The correlation of O_3 with other pollutants varies by season and location (World Health Organization, 2003).

There was evidence from controlled human and animal exposure studies of the potential for O_3 to cause adverse health effects. Epidemiological studies have also addressed the effects of short and long-term exposures to O_3 and provided important results (Xin et al., 2010). However, the health effects of O_3 had been less studied than those of PM and thus more research was needed, especially addressing the spatial and seasonal patterns and misclassification of individual exposure in association with health outcomes.

Besides health effects to human, O_3 also brought some effects to ecosystems and sensitive vegetation. When sufficient O_3 entered a sensitive plant, it can reduce the photosynthesis of the plant and thus slowing down the growth of the plant (United States Environmental Protection Agency, 2016c). Moreover, it also increased the risk of the plants to get disease, damage from insects and effects of other pollutants.

2.2.2 Particulate Matter

PM also called particle pollution, which was a term for a mixture of solid particles and liquid droplets found in the air. Some particles were large or dark enough to be seen by naked eye while some were so small until they can only be detected by using an electron microscope.

These particles came in many shapes and sizes and can be made up of different chemicals. Some were emitted directly from a source but most of the particles in the atmosphere were a result of complex reactions of chemicals from those pollutants like nitrogen oxide and SO_2 that emitted from human activities.

Basically, PM can be divided into 2 categories based on their size. Coarse dust particles (PM 10) were particles with diameter between 2.5 to 10 micrometers. While fine particles were particles with diameter less than 2.5 micrometers, which can only be seen with an electron microscope (United States Environmental Protection Agency, 2016d). The potential of causing health problems was directly linked to the size of the particles. The smaller the particles pose the greater problem as they can get deeper into the lungs and some may even get into the bloodstream if the size is small enough. The comparison between PM 2.5, PM 10 with human's hair was shown in 2.2.

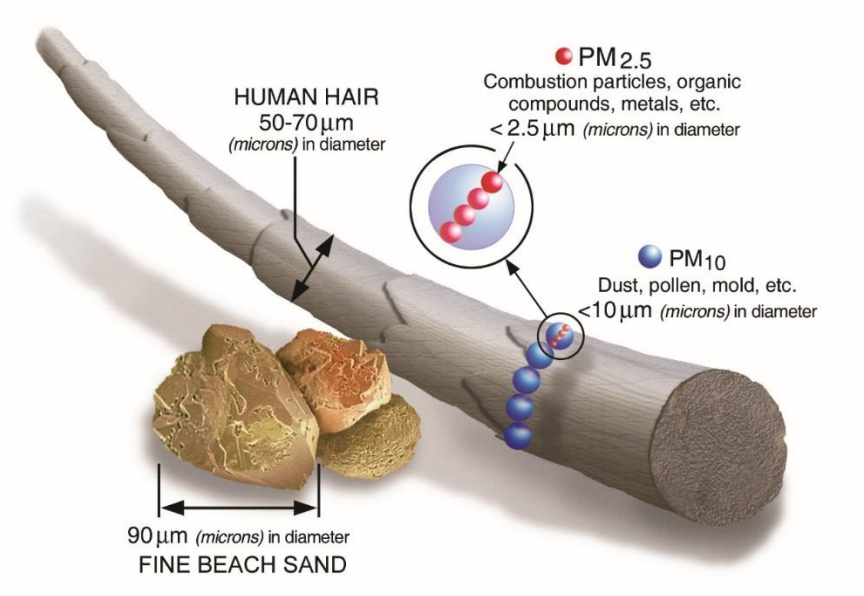

Figure 2.2 Comparison between PM with human's hair (United States Environmental Protection Agency, 2016d)

2.2.3 Carbon Monoxide

CO was a colourless and odourless gas that can be harmful when large amounts were inhaled. It was released when something was burned and the greatest sources in outdoor were cars and any vehicles or machinery that burn fossil fuels. CO was dangerous in term of it will reduce the amount of oxygen that can be transport in the blood stream to others organs like heart when breathing in CO (United States Environmental Protection Agency, 2016a). It will react with haemoglobin faster than oxygen in the lung because it had greater combining affinity. At very high levels, either indoors or outdoors, CO can cause dizziness, confusion, unconsciousness and even death.

2.2.4 Sulphur Dioxide

SO_2 was the component of greatest concern and was used as the indicator for the larger group of gaseous sulphur oxides (SO_x). This was because other gaseous SO_x concentrations were found to be much lower than SO_2. SO_2 in the atmosphere were mainly come from the burning of fossil fuels by power plants and other industrial facilities (United States Environmental Protection Agency, 2016e). SO_2 as one of the monitoring pollutants in the API system can harm the human respiratory system and make breathing difficult even for a short term exposures and this situation was even obvious for children, elderly and those who suffer from asthma. Besides, high concentration of SO_2 can harm trees and plants by damaging foliage and decreasing growth. In addition, SO_2 was also a source for PM if they react with other compounds in the atmosphere.

2.2.5 Nitrogen Dioxide

Same as SO_2, NO_2 was acting as the indicator for the larger group of nitrogen oxides (NO_x) and it major source was burning of fuels, cars and power plants. More or

less, NO_2 propose almost the same effect as SO_2. It can irritate airways in human respiratory system when high concentration was breathed. Moreover, NO_2 and other NO_x can form acid rain that harms sensitive ecosystems when interact with water, oxygen and other chemicals in atmosphere (United States Environmental Protection Agency, 2016b).

2.3 Water Quality Index

WQI was initially developed by Horton (1965). Since then, several authors had developed many different methods for the calculation of WQI such as Brown et al. (1970). The WQI was one criterion for surface water classifications based on the use of standard parameters for water characterization. It provided a comprehensive picture of the quality of water. It was a mathematical mechanism for summarizing the water quality data into simple terms that can be easily understand by public such as good or bad (Abdel-Satar et al., 2016). These terms can reflect the quality level for the measured water.

In 1974, the DOE recommended the WQI parameter for categorizing and estimating water quality. The water quality data were used to determine the water quality status as Table 2.3 and classify the rivers into classes based on WQI as Table 2.4 and Interim National Water Quality Standards for Malaysia (INWQS) every year. In addition, different country will have different parameters to be measured and used in the WQI calculation. For Malaysia, the WQI was computed based on six main parameters which were biochemical oxygen demand (BOD), chemical oxygen demand (COD), ammoniacal nitrogen (NH_3N), pH, dissolved oxygen (DO) and suspended solids (SS), while other parameters such as heavy metals and bacteria would be measured according to site requirement (Department of Environment, 2005).

However, the conventional method suggested by DOE to determine WQI requires lengthy transformations to estimate sub-indices (Department of Environment, 2005). In

addition, the sub-indices required the inclusion of different equations as shown in Table 2.5, which need lengthy effort and time to calculate the final WQI. Therefore, estimation of such a WQI was cumbersome and can lead to occasional mistakes (Mohammadpour et al., 2015).

Table 2.3 Status of water corresponding to WQI (Department of Environment, 2016)

WQI	Status
0 – 59	Polluted
60 – 80	Slightly Polluted
81 – 100	Clean

Table 2.4 Classes and its criteria of water corresponding to WQI (Department of Environment, 2016)

Class	Parameter						WQI
	BOD (mg/l)	COD (mg/l)	NH_3N (mg/l)	pH	DO (mg/l)	SS (mg/l)	
1	< 1	< 10	< 0.1	> 7	> 7	< 25	> 92.7
2	1 – 3	10 – 25	0.1 – 0.3	6 – 7	5 – 7	25 – 50	76.5 – 92.7
3	3 – 6	25 – 50	0.3 – 0.9	5 – 6	3 – 5	50 – 150	51.9 – 76.5
4	6 – 12	50 – 100	0.9 – 2.7	< 5	1 – 3	150 – 300	31 – 51.9
5	> 12	> 100	> 2.7	< 5	< 1	> 300	< 31

Table 2.5 Equations for sub-indices for WQI in Malaysia (Department of Environment, 2005)

Parameter	Value	Sub-index
DO (% saturation)	$X \leq 8$	$SI = 0$
	$8 < X < 92$	$SI = -0.395 + 0.03X^2 - 0.0002X^3$
	$X \geq 92$	$SI = 100$
BOD	$X \leq 5$	$SI = 100.4 - 4.23X$
	$X > 5$	$SI = (108e^{-0.055X}) - 0.1X$
COD	$X \leq 20$	$SI = 99.1 - 1.33X$
	$X > 20$	$SI = (103e^{-0.0157X}) - 0.04X$
NH_3N	$X \leq 0.3$	$SI = 100.5 - 105X$
	$0.3 < X < 4$	$SI = (94e^{-0.573X}) - 5\|X-2\|$
	$X \geq 4$	$SI = 0$
SS	$X \leq 100$	$SI = (97.5e^{-0.00676X}) + 0.05X$
	$100 < X < 1000$	$SI = (71e^{-0.0016X}) + 0.015$
	$X \geq 1000$	$SI = 0$
pH	$X < 5.5$	$SI = 17.2 - 17.2X + 5.02X^2$
	$5.5 \leq X < 7$	$SI = -242 + 95.5X - 6.67X^2$
	$7 \leq X < 8.75$	$SI = -181 + 82.4X - 6.05X^2$
	$X \geq 92$	$SI = 536 - 77X + 2.76X^2$

2.3.1 Dissolved Oxygen

DO were oxygen that dissolves in water by a purely physical process, proportional to the partial pressure in the gas in the contact with the water. It was dependent on the temperature and the concentration of dissolved salts, notably chlorides. Table 2.6 shows

the level of DO in the water corresponding to the temperature based on Henry's Law (Brown and Caldwell, 2001).

Table 2.6 Level of DO in water corresponding to temperature (Brown and Caldwell, 2001)

Temperature (°C)	DO (mg/l)
0	14.6
5	12.75
10	11.27
15	10.07
20	9.07
25	8.24
30	7.54

2.3.2 Biochemical Oxygen Demand

BOD was the amount of DO required by microorganisms that mainly bacteria, for the oxidation of organic material in a waste under aerobic conditions. The BOD test was a bioassay technique involving the measurement of oxygen consumed by the bacteria while stabilizing the organic matter in the waste as they would normally do in nature but under normal laboratory conditions. It normally conducted at 20°C for a 5 days period (Kunz, 2010).

2.3.3 Chemical Oxygen Demand

COD does not differentiate between biologically available and inert organic matter. It measures the total quantity of oxygen required to oxidize all organic material

into carbon dioxide and water (Brown and Caldwell, 2001). Thus, value of COD will always greater than value of BOD but at the time, COD can be measured within few hours while BOD measurement required 5 days.

2.3.4 Ammoniacal Nitrogen

NH_3N was one form of nitrogen and it is soluble in water and can end up in ground water and drinking water. It was an essential nutrient for living organisms but too much of it can be toxic (Rožić et al., 2000). The presence of excess nitrogen can cause serious distortions of the natural nutrient cycle between the living world, the soil, water and atmosphere. The most obvious consequences of excess nitrogen is eutrophication.

2.3.5 pH

pH was a measure of the amount of free hydrogen ions in water and it can be expressed as the negative logarithm of the molar concentration of hydrogen ions. Acidity of water increased as the value of pH decreased while the alkalinity increased as value of pH increased. The pH value of water will affect the solubility of many toxic and nutritive chemicals.

2.3.6 Suspended Solids

SS were those solid that suspended in the water that cannot pass through a filter. SS can be coming from many sources and sizes and it can be categorised as organic and inorganic. As the level of SS increase, a water body began to lose its ability to support a diversity of aquatic life. This was because SS absorb heat from sunlight which later increased the temperature of water and thus decrease the DO level as we know the warmer the water, the lesser the DO in the water (Rožić et al., 2000).

2.4 Support Vector Machine

Vapnik (1995) developed the theoretical foundation of SVM and it can be act as an alternative method to artificial neural networks (ANN). It had been successfully employed to solve the problems related to engineering and can provide an effective approach to improve prediction performance and achieve a global optimization solution simultaneously. SVM can be applied in many machine-learning applications such as solving classification, regression and time-series prediction problems in an efficient and stable way (Wang et al., 2008). It also purposed has high ability for generalization and was less prone to overfitting. In addition, it simultaneously minimized the estimation of error and model dimensions (Mohammadpour et al., 2015).

The SVM lead to a unique and global solution because of its formulation, which employed a structural risk minimization (SRM) principle as opposed to an empirical risk minimization principle, employed by conventional neural networks (Khan and Coulibaly, 2006). It was an approach that minimized the upper bound risk functional related to the generalization performance. Due to this theoretical basis, SVM had a greater ability to generalize compared to traditional neural network approaches. The practical applications, especially in the environmental prediction domain, suggested that SVM was effective and can produce more accurate prediction results than artificial neural networks models (Wang et al., 2008).

Overall speaking, the idea of SVM was select a numbers of data from the dataset, which called support vectors that can define a hyperplane separating the two classes of observations. The SVM will couple with a non-linear Kernel mapping procedure when the problem was not linearly separable, projecting the data points to a higher dimensional space, called feature space, where the classes were linearly separable as shown in Figure 2.3 (Sotomayor-Olmedo et al., 2013).

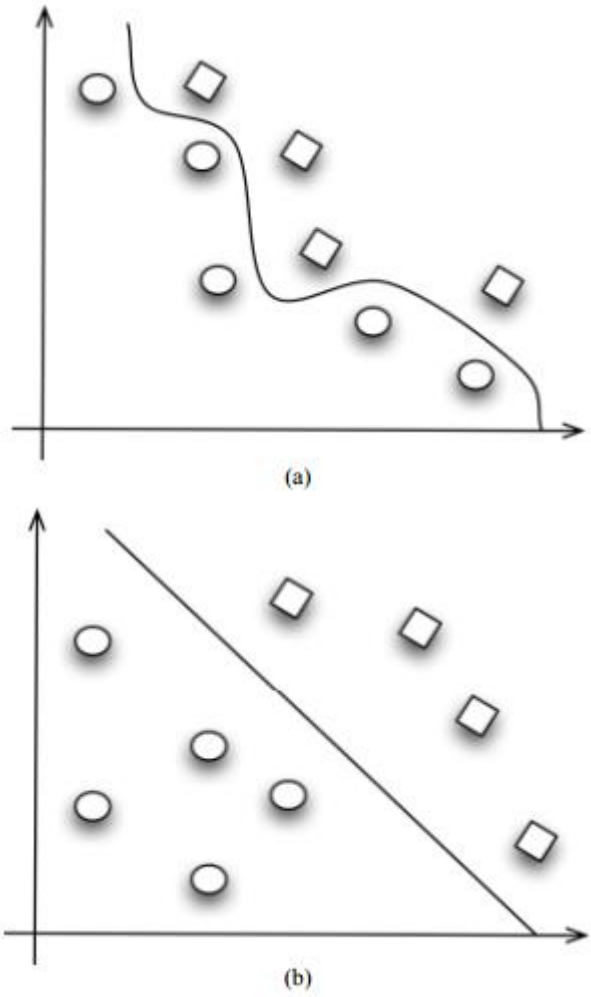

Figure 2.3 Feature map can simplify the classification a regression tasks. (a) Input space; (b) Feature space (Sotomayor-Olmedo et al., 2013)

The SVM will use suitable kernel function to map the original data into a high dimensional feature space where a maximal separating plane was constructed. In order to separate the data, two parallel hyperplanes can be developed on each side of the separating plane. Then SVM will simultaneously maximize the geometric margin and minimize the empirical classification error (Singh et al., 2011). Then, with the introduction of ε-intensive loss function, the SVM was extended to solve the regression problems (Pan et al., 2008).

In order to understand more on this technique, a brief discussion was as shown as below. For a set of training data (x_i,y_i), SVM was used to find out the function f(x) with high deviation (ε) from targets (y_i) and it should be as flat as possible at the same time. If f(x) was introduced as a linear discriminant function, SVM was then can be presented as (Smola and Schölkopf, 2004):

$$f(x) = (w, x) + b \tag{2.1}$$

where w was the weight vector ($w \epsilon R^n$), and b was the bias. The f(x) function was flatted by minimizing the values the w. Using a convex optimization problem, minimization can be expressed as (Mohammadpour et al., 2015):

$$\begin{cases} Minimize\ \frac{1}{2}\|w\|^2 \\ subject\ to \begin{cases} y_i - (w, x_i) - b \leq \varepsilon \\ (w, x_i) + b - y_i \leq \varepsilon \end{cases} \end{cases} \tag{2.2}$$

The last equation in some cases with more errors can introduced with slack variables ε_i, then minimization formula changes as follows:

$$\begin{cases} Minimize\ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*)\ subject\ to \begin{cases} y_i - (w, x_i) - b \leq \varepsilon + \xi_i \\ (w, x_i) + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases} \end{cases} \tag{2.3}$$

The C was a penalty parameter, and it should be defined by the user. This parameter determined the trade-off between the tolerable amount larger than ε and flatness of f(x). The Lagrangian form of minimization formula can be express as:

$$L = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) - \sum_{i=1}^{m}(\eta_i\xi_i + \eta_i^*\xi_i^*) - \sum_{i=1}^{m}\alpha_i(\varepsilon_i + \xi_i - y_i + (w, x_i) +$$

$$b) - \sum_{i=1}^{m}\alpha_i^*(\varepsilon_i + \xi_i - y_i + (w, x_i) + b); \begin{cases} \alpha_i, \eta_i \geq 0 \\ \alpha_i^*, \eta_i^* \geq 0 \end{cases} \tag{2.4}$$

where are α_i, η_i, α_i*, and η_i* are Lagrangian parameters. The saddle points of Equation 2.4 can be estimated as:

$$\frac{\partial L}{\partial b} = 0\ then\ \sum_{i=1}^{m}(\alpha_i^* - \alpha_i) = 0 \tag{2.5}$$

$$\frac{\partial L}{\partial w} = 0\ then\ w - \sum_{i=1}^{m}(\alpha_i^* - \alpha_i)x_i = 0 \tag{2.6}$$

$$\frac{\partial L}{\partial \xi_i^*} = 0 \; then \; C - \alpha_i^* - \eta_i^* = 0 \tag{2.7}$$

Dual maximization problem was determined by substituting Equations 2.5 to 2.7 into Equation 2.4 as:

$$Maximize \left\{ -, \frac{1}{2}, \sum_{i=1}^{m} (\alpha_i^* - \alpha_i)(\alpha_j - \alpha_j^*)(x_i, x_j), -, \varepsilon, \sum_{i=1}^{m} \left(\alpha_j + \alpha_j^*\right) + \sum_{i=1}^{m} y_i \left(\alpha_j - \alpha_j^*\right) subject \; to \; \sum_{i=1}^{m} (\alpha_j - \alpha_j^*) = 0 \; and \; \alpha_i, \alpha_i^* \epsilon [0, C] \right. \tag{2.8}$$

Finally, the SVM function can be expressed as:

$$f(x) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*)\left(x_i, x_j\right) + b \tag{2.9}$$

The kernel function was used to solve the nonlinear problem in the support vector regression. This function maps the data into higher dimension feature space (Vapnik, 1995). The support vector regression problem in the feature space was expressed using $K(x_i, x_j)$ instead of (x_i, x_j), then the SVM can be written as:

$$f(x) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) K\left(x_i, x_j\right) + b \tag{2.10}$$

There were four possible choices for the kernel function, they were linear, polynomial, sigmoid, and radial Gaussian. Combination of several parameters, like the type of kernel function, penalty parameter, C, and ε-insensitive loss function will determine the performance of SVM. In addition, the selection of kernel function was depending on the distribution of the data and generally can be selected through the trial and error approach (Widodo and Yang, 2007).

The C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the training error. For a low value of C, insufficient fitting will be placed on the training data, while the algorithm will overfit for too large of C (Wang et al., 2007). A well-performing and robust regression model is dependent on a proper choice of C in combination with ε (Üstün et al., 2005).

2.4.1 Least Square Support Vector Machine

The standard SVM was solved using complicated quadratic programming methods, which were often time consuming and difficult to implement adaptively which were also its major drawback. Least squares support vector machine (LS-SVM), as a modification of SVM, adopted the least squares linear system as its loss function and therefore solves a set of linear equations. At the same time, LS-SVM also had good convergence and high precision (Haifeng and Dejin, 2005). In order to obtain a high-level performance with LS-SVM models, some parameters such as regularization parameter γ and the kernel parameter corresponding to the kernel type, such as σ^2 should be tuned. These parameters can be determined using trial and error method (Baghban et al., 2016).

2.5 Prediction of Air Pollution Index And Water Quality Index

At present, monitoring and forecasting air pollution and water quality trends involve using a variety of approaches. Among all of the approaches, computational intelligence techniques like artificial neural networks, genetic algorithms, SVM, etc. were paid more and more attention in environmental time-series prediction researches because they can model non-linear systems well and are robust for the noise data, and so they can produce more accurate results (Wang et al., 2008).

In a study of prediction WQI in constructed wetlands, SVM and two others method were used and the results had shown that SVM technique was able to successfully predict WQI with high accuracy. The high value of R^2 with 0.9984 and low mean absolute error of 0.0052 indicated that the SVM model provided better prediction compared to the others two methods (Mohammadpour et al., 2015). Moreover the study highlighted that SVM can be successfully used as valuable methods for the prediction of water quality in