AUGMENTED SENSORS FOR PARTICULATE MATTER CONCENTRATION PREDICTION USING SUPERVISED LEARNING MODELS

Miss Chadaphim Photphanloet

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Applied Mathematics and Computational Science
Department of Mathematics and Computer Science
FACULTY OF SCIENCE
Chulalongkorn University
Academic Year 2019

Copyright of Chulalongkorn University

ตัวรับรู้เสริมสำหรับการทำนายความเข้มข้นละอองธุลีโดยใช้ตัวแบบการเรียนรู้แบบมีผู้สอน

น.ส.ชฎาพิมพ์ พตพันธ์เลิศ

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาคณิตศาสตร์ประยุกต์และวิทยาการคณนา ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

Thesis Title	AUGMENTED SENSORS FOR PARTICULATE MATTER CONCENTRATION PREDICTION USING SUPERVISED LEARNING MODELS
By	Miss Chadaphim Photphanloet
Field of Study	Applied Mathematics and Computational Science
Thesis Advisor	Associate Professor RAJALIDA LIPIKORN, Ph.D.

Accepted by the FACULTY OF SCIENCE, Chulalongkorn University in Partial Fulfillment of the Requirement for the Doctor of Philosophy

.... Dean of the FACULTY OF SCIENCE
(Professor POLKIT SANGVANICH, Ph.D.)

DISSERTATION COMMITTEE

.... Chairman
(Associate Professor PHANTIPA THIPWIWATPOTJANA, Ph.D.)

.... Thesis Advisor
(Associate Professor RAJALIDA LIPIKORN, Ph.D.)

.... Examiner
(Assistant Professor KRUNG SINAPIROMSARAN, Ph.D.)

.... Examiner
(Assistant Professor KITIPORN PLAIMAS, Ph.D.)

.... External Examiner
(Associate Professor Panjai Tantatsanawong, Ph.D.)

ชฎาพิมพ์ พตพันธ์เลิศ : ตัวรับรู้เสริมสำหรับการทำนายความเข้มข้นละอองธุลีโดยใช้ตัวแบบการเรียนรู้แบบมีผู้สอน. (AUGMENTED SENSORS FOR PARTICULATE MATTER CONCENTRATION PREDICTION USING SUPERVISED LEARNING MODELS) อ.ที่ปรึกษาหลัก : รศ. ดร.รัชลิดา ลิปิกรณ์

แบบจำลองการทำนายความเข้มข้นของละอองธุลีได้รับการวิจัย พัฒนาและนำไปประยุกต์ใช้กับข้อมูลภูมิประเทศต่างๆทั่วโลก ลักษณะของภูมิประเทศที่แตกต่างกันทำให้แบบจำลองมีความเหมาะสมสำหรับภูมิประเทศแต่ละแห่งแบบเฉพาะเจาะจง วิทยานิพนธ์ฉบับนี้นำเสนอวิธีการทำนายละอองธุลีที่มีเส้นผ่านศูนย์กลางเล็กกว่า 10 ไมครอนในจังหวัดน่านของประเทศไทย ซึ่งได้รวมวิธีการเลือกคุณสมบัติ แบบจำลองการเรียนรู้แบบมีผู้สอนและอัลกอริทึมการค้นหาเชิงลึก ซึ่งแตกต่างจากแบบจำลองการเรียนรู้แบบมีผู้สอนแบบดั้งเดิม ทั้งนี้วิธีการที่นำเสนอนั้นสามารถรับข้อมูลหลายมิติที่ประกอบด้วยข้อมูลความเข้มข้นของละอองธุลี สารมลพิษในอากาศ คุณภาพของอากาศ ซึ่งสารมลพิษในอากาศและคุณภาพของอากาศเป็นปัจจัยที่มีอิทธิพลต่อการทำนายความเข้มข้นของละอองธุลี ผลการทดลองแสดงให้เห็นว่าวิธีการที่นำเสนอนั้นมีประสิทธิภาพดีกว่าวิธีการอื่น ๆ เมื่อคาดการณ์ล่วงหน้าหนึ่งชั่วโมงโดยไม่จำเป็นต้องใช้ข้อมูลทิศทางกระแสลมและความเร็วกระแสลม วิธีการที่นำเสนอได้รับการพัฒนาโดยมีกรอบงานทั่วไปและสามารถนำไปประยุกต์ใช้กับการทำนายความเข้มข้นของละอองธุลีในจังหวัดน่าน

สาขาวิชา	คณิตศาสตร์ประยุกต์และวิทยาการคณนา	ลายมือชื่อนิสิต ....
ปีการศึกษา	2562	ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

5972809523 : MAJOR APPLIED MATHEMATICS AND COMPUTATIONAL SCIENCE

KEYWORD: Particulate matter concentration prediction, Modified depth-First Search, Feature selection, Supervised learning model, Genetic algorithm, Multilayer perceptron neural network

Chadaphim Photphanloet : AUGMENTED SENSORS FOR PARTICULATE MATTER CONCENTRATION PREDICTION USING SUPERVISED LEARNING MODELS. Advisor: Assoc. Prof. RAJALIDA LIPIKORN, Ph.D.

Particulate matter concentration prediction models have been researched, developed, and applied to data from various topography around the world. The characteristics of different topography make a model suitable for each specific topography. This dissertation proposes a novel method to predict particulate matter concentration with a diameter smaller than 10 microns in Nan Province of Thailand that integrates feature selection method, supervised learning model, and modified depth-first search algorithm. Unlike the traditional supervised learning models, the proposed method is able to accept multi-dimensional data as input which consist of particulate matter concentration, air pollutants, and air qualities. These features are the factors that influence particulate matter concentration prediction. The experimental results show that the proposed method performs better than other methods when predicting the concentration one hour ahead with no need of wind direction and wind speed data. The proposed method was developed with a general framework and could be applied to predict particulate matter concentration in Nan Province.

Field of Study:	Applied Mathematics and Computational Science	Student's Signature ....
Academic Year:	2019	Advisor's Signature ....

ACKNOWLEDGEMENTS

I would like to express the deepest appreciation to Assoc. Prof. Rajalida Lipikorn, Ph.D., my research advisor, for her valuable guidance, kind supervision, patience, motivation, enthusiasm, and immense knowledge. Her guidance helped me in completing my research and writing dissertation.

I would like to thank for data supported by Garavig Tanaksaranond, Ph.D., at the Department of Survey Engineering, Faculty of Engineering, Chulalongkorn University.

I would like to thank all committees for their precious comments.

I would like to thank Mr. Sathawit Suwannarut for topographic map support, friendly and helpful support in dissertation work.

I would like to thank my friends in Applied Mathematics and Computational Science (AMCS) and Machine Intelligence and Multimedia Information Technology Laboratory (MIMIT Lab) for their suggestion.

Additionally, this research was supported by the Science Achievement Scholarship of Thailand (SAST), under the collaboration of the Office of the Higher Education Commission and the meeting of Science Dean of Thailand.

Finally, I would like to thank my family for their understanding and support.

Chadaphim Photphanloet

TABLE OF CONTENTS

Page

ABSTRACT (THAI) iii

ABSTRACT (ENGLISH) iv

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS vi

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER 1 INTRODUCTION 1

1.1 Rationale 1

1.2 Objectives 2

1.3 Scope 2

1.4 Research methodology 3

1.5 Expected outcomes 4

CHAPTER 2 LITERATURE REVIEW 5

CHAPTER 3 THEORETICAL BACKGROUND 9

3.1 Environmental Science 9

  3.1.1 Air quality 9

  3.1.2 Air pollution 10

3.2 Relevant Mathematical Theories 11

  3.2.1 Feature selection 11

  3.2.2 Prediction model 21

  3.2.3 Performance evaluation 35

CHAPTER 4 METHODOLOGY 37

4.1 Data Preprocessing 38

4.2 PM_{10} Concentration Prediction 54

CHAPTER 5 RESULTS 59

5.1 The normalized data 59

5.2 The experimental results for PM_{10} concentration prediction 70

CHAPTER 6 DISCUSSION AND CONCLUSION 83

REFERENCES 86

APPENDIX 90

VITA 112

LIST OF TABLES

Page

Table 1 The level of health effect based on an average of every 24-hour period 11

Table 2 An example of forward selection and backward elimination for a set of relevant features 13

Table 3 Terminology for the genetic algorithm 14

Table 4 The definitions of variables for the multilayer perceptron neural network ... 25

Table 5 Examples of kernel functions 33

Table 6 The spans between latitude and longitude of the 14-sensor network system 41

Table 7 The distance between six sensors at monitoring stations 41

Table 8 The basic statistical values of PM_{10} concentrations for every 24 hours of each monitoring station 52

Table 9 Parameter settings for the feature selection methods 56

Table 10 Parameter settings for supervised learning models 56

Table 11 The definitions of variables for the PM_{10} concentration prediction 57

Table 12 The PM_{10} concentration prediction results of each monitoring station using three models. 71

Table 13 The PM_{10} concentration prediction results for the next hour of each observed station using feature selection and supervised learning models. 72

Table 14 The PM_{10} concentration prediction results for the next hour of each observed station using MDFS algorithm. 73

Table 15 The PM_{10} concentration prediction results for the next hour of each observed station using MDFS algorithm with the actual data. 79

LIST OF FIGURES

Page

Figure 1 Comparative size of particulate matter (from https://www.epa.gov/pm-pollution/particulate-matter-pm-basics)............ 10

Figure 2 Population, chromosomes, and genes............ 14

Figure 3 The process of the genetic algorithm............ 15

Figure 4 The binary encoding............ 16

Figure 5 The value encoding............ 16

Figure 6 The permutation encoding............ 17

Figure 7 Roulette wheel selection............ 18

Figure 8 Examples of crossover methods............ 20

Figure 9 An example of bit-flip mutation............ 21

Figure 10 Multilayer perceptron neural network............ 24

Figure 11 Architecture of backpropagation neural network............ 25

Figure 12 Architecture of feedforward neural network from the input layer to the hidden layer............ 26

Figure 13 Architecture of feedforward neural network from the hidden layer to the output layer............ 27

Figure 14 Architecture of error estimation for backpropagation neural network............ 28

Figure 15 Architecture of weight adjustment in backpropagation neural network............ 28

Figure 16 The soft margin liner SVR............ 30

Figure 17 Overview and the flowchart of the proposed method............ 37

Figure 18 The location of Nan Province............ 38

Figure 19 Air pollution in Nan Province as a case study............ 39

Figure 20 Positions of six sensors at monitoring stations 40

Figure 21 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_1 43

Figure 22 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_2 45

Figure 23 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_3 46

Figure 24 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_4 48

Figure 25 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_5 50

Figure 26 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_6 51

Figure 27 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_1 61

Figure 28 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_2 62

Figure 29 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_3 64

Figure 30 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_4 66

Figure 31 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_5 67

Figure 32 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_6 69

Figure 33 The process of selecting data from the neighboring stations using the MDFS algorithm for PM_{10} concentration prediction of each monitoring station 77

Figure 34 The process of selecting data from the neighboring stations using the MDFS algorithm with the actual data for PM_{10} concentration prediction of each monitoring station 82

CHAPTER 1 INTRODUCTION

1.1 Rationale

Atmospheric aerosol particles, also known as atmospheric particulate matter (PM), are important factors that influence the weather and climate changes which have an impact on life and Earth. According to the United States Environmental Protection Agency, particulate matters that affect human health and ecosystems are PM_{10} and $PM_{2.5}$. PM_{10} is inhalable particles with diameters less than or equal to 10 micrometers, and $PM_{2.5}$ is fine inhalable particles with diameters less than or equal to 2.5.

In 2000, Friedlander et al. presented that the main sources of particulate matter are from natural phenomena and anthropogenic [1]. Over the past few decades, particulate matter, especially PM_{10}, has become one of the most interesting subjects for research due to its effects on human health and ecosystems. It is commonly known that particulate matter can penetrate into the respiratory system and cause respiratory diseases [2]. Dockery et al [3] and Kingdom et al. [4] showed clear scientific evidence that PM_{10} with low concentration in ambient air can damage human health.

Air pollutant quality models are essential and important in the management of weather for integrated pollution and environment management program. In Thailand, the government has paid great attention to solve the pollution problem and defined the problem as particulate matter-related policy in the 20-year national strategy (2017 - 2036 A.D.). Air pollutant quality simulation in different management scenarios using models makes a great support to various development planning; therefore, the development of atmospheric particulate matter quality models in Thailand are needed.

Air pollutant models have been developed and used for specific topography because each topography has unique structure and characteristics. Thus an accurate model for each topography must be developed and the performance based on real data collected from each area should be evaluated. Since particulate matter pollution is one of the major problems in Nan province of Thailand because most of the farmers burn their corn fields after harvesting in order to prepare for the next plantation which cause PM_{10} concentration to be above the standard, these problems motivate us to uses Nan province as a case study for developing the atmospheric particulate matter quality prediction models, especially the models for predicting PM_{10} concentration. This would be a challenge to develop an appropriate model that can predict PM_{10} concentration of Nan province under different scenarios based on historical PM_{10} concentration records.

1.2 Objectives

1. Design and develop a method for predicting PM_{10} concentration in the corn fields of Nan Province after harvesting.

2. Derive the most suitable set of input parameters for predicting PM_{10} concentration.

3. Predict PM_{10} concentration in the areas of Nan Province of Thailand as a case study using the proposed method.

1.3 Scope

In this dissertation, the method is constrained as follows:

1. The proposed method uses data obtained from the online sensor network system at each monitoring station which consist of two air quality data and three air pollutant data; i.e., temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from the beginning of February 2017 to the end of April 2017.

2. Data from only six sensors are used in the experiments because these sensors have complete data recorded.

3. The particulate matter concentration prediction is considered as a regression problem; therefore, three supervised learning models for regression problems are

used; i.e., multiple linear regression, multilayer perceptron neural network, and support vector regression.

1.4 Research methodology

The research procedures are organized into seven parts as follows:

1. Review related literatures on particulate matter, especially PM_{10}, and study the background knowledge about the definitions and the properties of feature selection and the existing models for PM_{10} concentration prediction.

2. Gather data from the sensor at each monitoring station in Nan Province of Thailand and preprocess the data to be ready for the experiments.

3. Implement the supervised learning models for predicting PM_{10} concentration at each monitoring station which include multiple linear regression, multilayer perceptron neural network, and support vector regression.

4. Improve the accuracy of PM_{10} concentration prediction by combining feature selection methods to select only relevant data from the sensor at each monitoring station.

5. Conduct the experiments on PM_{10} concentration prediction at each monitoring station using the combination of feature selection methods and the supervised learning models.

6. Improve the accuracy of PM_{10} concentration prediction at each monitoring station by including data from the nearby stations using the modified depth-first search algorithm.

7. Prepare a manuscript for submitting to the journal and write the dissertation.

1.5 Expected outcomes

If the concentration of air pollutants, especially PM_{10}, can be predicted in advance, the warning or alert can be announced to the residents of the nearby areas so that they can prepare and protect themselves from air pollution. Moreover, the aim of this research is to find the optimal features that have significant effects on PM_{10} concentration such that we can prevent dangerous situations to arise by controlling these features.

CHAPTER 2 LITERATURE REVIEW

In this chapter, related work and literatures associated with this dissertation are reviewed. Modelling of particulate matter concentration prediction is considered. The existing stage-of-the-art methods for prediction are examined to find out the advantages and the disadvantages of the models. Finally, the proposed model and its contribution are pointed out.

Prediction models of particulate matter concentration are the effective tools for predicting and simulating the air pollution in every geographical area. Particulate matter is primarily generated through natural processes or from human activity [5, 6]. Since it can be harmful to human health as well as to the environment and other living things, the issue of particulate matter in the atmosphere, especially PM_{10}, has increasingly drawn the attention of the public and the scientific community. One particular problem caused by PM_{10} is the harm caused to the human respiratory system [2]. The evidence put forward by scientists suggests that when PM_{10} concentration levels in the air the standard level during each period of time, human health will be adversely affected [3]. One way to limit the damage would be to find a means of accurately predicting PM_{10} concentration levels for the upcoming days in order to enforce prevention and control measures to protect the public during the worst periods.

In 2011, Kuo-Ping Lin et al. [7] developed the preprocessing procedures for data and used immune algorithm for optimization to accurately predict the air pollutant concentration called support vector regression with logarithm preprocessing procedure and immune algorithm. As the results, the proposed model can reliably predict air pollutant concentrations.

In 2014, Chen Xi Zhao et al. [8] presented the temporal and spatial distribution of pollution status and association of meteorological factors and particulate matter. Meteorological variables such as average daily relative humidity, precipitation, temperature, and wind speed were selected and their relationship with $PM_{2.5}$ and PM_{10} concentrations were examined using Spearman rank correlation analysis.

Concentrations of $PM_{2.5}$ and PM_{10} have been shown to be positively associated with relative humidity and temperature and correlated negatively with wind speed. From this literature, relative humidity and wind speed are the main determinants that influence PM_{10} and $PM_{2.5}$ concentration distributions.

In 2015, Han Li et al. [9] considered the relationship between factors of meteorology and the particulate matter during the summer in Shijiazhuang. The results reveal that the meteorological factors, such as precipitation, wind speed, air pressure, and air temperature, can influence the concentration of atmospheric pollutants. In this literature, it was found that wind speed exhibits a slight impact on air pollution.

In order to develop an accurate forecasting model, Auder et al., in 2016 [10] suggested sequential aggregation of the output from several PM_{10} forecasting models and used them as input to another forecasting model for PM_{10} concentration. The input data included the average daily concentrations of PM_{10} at each monitoring station, while the average daily data from other measures could also be added to the input data for forecasting PM_{10} concentration of the next day. Different models could be developed on this basis. The findings indicate that this approach is capable of improving the expert system process and can generate better warnings and predictions.

In 2017, German Hernandez et al. [11] presented the climate condition and PM_{10} concentration analysis. It was found that relative humidity and temperature have relationships with particulate matter. Moreover, the relationships between particulate matter of different sizes are strong. The temperature over a diurnal period was found to have a negative correlation with PM_{10}; whereas, relative humidity has a positive correlation with PM_{10}.

In 2018, Fabiana Franceschi et al. [12] determined the most important meteorological variables on air pollution based on data mining algorithms. In this research, principal component analysis was employed in order to assess the variables, while data were grouped by using the k-means clustering. The prediction models then used the results of these processes as the input. The predicted results demonstrate that the proposed models can be used as the indicators which can provide advance warnings when high level of air pollution is going to occur because they are able to predict the incidence of high pollution with a relatively good degree of accuracy.

Working in the Oviedo area, García Nieto et al. in 2018 [13] developed a model capable of predicting PM_{10} concentration levels using a vector support machine along with a multilayer perceptron neural network, a vector autoregressive moving average model, and an autoregressive integrated moving average model. These models used pollution concentration data in the form of monthly averages as the input data. The findings suggest that PM_{10} concentrations can best be predicted through the use of a vector support machine that uses the radial basis function as a kernel function and uses the idea of cross-validation to serve as a mean of estimating accuracy levels.

In 2019, Zhigen Shang et al. [14] proposed a novel model for prediction based on the methods of regression and classification tree, and the ensemble extreme learning machine. Because hourly concentrations of $PM_{2.5}$ have several patterns of shift, it is useful to divide the entire dataset into several subsets with similar properties and to train a local prediction model for each subset. They used a shallow hierarchical regression tree, regression and classification tree to divide the dataset. Then, ensemble extreme learning machine models were built at each node of the tree using the node training samples, and the number of hidden neurons were chosen to minimize validation errors on the subtree’s leaves that take the node as root. Finally, the global and the local ensemble extreme learning machines were compared for each tree leaf on the path from the root to the leaf, and the one with the minimum validation error on the leaf was selected. The new model enhanced the ability to handle multiple patterns of transition.

In this research, we propose a novel PM_{10} concentration prediction model to predict the concentration at any station of interest using the combination of data from the station of interest and its neighboring stations. The proposed method can accurately predict PM_{10} concentration without knowing the wind speed and the wind direction.

Debate has been focused on the basic issue of whether a better model for PM_{10} concentration prediction such that we can prevent and protect our health from the effects of particulate matter. Thus, the contribution of this study is to develop a new model to predict PM_{10} concentration ahead of time from available air quality and air pollution data in order to provide the warning to local residents about the level of air pollution.

CHAPTER 3 THEORETICAL BACKGROUND

In this chapter, we present the principle knowledge which contains two parts of important background in Environmental Science and relevant mathematical theories. Environmental Science part explains the definition of air quality, air pollution, and particulate matter standard regulatory in Thailand according to regulations. The relevant theoretical mathematical part describes the fundamental concepts of techniques used in this dissertation.

3.1 Environmental Science

3.1.1 Air quality

Air quality is the degree to which air is sufficient or safe enough to keep humans, animals or plants healthy. In this dissertation, two types of air quality are mentioned; i.e., temperature and humidity.

Air temperature is a fundamental factor in studying weather because the temperature changes every period, such as year, season, month, day, and even hour. Temperature is a physical property of matter at a given moment that quantitatively expresses hot and cold weather. The factors that affect the temperature consist of ground and water, elevation, latitude, geography, cloud volume, and albedo of the surface. Air temperature is a proportion of temperature at various degrees of the Earth's climate. It is represented by numerous components, including approaching sunlight based radiation, moistness, and height.

Humidity is another air quality used in PM prediction, it represents the amount of water vapor in the air. The humidity of air changes over time depending on the pressure and the temperature. As the amount of water vapor in the air depends on the temperature of the air, thus warm climate can maintain more water vapor than cold air.

3.1.2 Air pollution

Air pollution is the mixture of toxic chemicals or compounds that contaminates the Earth's atmosphere and can be very harmful to all mankind. In this dissertation, three types particulate matter of air pollutant are mentioned.

Atmospheric particulate, also referred to as particulate matter (PM), is any compound besides uncombined water that exists in a finely divided structure as fluid or mass. PM is an essential element that impacts the climate and the changes of climate can have a big impact on life and Earth. Sources of particulate matter in the atmosphere are divided into two major types; i.e., natural source, and man-made source. Examples of natural sources are soil, sand, stone, steam, smoke from forest fires, salt dust of the sea, and etc.; whereas, examples of man-made sources are transportation and traffic, construction, industrial enterprises, incineration in open areas for agriculture, and etc.

The National Environment Board, Pollution Control Department of Thailand classifies particulate matter based on its size, in accordance with the United States Environmental protection corporations, into two classes: PM_{10} and $PM_{2.5}$. PM_{10} is referred to any particulate matter with a diameter smaller than 10 microns while $PM_{2.5}$ is referred to any particulate matter with a diameter smaller than 2.5 microns as shown in Figure 1.

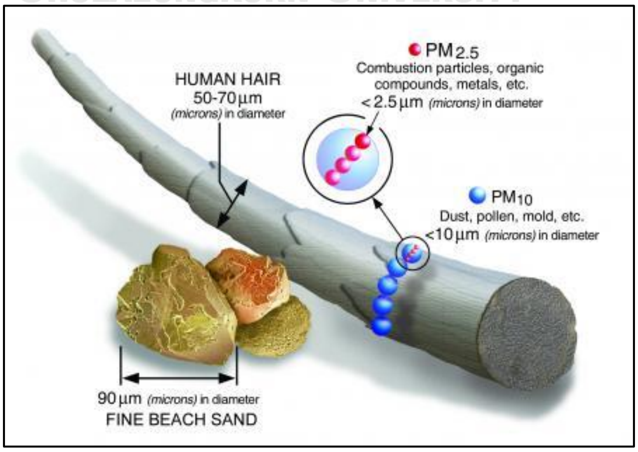

PM2.5
Combustion particles, organic compounds, metals, etc.
<2.5µm (microns) in diameter
HUMAN HAIR
50-70µm
(microns) in diameter
PM10
Dust, pollen, mold, etc.
<10µm (microns) in diameter
90µm (microns) in diameter
FINE BEACH SAND

Figure 1 Comparative size of particulate matter (from https://www.epa.gov/pm-pollution/particulate-matter-pm-basics)

The National Environment Board, Pollution Control Department of Thailand has issued standard regulations for the concentration of particulate matter as follows: particulate matter with a diameter smaller than 100 microns shall not surpass 330 $\mu g/m^3$ in 24-hour average, particulate matter with a diameter smaller than 10 microns shall not surpass 120 $\mu g/m^3$ in 24-hour average, and particulate matter with a diameter smaller than 2.5 microns shall not surpass 50 $\mu g/m^3$ in a 24-hour average. The concentration of particulate matter has an impact of our health and the level of health effect caused by PM_{10} and $PM_{2.5}$ concentrations that are continuously measured from the average of every 24-hour period are shown as Table 1.

Table 1 The level of health effect based on an average of every 24-hour period

$PM_{2.5}$ $(\mu g/m^3)$	PM_{10} $(\mu g/m^3)$	The level of health effects
0-25	0-50	Safe
26-37	51-80	Stay alerted
38-50	81-120	Affecting health
51-90	121-180	Significant health effects
>90	>180	Severe health effects

3.2 Relevant Mathematical Theories

3.2.1 Feature selection

One critical step in the prediction process is how to select a suitable subset of features that will be used as input data. In most cases, using all of the features for prediction gives lower accuracy rate than using only relevant features that are significant for prediction, thus, the first step of the proposed method is to select the most suitable subset of features which are unknown. If any significant features are to be inadvertently omitted from the process, the possibility of achieving accurate prediction would be greatly reduced. Conversely, if too many irrelevant features are included in the input data, the accuracy can be declined, and the time complexity for the training will be higher as well. In addition, the problem of overfitting, as reported by Maier and Dandy in 2000 [15], can also result from excessive input features. In this research, three feature selection techniques are presented: forward selection (FS), backward elimination (BE), and genetic algorithm (GA). These feature selections are used to select only the features that have direct impact upon the concentration of particulate matter. We propose to apply feature selection because if we use a combination of all features as input data, the computational complexity will be 30 factorials for each round of prediction.

1. Forward selection

Forward selection is a stepwise process that starts from adding features to the model one at a time. At each step, each feature that is not already added to the model is tested for inclusion into the model using correlation between dependent and independent features. Forward selection starts from a null set of features. Each feature is added to the model one at a time as long as the model's p-value is below some preset significance level; for example, 0.05. This process stops when adding a new feature to the model gives non-statistical significance level (i.e., p-value>0.05) [16, 17]. An example of forward selection for a subset of features is shown in Table 2. However, the drawback of forward selection is that the correlation between features in an existing set is not taken into consideration because each set of new features is built from a new feature that is added to the set in each round; thus, it is possible that adding new features into the existing set does not guarantee the improvement of the model.

2. Backward elimination

Since forward selection has some drawbacks, an alternative approach which avoids this problem is backward elimination. The backward elimination starts from fitting a model with all of the features. Then each irrelevant feature is eliminated from the model one at a time by considering the feature with the maximum p-value. If the maximum p-value is more than the preset significance level (i.e., p-value > 0. 05) then this feature is eliminated else the backward elimination stops [16, 17]. The process is repeated until an optimal set of features is found. An example of backward elimination for feature selection is shown in Table 2.

Table 2 An example of forward selection and backward elimination for a set of relevant features.

Forward selection	Backward elimination
Input: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial set of features: $\{\quad\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_3\}$ $\Rightarrow$ Add a feature: $\{A_1, A_3, A_4\}$	Input: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial set of features: $\Rightarrow \{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_2, A_4, A_5, A_6\}$ $\Rightarrow$ Remove a feature: $\{A_1, A_2, A_5, A_6\}$

3. Genetic algorithm

The genetic algorithm (GA) is the method of finding the best answer for the problem by mimicking natural evolution, which is based on the natural evolution theory of Charles Darwin [18, 19], who said that the stronger the person has, the higher probability of surviving and having a chance in the inheritance of strong characteristics into the next generation. This concept was applied in search and optimization problems by Holland in 1975 [20] and developed by Goldberg in 1989 [21]. Genetic algorithm is one of the artificial intelligence techniques that simulates biological process or natural evolution in the birth of new generations based on the genetic evolution in the inheritance of various characteristics into the next offspring. Since the genetic algorithm is based the theory of evolution, the terms used in GA are shown in Table 3 and the structures of population, chromosomes, and genes are shown in Figure 2.

Table 3 Terminology for the genetic algorithm

Term	Description
Chromosome	An individual contains a set of genes
Gene	A feature (parameter) on a chromosome
Population	A set of chromosomes

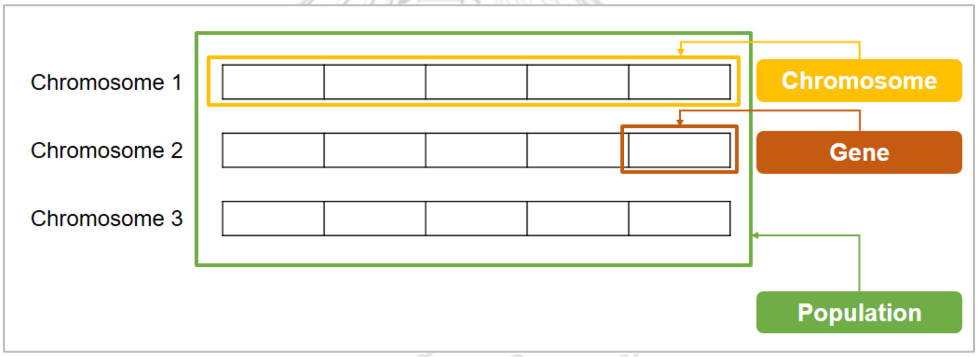

Chromosome 1
Chromosome 2
Chromosome 3
Chromosome
Gene
Population

Figure 2 Population, chromosomes, and genes

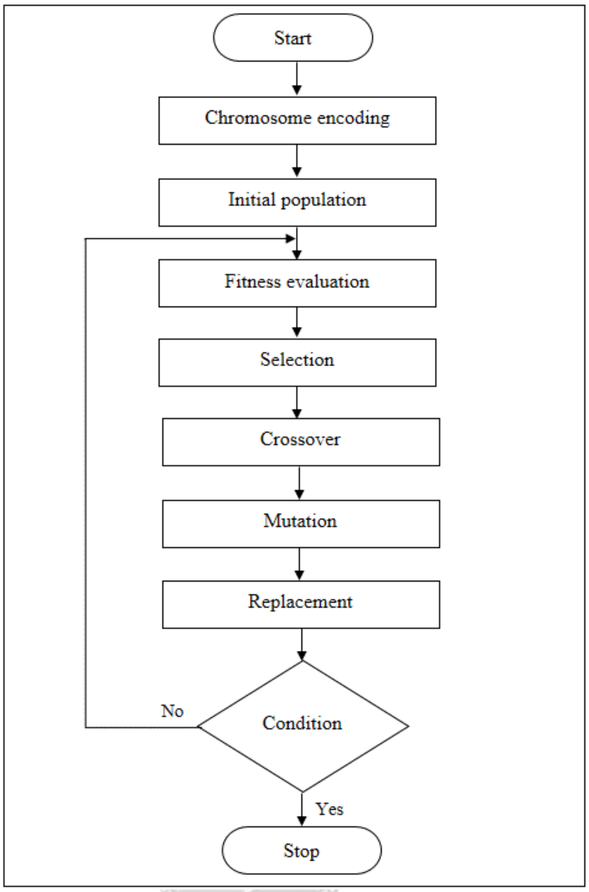

Start
Chromosome encoding
Initial population
Fitness evaluation
Selection
Crossover
Mutation
Replacement
No
Condition
Yes
Stop

Figure 3 The process of the genetic algorithm

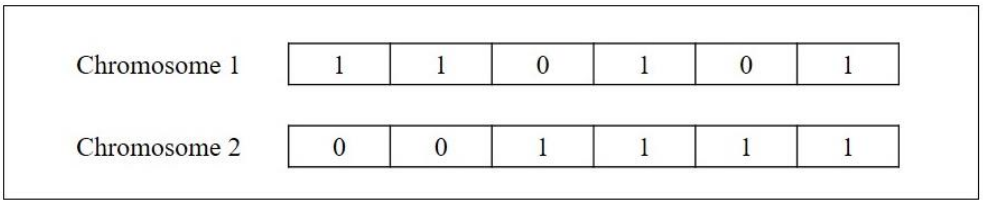
From Figure 3, the process details of genetic algorithm consists of chromosome encoding, population initialization, fitness evaluation, genetic operations which consists of crossover and mutation, and replacement. The process of the genetic algorithm can be explained as follows:

Step 1: Chromosome Encoding

Chromosome encoding is an important step in the genetic algorithm. It is the process of representing a chromosome in a form of a string of values that contain useful information. Each position in a string represent each gene in a chromosome that contains a feature of the solution. The values of each string depend on the problem to be solved; thus, there are several chromosome encoding methods to be used depending on a particular problem. In this section, three chromosome encoding methods are explained as follows:

1.1) Binary Encoding

This is the basic encoding method that assigns the value to each position of a string with 0 or 1 as shown in Figure 4. Chromosomes obtained from this method are string of 0s and 1s where each gene in a chromosome represents a feature of a solution to the problem.

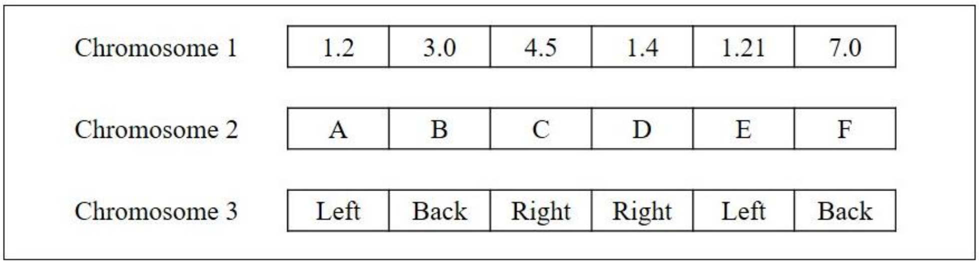


Chromosome 1	1	1	0	1	0	1
Chromosome 2	0	0	1	1	1	1

Figure 4 The binary encoding

1.2) Value Encoding

For the value encoding method, a chromosome is encoded as a string of values, such as real numbers, characters, or words when binary encoding is not capable of representing such complicated values as shown in Figure 5. Chromosome 1 in Figure 5 is a sample of an encoding for a problem of finding the weights in neural networks by representing each gene with real number corresponding to the weights for the inputs. Chromosome 2 in Figure 5 is a sample of using the alphabet to encode a chromosome for the bus route management problem. Chromosome 3 in Figure 5 is a sample of encoding a chromosome for a problem of the route finding problem.

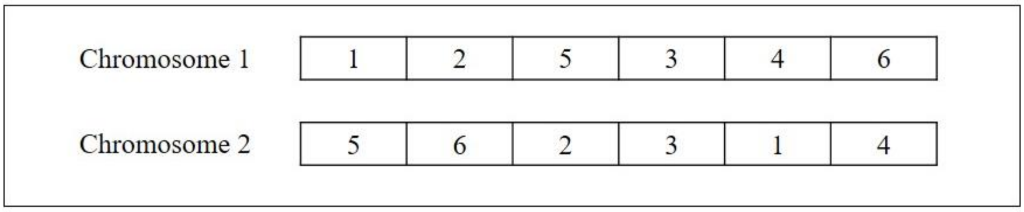


Chromosome 1	1.2	3.0	4.5	1.4	1.21	7.0
Chromosome 2	A	B	C	D	E	F
Chromosome 3	Left	Back	Right	Right	Left	Back

Figure 5 The value encoding

1.3) Permutation Encoding

All genes in a chromosome are encoded as a string of positive integers to represent numbers in sequence. Permutation Encoding is suitable for use in the problem of arranging a set of data whose values must not be repeated in an individual string, such as the traveling salesman problem or the scheduling problem. Figure 6 shows chromosomes that are encoded by permutation encoding for the traveling salesman problem where each number represents a town to be visited. The positions of the genes give the sequence of the towns to be visited. Chromosome 1 and chromosome 2 are some possible sequences of towns for the salesman to visit.



Chromosome 1	1	2	5	3	4	6
Chromosome 2	5	6	2	3	1	4

Figure 6 The permutation encoding

Step 2: Population Initialization

In this step, a population or a set of individuals is randomly generated. The size of a population depends on the number of genes. In general, the size of population is suggested to be approximately 1.5 - 2 times the number of genes, and the maximum size should be no more than 100 [22].

Step 3: Fitness evaluation

For this step, the fitness score that determines each chromosome's suitability can be calculated from the fitness function. In this dissertation, prediction models for the particulate matter are used as fitness functions, i.e., multiple linear regression, multilayer perceptron neural network and support vector regression, to calculate the fitness score of each chromosome (f_t).

Step 4: Selection

In this step, the fittest chromosomes with high fitness scores are selected and their genes will be passed to the next generation. Two pairs of chromosomes are selected as 'parents' for the reproduction. Chromosomes with high fitness scores have more chance of being selected. Two selection techniques that are widely used are as follows:

1) Roulette Wheel Selection

In this method, the selection principle is an imitation of playing roulette. A roulette wheel assigns a proportion to each chromosome based on its fitness score that is calculated from Equation (3.1). Figure 7 shows an example of roulette wheel selection where the number of chromosomes of the population size is equal to 5.

$$fitness(t) = \frac{f_t}{\sum_{t=1}^{P} f_t} \tag{3.1}$$

where f_t is the fitness score of a chromosome t, $t = 1, 2, 3, \ldots, P$, and P is the population size.

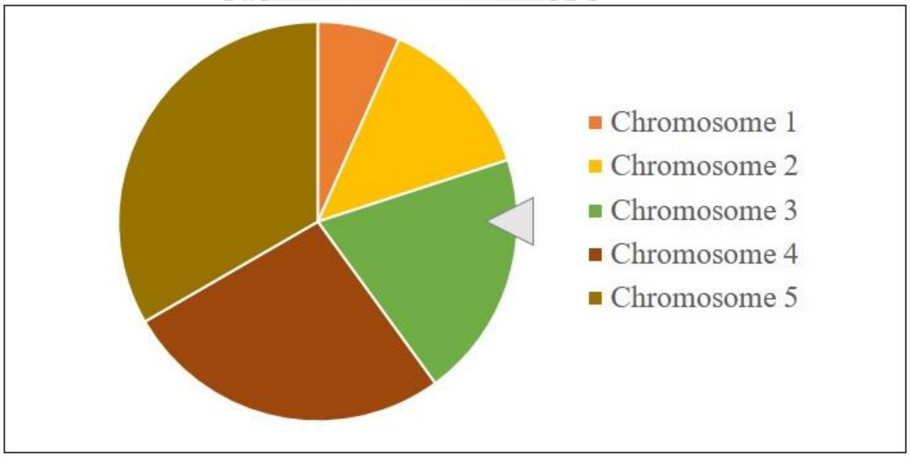


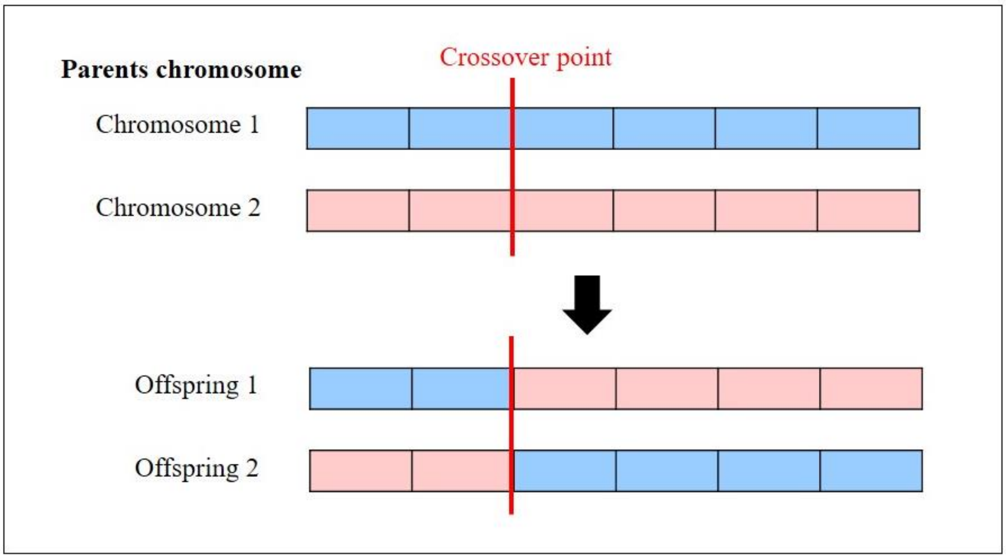
Figure 7 Roulette wheel selection

2) Tournament Selection

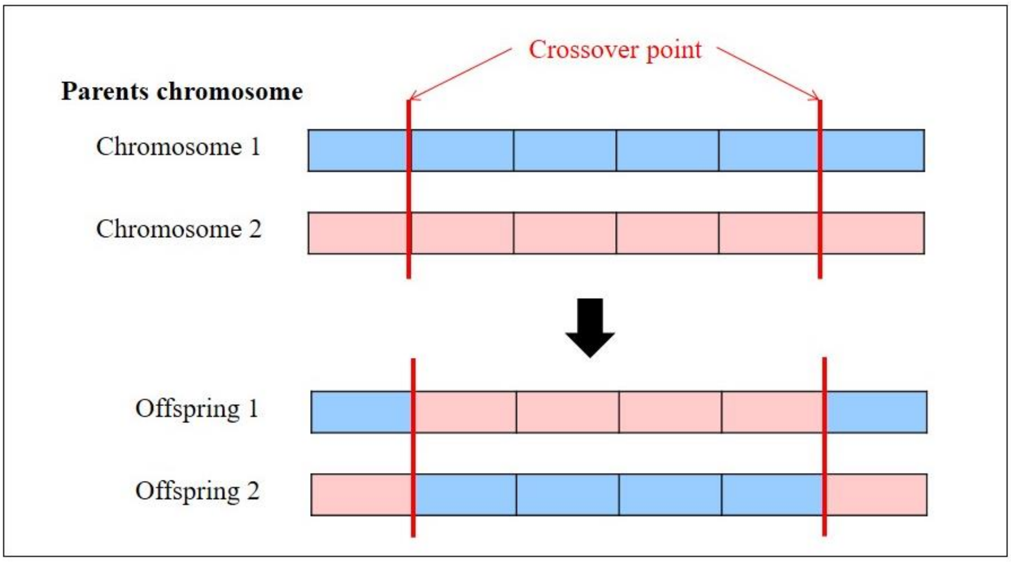
The tournament selection method is used to select the fittest chromosomes to be pairs of parents similar to the sport events. The fundamental steps of tournament selection are as follows: (i) determine the tournament size (the number of chromosomes for each round of the tournament) K, (ii) randomly select K chromosomes from the population, (iii) select the best chromosome from K chromosomes with probability p, which has in the interval $[0,1]$, (iv) select the second best chromosome from K chromosomes with probability $p(1-p)$, which has the desired amount of chromosomes equal to parent's size.

Step 5: Crossover

Crossover is the most important step for GA. At first, a crossover point, where the exchange of genes between each pair of parents to be mated in order to create offspring, is chosen at random based on the crossover probability in the range of $[0,1]$ [23, 24]. Each pair of parents create two offspring through the exchange of genes between parents until reaching the crossover point. There are many crossover methods to generate offspring, for example, one-point crossover and two-point crossover. Figure 8(i) shows the one-point crossover method where a single crossover point is randomly selected. Figure 8(ii) shows the two-point crossover method where two crossover points are randomly selected and the genes of two parents within between the two crossover points are exchanged. As a result, each pair of parents generates two new offspring, and they are delivered to the new population. Therefore, the population size of the new generation is the same as the initial population size.


Parents chromosome
Crossover point
Chromosome 1
Chromosome 2
Offspring 1
Offspring 2

(i) The one-point crossover


Crossover point
Parents chromosome
Chromosome 1
Chromosome 2
Offspring 1
Offspring 2

(ii) The two-point crossover

Figure 8 Examples of crossover methods

Step 6: Mutation

Mutation is the next step to perform after crossover. The purpose of mutation is to change the genes of the existing chromosome to maintain diversity and to avoid duplicate chromosome problem. The general concept of mutation is to randomly select the genes and change the values of the genes under the probability in the range of $[0,1]$ [24]. At present, there are many mutation techniques, such as bit-flip mutation, inversion mutation, insertion mutation, uniform mutation, and others. An example of bit-flip mutation on a binary encoded chromosome is shown in Figure 9 where two genes are randomly selected and the values of the genes are flipped.

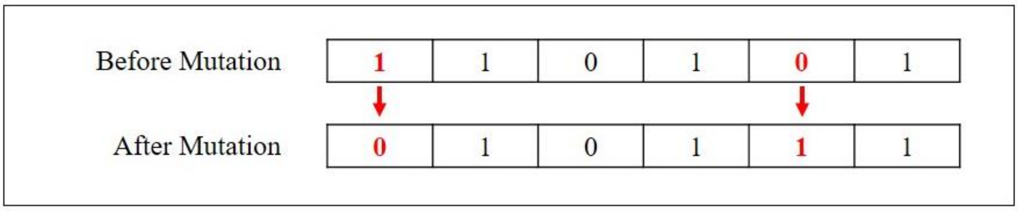


Figure 9 An example of bit-flip mutation

Step 7: Replacement

The current population is replaced with the new population of the same size. The algorithm then repeats steps 3 – 7 until it reaches a termination condition; i.e., reaching the maximum number of iterations or converging to the solution which means that the offspring of the current population are not quite different from the offspring of the previous population.

3.2.2 Prediction model

The prediction of particulate matter concentration may be considered a regression problem, since the predictive model aims to approximate the mapping function from the input characteristics to the continuous output of particulate matter concentration and the relationship between data is unknown; thus, multiple linear regression, multilayer perceptron neural network, and support vector regression are used in this dissertation. The reason that these three models are used is they return quantitative results, their output can be assessed using the most common root mean squared error [25].

1. Multiple linear regression

The outcomes of dependent variables can be predicted through the use of a number of independent variables in a multiple linear regression approach. This is a statistical technique which models the linear relationship that arises between the independent variables and the desired dependent variable [26]. Suppose $\mathbf{X_i}$ is a vector of independent variables where $i = 1, 2, 3, \ldots, k,$ which has a relationship with a vector of output, $\mathbf{Y}$. The MLR model for this study is expressed in Equation (3.2).

$$\mathbf{Y} = \beta_0 + \beta_1\mathbf{X_1} + \beta_2\mathbf{X_2} + \cdots + \beta_k\mathbf{X_k} + \mathbf{e} \tag{3.2}$$

where $\mathbf{Y}$ is a vector of dependent variables, $\mathbf{X_i}$ is the ith vector of independent variables, $\boldsymbol{\beta_i}$ is the weight of individual independent variables for $i = 1, \ldots, k$, $\boldsymbol{\beta_0}$ is a constant, $\mathbf{e}$ is an error value, and k is the number of independent variables. The conditions for the MLR model are as follows: (1) the error value has a normal distribution, (2) the average of the error value is equal to zero, (3) the variance of error value is steady, but has unknown value, and (4) the variances at i and q are independent where $i \neq q$ for $i, q = 1, \ldots, k$.

The optimum values of β_0 and β_i can be found when the least square error occurs [27]. Let Equation (3.3) represent the dependent variable y_ρ where $\rho = 1,2, \ldots, m$ and m is the total number of data.

$$y_\rho = \beta_0 + \beta_1 x_{1\rho} + \beta_2 x_{2\rho} + \cdots + \beta_k x_{k\rho} + e_\rho \tag{3.3}$$

Suppose an approximation of Equation (3.3) takes the form in Equation (3.4)

$$\hat{y}_\rho = a + b_1 x_{1\rho} + b_2 x_{2\rho} + \cdots + b_k x_{k\rho} \tag{3.4}$$

Then the aggregate of the squares of the errors is described by

$$\sum_{\rho=1}^{m} \left(y_\rho - \hat{y}_\rho\right)^2 = \sum_{\rho=1}^{m} \left(y_\rho - a - b_1 x_{1\rho} - b_2 x_{2\rho} - \cdots - b_k x_{k\rho}\ \right)^2 \tag{3.5}$$

and the minimum value with respect to a and b_i for $i = 1, \dots, k$ of Equation (3.5) is given by:

$$\frac{\partial}{\partial a}\sum_{\rho=1}^{m}\left(y_\rho - \hat{y}_\rho\right)^2 = 0\;;$$

$$na + b_1\sum_{\rho=1}^{m} x_{1\rho} + b_2\sum_{\rho=1}^{m} x_{2\rho} + \cdots + b_k\sum_{\rho=1}^{m} x_{k\rho} = \sum_{\rho=1}^{m} y_\rho \qquad (3.6)$$

$$\frac{\partial}{\partial b_1}\sum_{\rho=1}^{m}\left(y_\rho - \hat{y}_\rho\right)^2 = 0\;;$$

$$a\sum_{\rho=1}^{m} x_{1\rho} + b_1\sum_{\rho=1}^{m} x_{1\rho}^2 + b_2\sum_{\rho=1}^{m} x_{1\rho}x_{2\rho} + \cdots + b_k\sum_{\rho=1}^{m} x_{1\rho}x_{k\rho} = \sum_{\rho=1}^{m} x_{1\rho}y_\rho \qquad (3.7)$$

$$\frac{\partial}{\partial b_2}\sum_{\rho=1}^{m}\left(y_\rho - \hat{y}_\rho\right)^2 = 0\;;$$

$$a\sum_{\rho=1}^{m} x_{2\rho} + b_1\sum_{\rho=1}^{m} x_{1\rho}x_{2\rho} + +b_2\sum_{\rho=1}^{m} x_{2\rho}^2 + \cdots + b_k\sum_{\rho=1}^{m} x_{2\rho}x_{k\rho} = \sum_{\rho=1}^{m} x_{2\rho}y_\rho \qquad (3.8)$$

$$\vdots$$

$$\frac{\partial}{\partial b_k}\sum_{\rho=1}^{m}\left(y_\rho - \hat{y}_\rho\right)^2 = 0\;;$$

$$a\sum_{\rho=1}^{m} x_{k\rho} + b_1\sum_{\rho=1}^{m} x_{1\rho}x_{k\rho} + b_2\sum_{\rho=1}^{m} x_{1\rho}x_{k\rho} + \cdots + b_k\sum_{\rho=1}^{m} x_{k\rho}^2 = \sum_{\rho=1}^{m} x_{k\rho}y_\rho \qquad (3.9)$$

Putting equations together from Equation (3.6) to Equation (3.9) are called the normal equation system which is used to solve for a and b_i for $i = 1, \dots, k$.

2. Multilayer perceptron neural network

One of several neural networks in feed forward artificial neural network category is a multilayer perceptron neural network (MLP) [28]. MLP essentially comprises of three parts: network layers, weight adjustment, and activation function. The network layer consists of an input layer, hidden layers, and an output layer. MLP for prediction uses a supervised learning technique for training, called backpropagation [29] which is one of the popular techniques for multilayer neural network because it can solve the problems of data sets with linear and nonlinear characteristics. The weight adjustment process uses the error values to adjust the weights of all nodes in the layer and all nodes in the adjacent layer. Each node is fully connected to all nodes in the next layer as shown in Figure 10. Figure 11 shows the weight adjustment process of the backpropagation neural network.

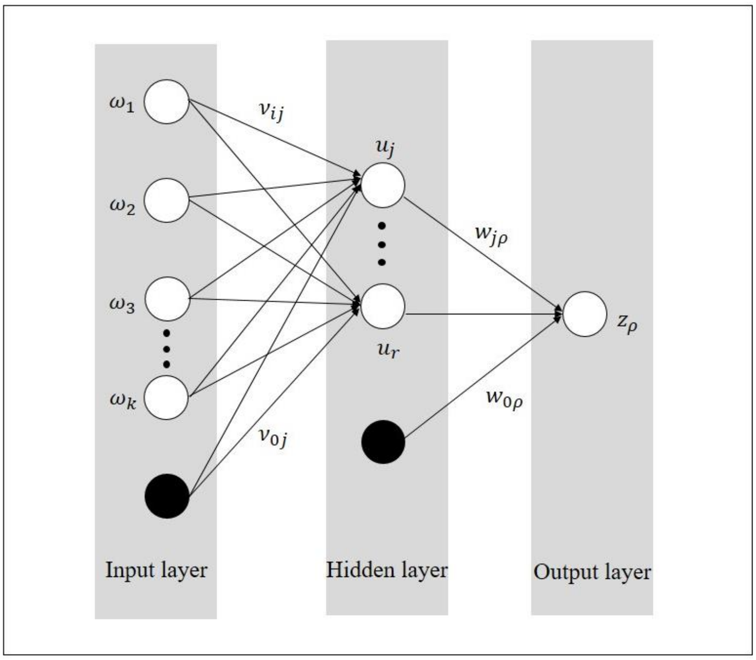


Figure 10 Multilayer perceptron neural network

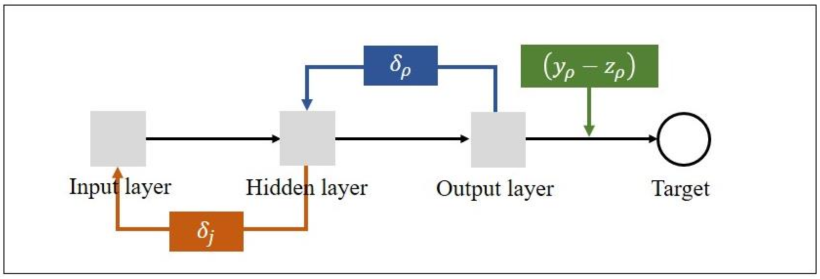


Figure 11 Architecture of backpropagation neural network

Table 4 The definitions of variables for the multilayer perceptron neural network

ω_i	The ith node of an input layer where $i = 1,2,\ldots,k$
v_{ij}	The weight from the ith input node to the jth hidden node
v_{0j}	The bias value from the input layer to the jth hidden node
u_j	The aggregate value of the jth hidden node where $j = 1,2,\ldots,r$
λ_j	The result from using the activation function of the jth hidden node
$w_{j\rho}$	The weight from the jth hidden node to the ρth output node
$w_{0\rho}$	The bias value from the hidden layer to the ρth output node
h_ρ	The aggregate value of the ρth output node where $\rho = 1,2,\ldots,m$
z_ρ	The ρth output node from using activation function of the ρth hidden node
y_ρ	The actual value where $\rho = 1,2,\ldots,m$
δ_ρ	The error between the output layer and the hidden layer
δ_j	The error between the hidden layer and the input layer
$f'(S)$	The derivative of the sum of aggregate values in a neural network model
e^q	The error of each row of data
k	The total number of input nodes
r	The total number of nodes in the hidden layer
m	The total number of output nodes

The process of backpropagation neural network using the defined variables as shown in Table 4 is as follows:

(1) Determine the numbers of nodes in the input layer, the hidden layers, and the output layer which are expected to be suitable for the data set of each problem depending upon the number of independent variables and the number of dependent variables.

(2) Determine a learning rate parameter (η) within the interval of $[0, 1]$.

(3) Randomly select the initial weight for every connection between two consecutive layers in the range of $[-1, 1]$.

(4) Determine the maximum number of iterations to be used in learning (R) and the acceptable error (δ) for stopping condition of the process.

(5) Enter input data into MLP.

(6) Calculate the value of each node in the first hidden layer according to the method of feed forward neural network as shown in Figure 12; i.e., the value of the jth hidden node is calculated from the sum of the multiplications of input data and the weights using sigmoid function as the activation function from Equations (3.10) and (3.11), respectively.

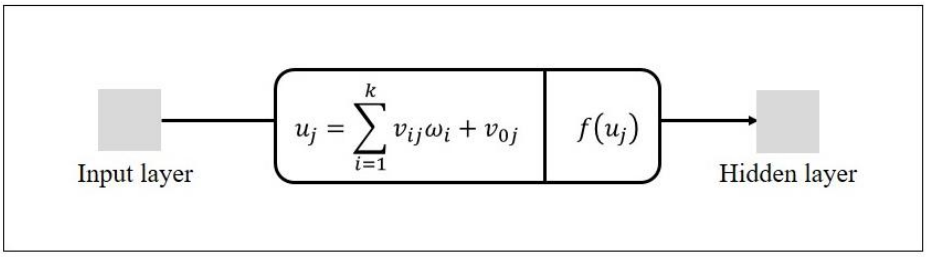


Figure 12 Architecture of feedforward neural network from the input layer to the hidden layer

$$u_j = \sum_{i=1}^{k} v_{ij}\omega_i + v_{0j} \tag{3.10}$$

$$\lambda_j = f(u_j) = \frac{1}{1+e^{-u_j}} \tag{3.11}$$

for $i = 1,2,\dots,k$ and $j = 1,2,\dots,r$

(7) Calculate the value of each node in the output layer according to the method of feed forward neural network as shown in Figure 13; i.e., the value of the ρth output node is calculated from the sum of the multiplications of hidden node values and the weights using the sigmoid function as the activation function from Equations (3.12) and (3.13), respectively.

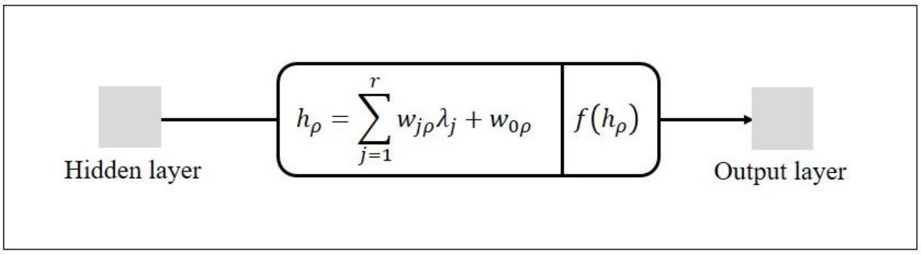


Figure 13 Architecture of feedforward neural network from the hidden layer to the output layer

$$h_\rho = \sum_{j=1}^{r} w_{j\rho}\lambda_j + w_{0\rho} \tag{3.12}$$

$$z_\rho = f(h_\rho) = \frac{1}{1+e^{-h_\rho}} \tag{3.13}$$

for $j = 1,2,\dots,r$ and $\rho = 1,2,\dots,m$

(8) Calculate the error of each node in the output layer as shown in Figure 14; i.e., the error of the ρth output node is calculated from the multiplications of the difference between the actual value and the ρth output value and the derivative of the ρth output node as expressed in Equation (3.14). This error is used to adjust the weight $w_{j\rho}$ in Equation (3.12).

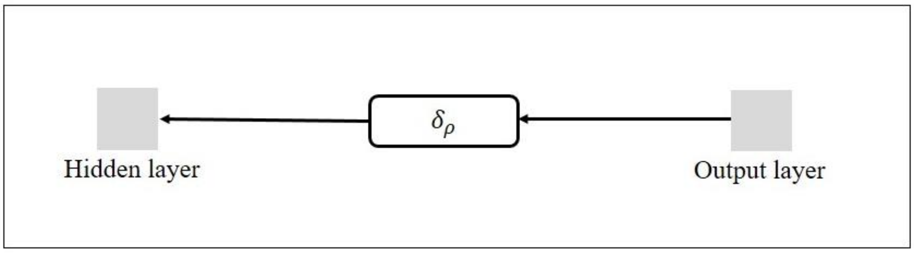


Figure 14 Architecture of error estimation for backpropagation neural network

$$\delta_\rho = (y_\rho - z_\rho) \cdot f'(h_\rho) \tag{3.14}$$

(9) Calculate the value of δ_j for each node in the hidden layer and use this value to adjust the weight of each node in the input layer as shown in Figure 15; i.e., the value of δ_j is calculated from Equation (3.15).

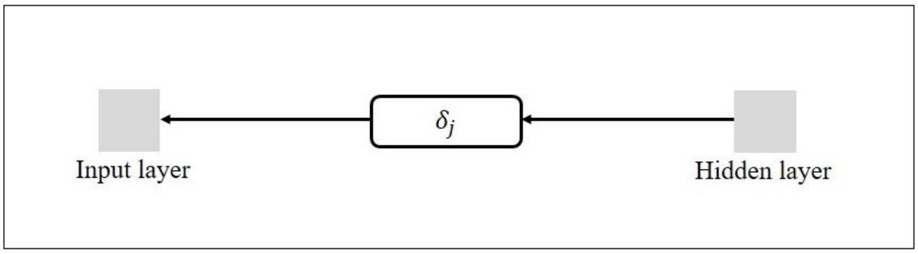


Figure 15 Architecture of weight adjustment in backpropagation neural network

$$\delta_j = \sum_{j=1}^{r} (\delta_\rho \cdot w_{j\rho}) f'(u_j) \tag{3.15}$$

(10) Adjust the weight for each input node in the $(q+1)^{th}$ iteration using Equation (3.16) and adjust the weight for each node in the hidden layer using Equation (3.17).

$$v_{ij}^{(q+1)} = v_{ij}^{(q)} + \eta \cdot \delta_j \cdot \omega_i \tag{3.16}$$

$$w_{j\rho}^{(q+1)} = w_{j\rho}^{(q)} + \eta \cdot \delta_\rho \cdot z_\rho \tag{3.17}$$

(11) Calculate the average error as shown in Equation (3.18).

$$e^q = \frac{1}{2} \sum_{\rho=1}^{m} (y_\rho - z_\rho)^2 \tag{3.18}$$

(12) Calculate the mean squared error (MSE) as shown in Equation (3.19). The value of MSE is used to verify that the error in each iteration is less than the acceptable error.

$$\mathrm{MSE} = \frac{1}{Q}\sum_{q=1}^{Q} e^{q} \tag{3.19}$$

where Q is the maximum number of iteration. If MSE is less than the acceptable error then the process stops. If MSE is greater than the acceptable error then the process repeats steps (5)–(12) until it reaches a maximum number of iterations or MSE is less than the acceptable error.

3. Support vector regression

The learning algorithm of support vector machine (SVM) can be divided into support vector classification (SVC) and support vector regression (SVR). SVC is a potential method for classifying data into two or more classes; whereas, SVR is an approximation method that estimates or predicts a real value from the given data known as support vectors.

In regression problems, the expected output is the prediction of a real value. The idea of SVR can be applied to these problems. Suppose our training data are represented as $\{(x_1, y_1), \dots, (x_l, y_l)\} \subseteq \mathbb{R}^n \times \mathbb{R}$, where $x_i \in \mathbb{R}^n$ is a vector of values of the input space, n is the dimension of input space, and $y_i \in \mathbb{R}$ is the value of the output space. A linear function, $f(x)$, is defined by

$$f(x) = \langle w, x \rangle + b \quad \text{for } w \in \mathbb{R}^n, b \in \mathbb{R} \tag{3.20}$$

where $\langle \cdot,\cdot \rangle$ denotes the dot product in $\mathbb{R}^n$ and $\|w\|^2 = \langle w, w \rangle$ and b is a constant value.

According to Vapnik in 1995 [30], our aim is to discover a function $f(x)$ for SVR that has a maximum deviation, ε, from the actual data (y_i) for all training data and is as flat as possible at the same time. In other words, as long as the deviations are less than ε, we do not care about errors, and we do not consider any deviation that is greater than ε. The loss function for SVR in this research is ε-insensitive loss function, $|\xi|_\varepsilon$, which can be expressed by

$$|\xi|_\varepsilon = |f(x) - y|_\varepsilon := \begin{cases} 0 & \text{if } |f(x) - y| \leq \varepsilon \\ |f(x) - y| - \varepsilon & \text{otherwise} \end{cases} \tag{3.21}$$

To solve the problem, the convex optimization, as shown in Equation (3.22), can be used. The convex optimization problem is feasible by adapting the use of soft margin loss function to SVR, adding slack variables (ξ_i, ξ_i^*) to cope with other infeasible constraints for the optimization problem.

$$\begin{aligned} &\text{minimize} \quad \left\{\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)\right\} \\ &\text{subject to} \quad y_i - \langle w, x_i\rangle - b \leq \varepsilon + \xi_i \\ &\qquad \langle w, x_i\rangle + b - y_i \leq \varepsilon + \xi_i^* \\ &\qquad \xi_i, \xi_i^* \geq 0 \end{aligned} \tag{3.22}$$

The constant C defines the trade-off between the flatness of $f(x)$ and the deviation value up to where the deviations greater than $\mathcal{E}$ are tolerated. Figure 16 shows an example of the soft margin liner SVR.

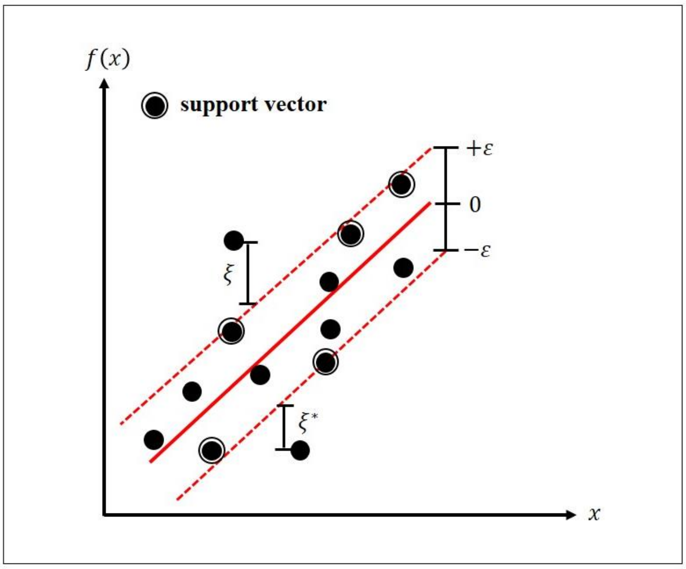


Figure 16 The soft margin liner SVR

According to Mangasarian in 1969 [31], McCormick in 1983 [32], and Vanderbei in 1997 [33], the construction of Lagrange function from an objective function, which is called the primary objective function and the corresponding constraints, is done by adding a set of dual variables. This function is shown to have a saddle point with respect to the solution's primal and dual variables which is proceeded as follows:

$$L(w,b,\xi_i,\xi_i^*) := \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i+\xi_i^*) - \sum_{i=1}^{l}(\eta_i\xi_i+\eta_i^*\xi_i^*) \right.$$

$$-\sum_{i=1}^{l}\alpha_i(\varepsilon+\xi_i-y_i+\langle w,x_i\rangle+b)$$

$$\left. -\sum_{i=1}^{l}\alpha_i^*(\varepsilon+\xi_i^*+y_i-\langle w,x_i\rangle-b) \right\} \tag{3.23}$$

subject to $\eta_i,\eta_i^*,\alpha_i,\alpha_i^* \geq 0$

where L is the Lagrange and $\eta_i,\eta_i^*,\alpha_i,\alpha_i^*$ are the Lagrange multipliers. Thus the partial derivatives of L with respect to the primal variables $b,\ w,\ \xi_i,\ \xi_i^*$ can be calculated from Equations (3.24) – (3.27) and set them equal to 0 for optimality.

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l}(\alpha_i^*-\alpha_i) = 0 \tag{3.24}$$

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\,x_i = 0 \tag{3.25}$$

$$\frac{\partial L}{\partial \xi_i} = C - \sum_{i=1}^{l}\eta_i - \sum_{i=1}^{l}\alpha_i = 0 \tag{3.26}$$

$$\frac{\partial L}{\partial \xi_i^*} = C - \sum_{i=1}^{l}\eta_i^* - \sum_{i=1}^{l}\alpha_i^* = 0 \tag{3.27}$$

Substituting Equation (3.24) to Equation (3.27) into Equation (3.23) yields the dual optimization problem as follows:

$$\text{maximize} \quad \left\{ -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j\rangle - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \right\} \tag{3.28}$$

$$\text{subject to} \quad \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \ \text{ and } \ 0 \le \alpha_i, \alpha_i^* \le C$$

Substituting Equation (3.24) into Equation (3.20), we can rewrite as follows:

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\,\langle x_i, x\rangle + b \tag{3.29}$$

where $b = -\frac{1}{2}\sum_{i=1}^{l}(\alpha_i - \alpha_i^*)[\langle x_i, x_r\rangle + \langle x_i, x_s\rangle]$, and x_r and x_s are support vectors at the top and the bottom planes, respectively.

For nonlinear problems, the dimension of a feature space is always high, thus it is impossible to compute the inner product directly in a feature space. Kernel function is a key step for SVR. Kernel function could be achieved by mapping from an input space into a feature space that is

$$K(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j)\rangle \tag{3.30}$$

Table 5 shows some of the kernel functions. It suffices to know $K(x_i, x_j)$ rather than Φ explicitly which allows us to restate the SVR optimization problem as:

$$\text{maximize} \quad \left\{ -\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \Phi(x_i), \Phi(x_j)\rangle - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{l} y_i(\alpha_i - \alpha_i^*) \right\} \tag{3.31}$$

$$\text{subject to} \quad \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \ \text{ and } 0 \le \alpha_i, \alpha_i^* \le C$$

Likewise, the expansion of Equation (3.29) can be written as

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) \langle \Phi(x_i), \Phi(x) \rangle + b \tag{3.32}$$

where $b = -\frac{1}{2}\sum_{i=1}^{l}(\alpha_i - \alpha_i^*)[\langle \Phi(x_i), \Phi(x_r)\rangle + \langle \Phi(x_i), \Phi(x_s)\rangle]$, and x_r and x_s are support vectors at the top and the bottom planes, respectively.

Table 5 shows examples of kernel functions that are widely used for SVR but the kernel function that is suitable for this study is a linear function according to the data.

Table 5 Examples of kernel functions

Kernel Functions	**Formula**
Linear	$K(x_i, x_j) = \langle x_i \cdot x_j \rangle$
Polynomial	$K(x_i, x_j) = \left(1 + \langle x_i \cdot x_j \rangle\right)^d$ where d is the degree of the polynomial
Gaussian kernel	$K(x_i, x_j) = e^{\left(-\frac{\lVert x_i - x_j \rVert^2}{2\sigma^2}\right)}$ where σ is the constant
Gaussian radial basis function	$K(x_i, x_j) = e^{-\gamma \lVert x_i - x_j \rVert^2}$ where $\gamma > 0$
Multi-quadratic	$K(x_i, x_j) = -\sqrt{\lVert x_i - x_j \rVert^2 + c^2}$ where c is the constant
Thin plate spline	$K(x_i, x_j) = \lVert x_i - x_j \rVert^2 \ln \lVert x_i - x_j \rVert$
Moderate decreasing	$K(x_i, x_j) = ke^{\left(\frac{\gamma}{\lVert x_i - x_j \rVert^2 + \sigma^2}\right)^{-1}}$ where k and σ are the constant

4. Modified depth first search

Depth first search (DFS) offers a way of graph traversal to make it possible to conduct the searches for the next vertex of a graph by moving deeper along a path which begins at any chosen vertex, known as the parent vertex, and passes through all other vertices on the path until no further vertices exist. If there is any vertex which has not yet been visited, DFS algorithm will return to the vertex from which other possible paths can be taken. This process is repeated until all of the vertices have been visited. Whenever there is a choice of paths, the algorithm will initially select the path that offers the shortest distance between the current and subsequent vertices [34, 35].

On the basis of DFS, this study presents a novel algorithm called the modified depth-first search (MDFS) which can be used to select the next vertex to visit based on the root mean square error. In this research, each vertex of a graph represents a station where PM_{10} and other data are measured and collected. These data are used to predict PM_{10} concentration. The concept of the prediction model is to include data from the nearby stations that influence PM_{10} concentration at the station of interest. MDFS selects the next station to be visited by calculating the RMSE from Equation (3.33) which is the mean difference between the observed value and the predicted value. The algorithm selects the station with the minimum RMSE. MDFS is not the same as DFS because MDFS never backtracks. It only stops when there is no more station to visit. The advantage of MDFS is that it can select the neighboring stations whose data can affect the prediction, even when the wind speed or the direction are not known. The process of MDFS algorithm could be described as follows:

(1) Select the station of interest which is the start vertex of a graph from any stations and specify the initial radius that defines the region of neighboring stations to be considered.

(2) For each of the neighboring stations within the region specified by the radius:

2.1 Combine data from a neighboring station with the data from the station of interest.

2.2 Use the combined data from 2.1 as the input to the model.

2.3 Calculate the RMSE from the prediction result.

2.4 Use the minimum RMSE as the mean to select the next station to visit.

(3) Set the station which has the lowest RMSE as the new station of interest and increase the radius by a constant value.

(4) Repeat step (2) – (3) until no more station to visit.

3.2.3 Performance evaluation

Several statistical models had been implemented for prediction models in this research. To evaluate the performance of the models, the predictive values obtained from those statistical models are compared with the observed data. Four standard statistical models are used as assessment criteria, namely the root mean square error (RMSE) [36, 37], the Pearson correlation coefficient (R^2) [38], the mean absolute error (MAE) [39, 40], and the mean absolute percentage error (MAPE) [41]. RMSE represents the model's error, which can be determined by

$$\mathrm{RMSE} = \sqrt{\frac{1}{T}\sum_{j=1}^{T}\left(y_j - \hat{y}_j\right)^2} \tag{3.33}$$

where $\hat{y}_j$ and y_j are the predicted value and the observed value for $j = 1, \ldots, T$, and T is the number of data to be predicted. RMSE can measure goodness-of-fit and describes the predicted average error. The next statistical model that we use to find the relationship between the predicted values and the observed values is the Pearson correlation coefficient (R^2) which can be calculated by

$$\mathrm{R}^2 = \frac{\sum_{j=1}^{T}\left(\hat{y}_j - \mu_{\hat{y}}\right)\left(y_j - \mu_y\right)}{\sqrt{\sum_{j=1}^{T}\left(\hat{y}_j - \mu_{\hat{y}}\right)^2}\sqrt{\sum_{j=1}^{T}\left(y_j - \mu_y\right)^2}} \tag{3.34}$$

where $\mu_{\hat{y}}$ and μ_y are the average predicted value and the average observed value.

MAPE and MAE have been used to show the dispersion average between the observed values and the predicted values which can be described as

$$\text{MAPE} = \frac{100}{T}\sum_{j=1}^{T}\left|\frac{y_j - \hat{y}_j}{y_j}\right| \tag{3.35}$$

$$\text{MAE} = \frac{1}{T}\sum_{j=1}^{T}\left|y_j - \hat{y}_j\right| \tag{3.36}$$

To determine the prediction results from MAPE and MAE, smaller values of MAPE and MAE yield better results.

CHAPTER 4 METHODOLOGY

In this chapter, the methodology is divided into two main parts: data preprocessing and PM_{10} concentration prediction. Data preprocessing is the first part to prepare data for prediction which can be divided into two parts: data collection and normalization. The next part is to predict PM_{10} concentration using the proposed method. The methodology of each part is shown in Figure 17 and the details are explained below.

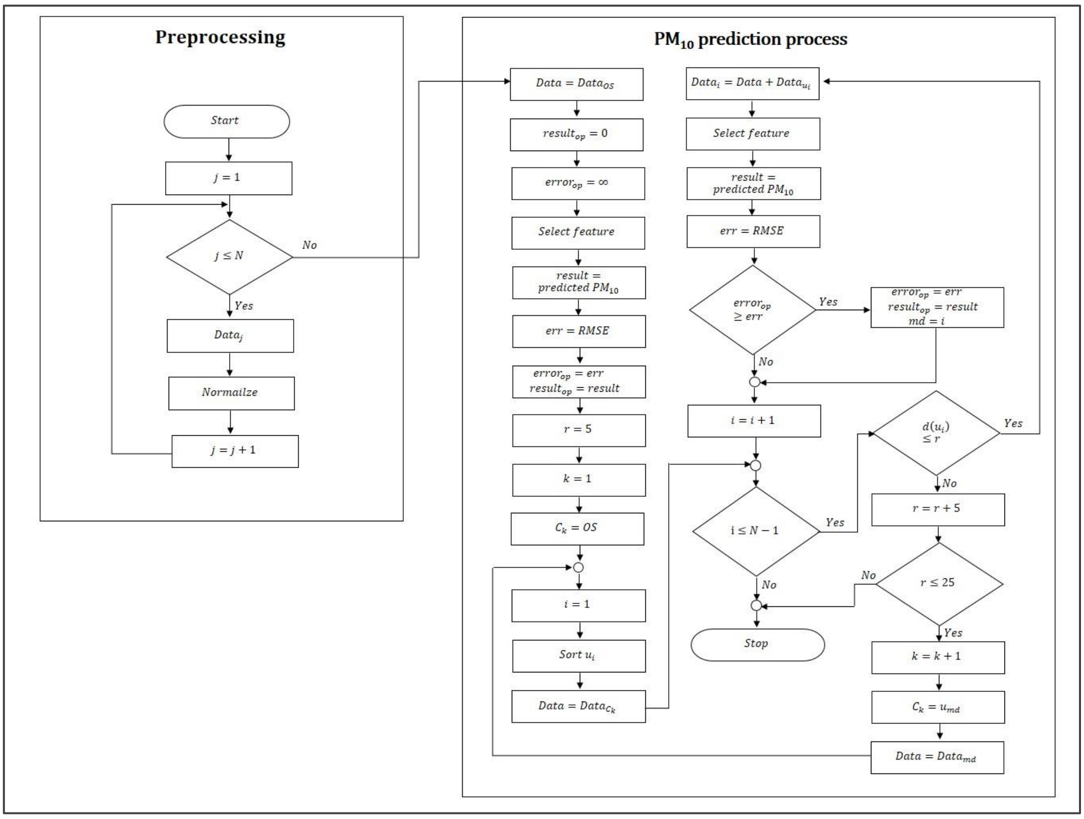

Preprocessing
Start
j = 1
j ≤ N
No
Yes
Data_j
Normailze
j = j + 1
PM10 prediction process
Data = Data_OS
result_op = 0
error_op = ∞
Select feature
result = predicted PM10
err = RMSE
error_op = err
result_op = result
r = 5
k = 1
C_k = OS
i = 1
Sort u_i
Data = Data_Ck
Data_i = Data + Data_ui
Select feature
result = predicted PM10
err = RMSE
error_op ≥ err
Yes
error_op = err
result_op = result
md = i
No
i = i + 1
i ≤ N − 1
Yes
No
Stop
d(u_i) ≤ r
Yes
No
r = r + 5
r ≤ 25
No
Yes
k = k + 1
C_k = u_md
Data = Data_md

Figure 17 Overview and the flowchart of the proposed method

4.1 Data Preprocessing

The evaluation of the proposed method is carried out using data derived from the sensor network system in northern region of Thailand which monitors the region's air quality. This study uses data obtained from Nan province since air pollution problem is a serious issue in this area as a consequence of burning the corn fields at the end of every harvesting season. This causes PM_{10} concentration to be high.

The location of Nan province in Thailand can be seen in Figure 18, while Figure 19 presents the effects of the burning process which leads to clearly visible smog across the province. The current online sensor network system provides air pollution monitoring by assessing haze levels [42]. Each of the monitoring stations has a sensor which measures the air quality data and the air pollutant data before sending the data to the Cloud for storage. There are two types of air quality recorded: humidity and temperature, and three types of air pollutants: PM_1, $PM_{2.5}$, and PM_{10}.

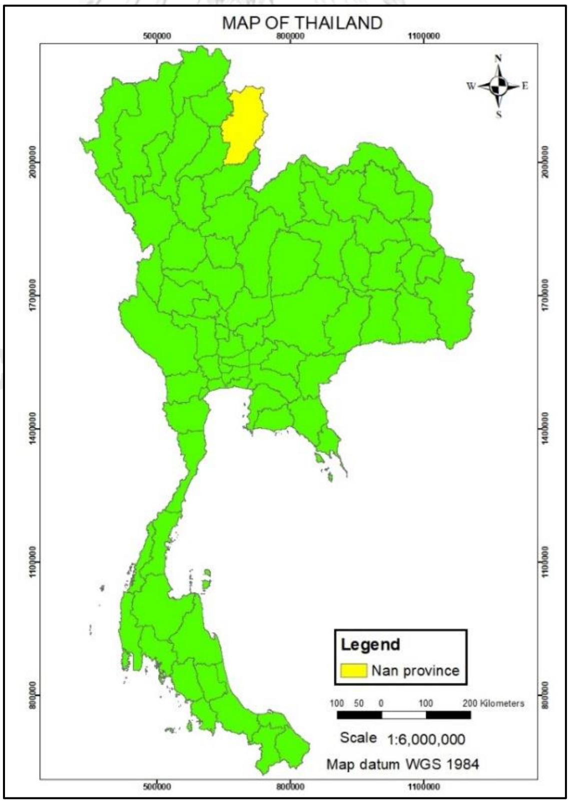


Figure 18 The location of Nan Province

Figure 19 Air pollution in Nan Province as a case study

This case study uses the data from six sensors installed at the stations around Nan province. This area is suitable for collecting data because it normally has high levels of PM_{10} pollution, and the data are generally complete. The use of only six sensors is because these sensors are located in a cluster with relatively good network connection, allowing for reliable recording of air quality data. These sensors are shown in Figure 20 as red triangles and the sensors which are not selected are represented by blue pentagons. The unselected sensors have problems with missing data and are located further away from the chosen sensors. Table 6 shows the geographical coordinates for each sensor, while Table 7 shows how far apart each of the six sensors are at various monitoring stations. The available raw data were collected every second during the period from the start of February 2017 to the end of April 2017. This short time interval between measurements leads to redundancy of data; therefore, hourly records using the averages for each hour were taken instead. Since the data were recorded every second, the average of each hour is calculated by finding the average of each minute and then finding the average of each hour from the average of 60 minutes; however, some data are missing; therefore, the denominator is equal to the actual number of data. Figure 21-26 shows the hourly data of the sensors S_1, $S_2, S_3, S_4, S_5,$ and $S_6,$ respectively. This approach permits predictions to be made for the upcoming 24-hour period, or the upcoming one-month period for any chosen monitoring station in terms of PM_{10} concentration by manipulating the available data.

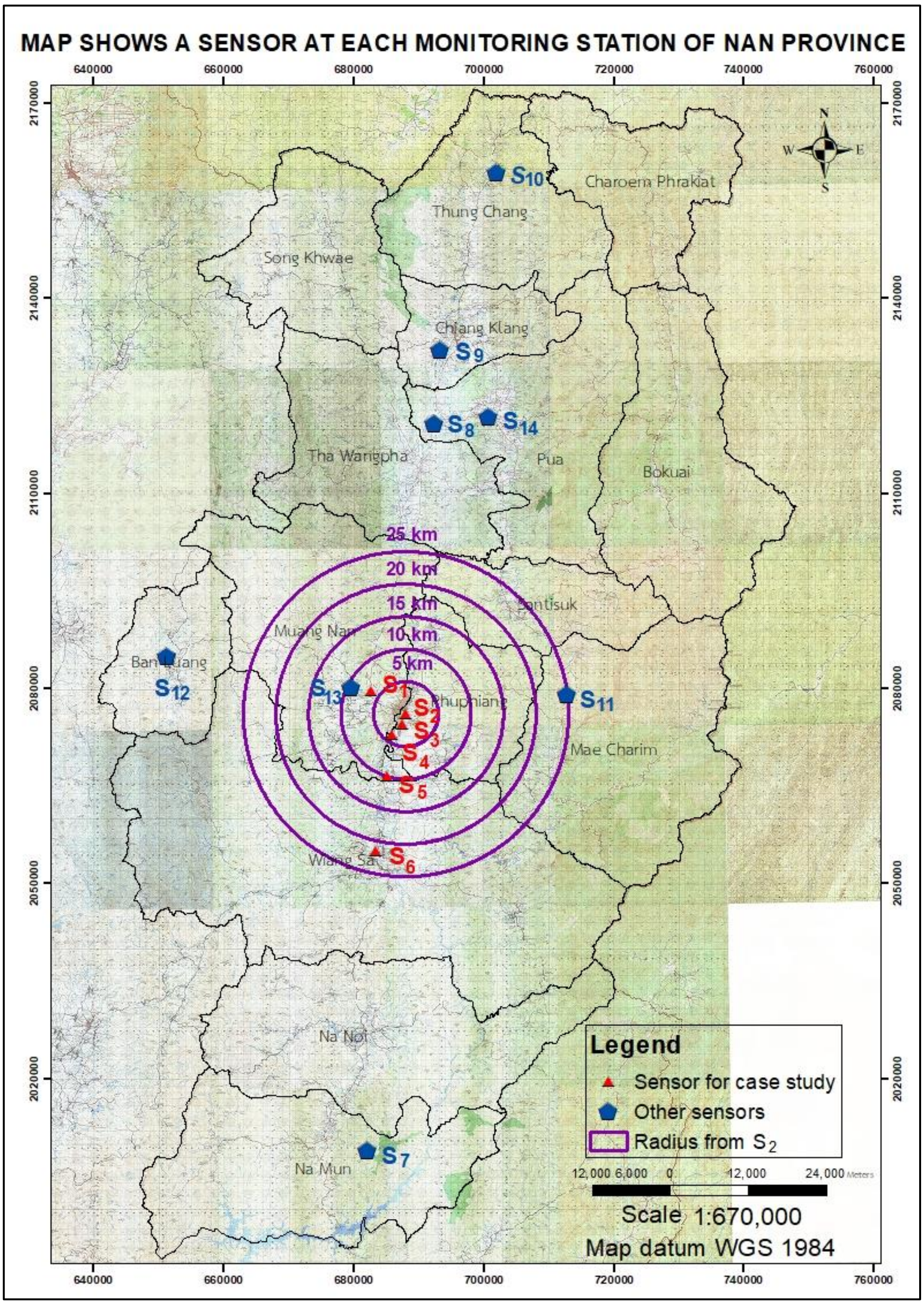

MAP SHOWS A SENSOR AT EACH MONITORING STATION OF NAN PROVINCE
640000
660000
680000
700000
720000
740000
760000
2170000
2140000
2110000
2080000
2050000
2020000
N
E
S
W
S10
Charoem Phrakiat
Thung Chang
Song Khwae
Chiang Klang
S9
S8
S14
Tha Wangpha
Pua
Bokuai
25 km
20 km
15 km
10 km
5 km
Santisuk
Muang Nan
Ban Luang
S12
S13
S1
S2
S3
S4
S5
S6
Phuphiang
S11
Mae Charim
Wiang Sa
Na Noi
S7
Na Mun
Legend
Sensor for case study
Other sensors
Radius from S2
12,000 6,000 0 12,000 24,000 Meters
Scale 1:670,000
Map datum WGS 1984

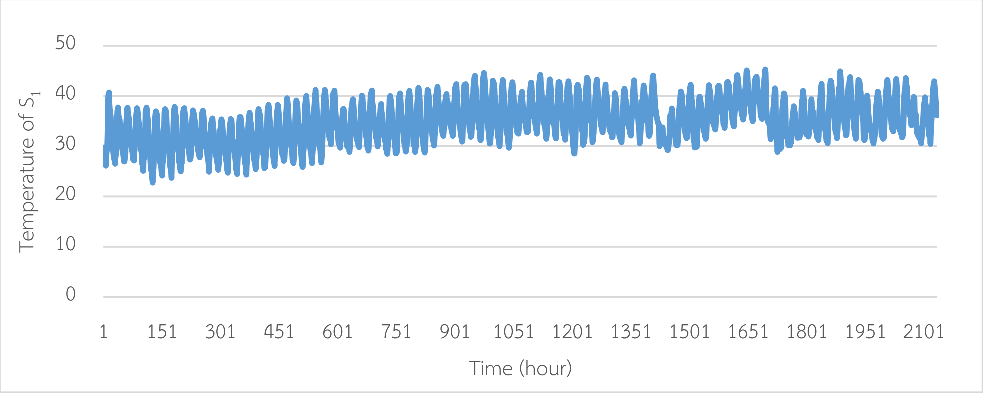
Figure 20 Positions of six sensors at monitoring stations

Table 6 The spans between latitude and longitude of the 14-sensor network system

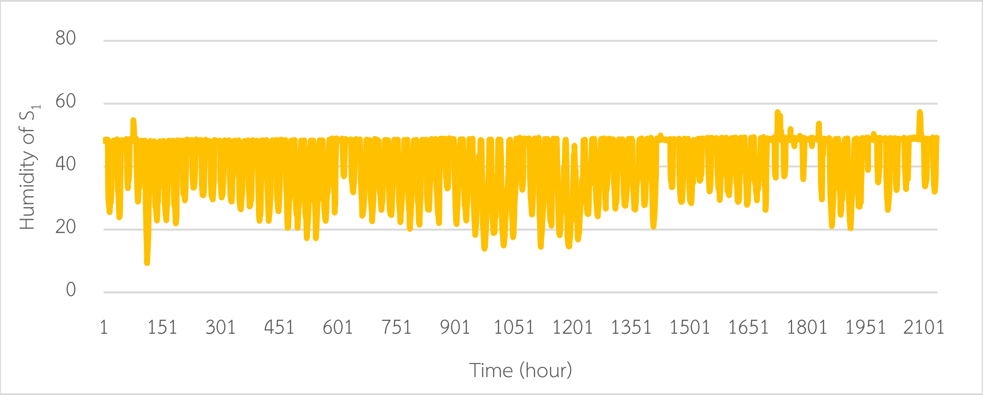
Sensor	Northern latitude		Eastern longitude	
	Deg° Min	Deg° Min' Sec"	Deg° Min	Deg° Min' Sec"
S_1	N18°47.9796	N18°47'58.776"	E100°43.998	E100°43'59.88"
S_2	N18°46.05402	N18°46'3.241"	E100°47.00214	E100°47'0.128"
S_3	N18°45.1692	N18°45'10.152"	E100°46.65	E100°46'39"
S_4	N18°44.3466	N18°44'20.796"	E100°45.81	E100°45'48.6"
S_5	N18°40.8708	N18°40'52.248"	E100°45.348	E100°45'20.88"
S_6	N18°34.626	N18°34'37.56"	E100°44.262	E100°44'15.72"
S_7	N18°9.654	N18°9'39.24"	E100°43.332	E100°43'19.92"
S_8	N19°10.1988	N19°10'11.928"	E100°49.77	E100°49'46.2"
S_9	N19°16.368	N19°16'22.08"	E100°50.382	E100°50'22.92"
S_{10}	N19°31.1028	N19°31'6.168"	E100°55.536	E100°55'32.16"
S_{11}	N18°47.5458	N18°47'32.748"	E101°1.152	E101°1'9.12"
S_{12}	N18°51.0858	N18°51'5.148"	E100°26.136	E100°26'8.16"
S_{13}	N18°48.3186	N18°48'19.116"	E100°42.264	E100°42'15.84"
S_{14}	N19°10.785	N19°10'47.1"	E100°54.552	E100°54'33.12"

Table 7 The distance between six sensors at monitoring stations

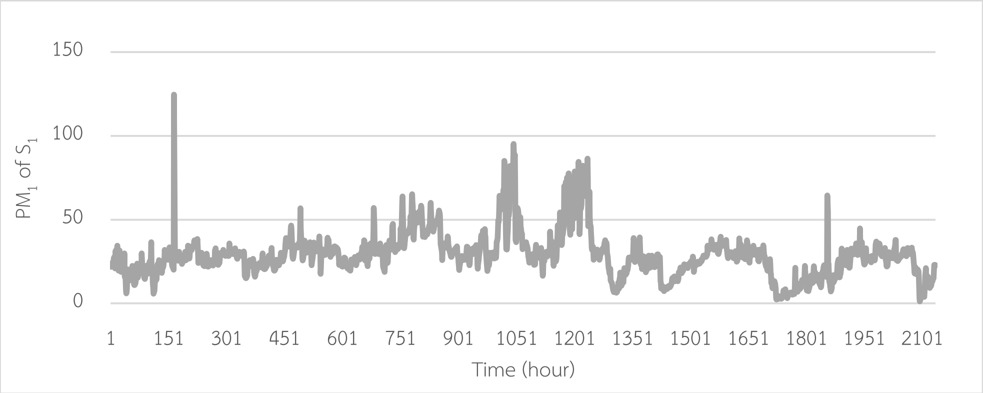
Sensor	Distance between sensor (kilometer)				
	S_2	S_3	S_4	S_5	S_6
S_1	6.365	6.984	7.446	13.39	24.75
S_2		1.752	3.793	10.03	21.72
S_3			2.121	8.287	19.98
S_4				6.492	18.22
S_5					11.73



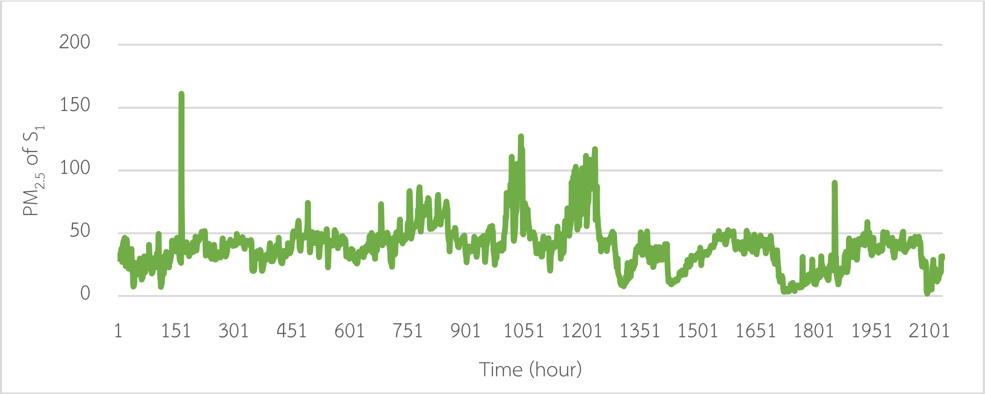
(a) The hourly data of temperature



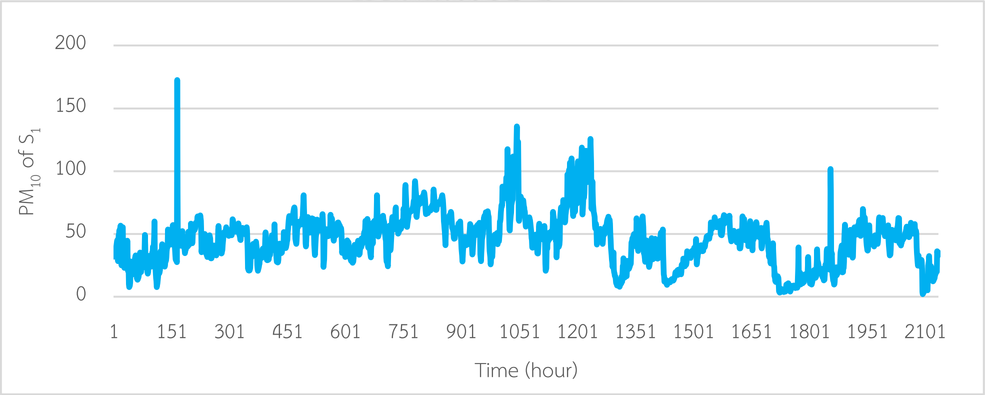
(b) The hourly data of humidity



(c) The hourly data of PM_1

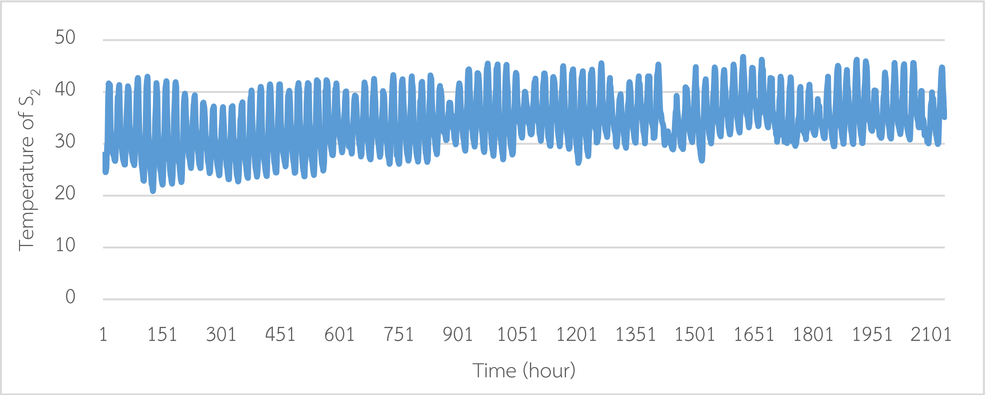


(d) The hourly data of $PM_{2.5}$

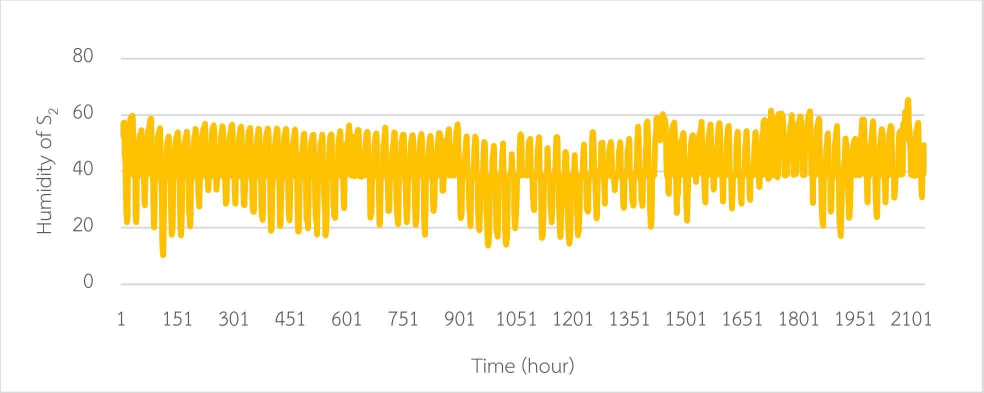


(e) The hourly data of PM_{10}

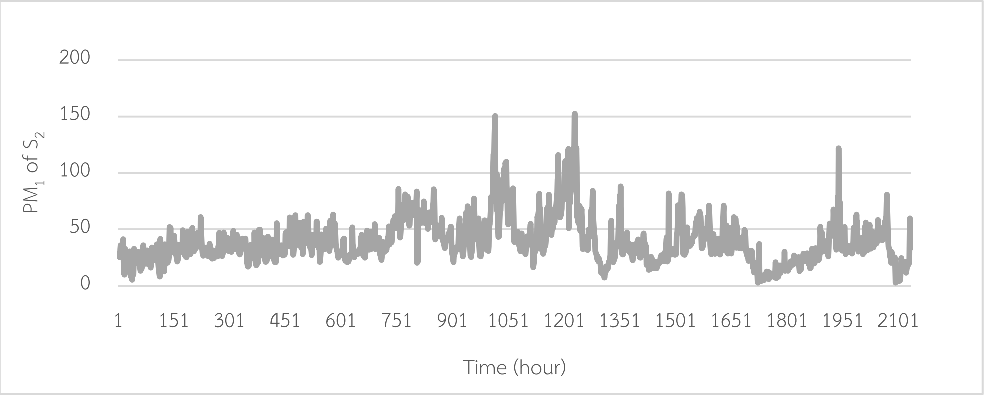
Figure 21 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_1



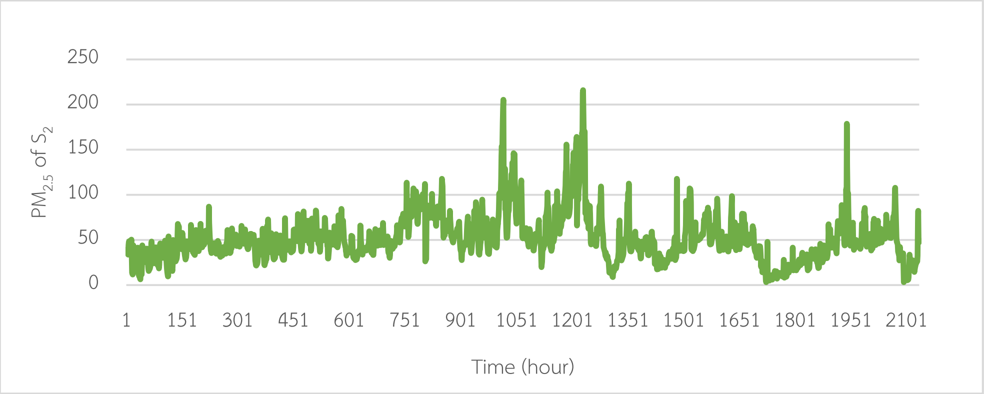
(a) The hourly data of temperature



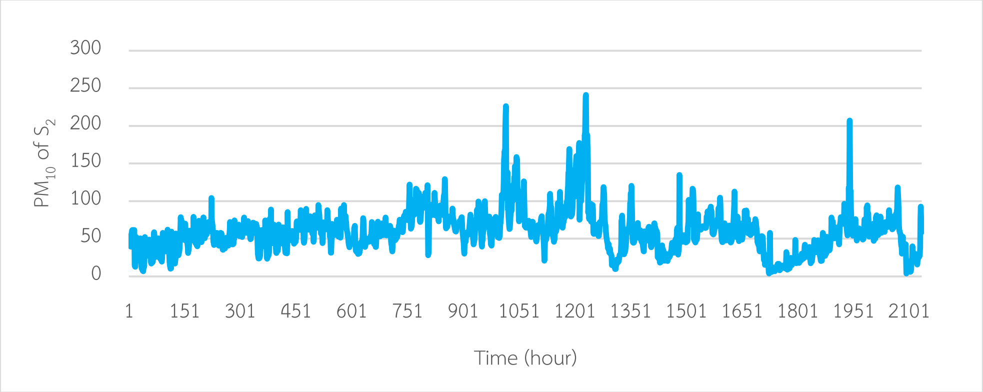
(b) The hourly data of humidity



(c) The hourly data of PM_1

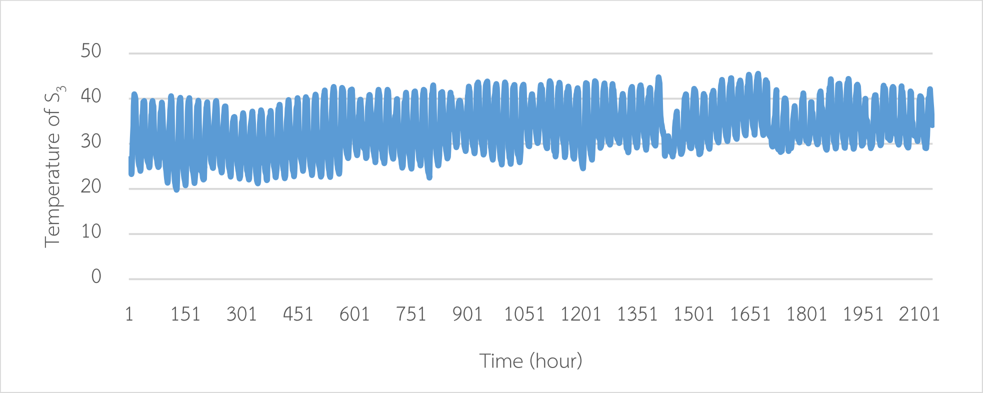


(d) The hourly data of $PM_{2.5}$

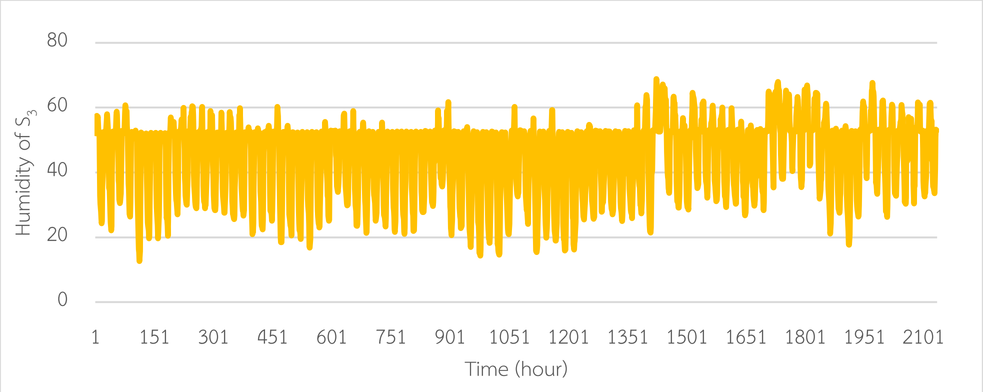


(e) The hourly data of PM_{10}

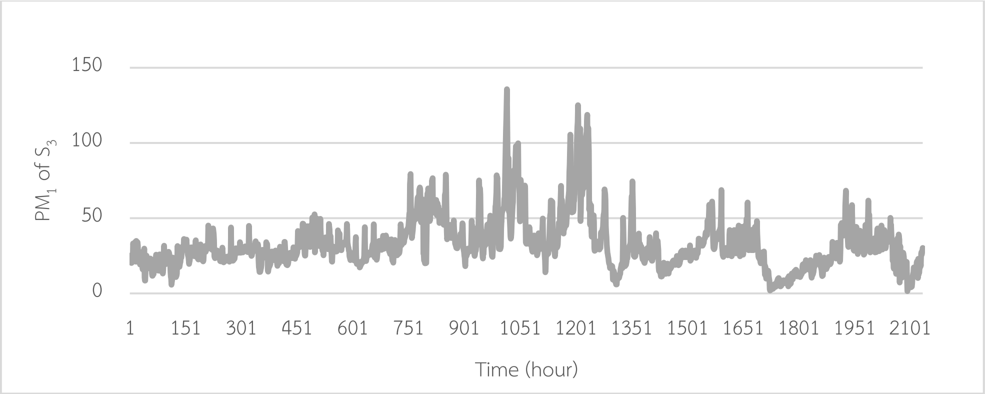
Figure 22 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_2



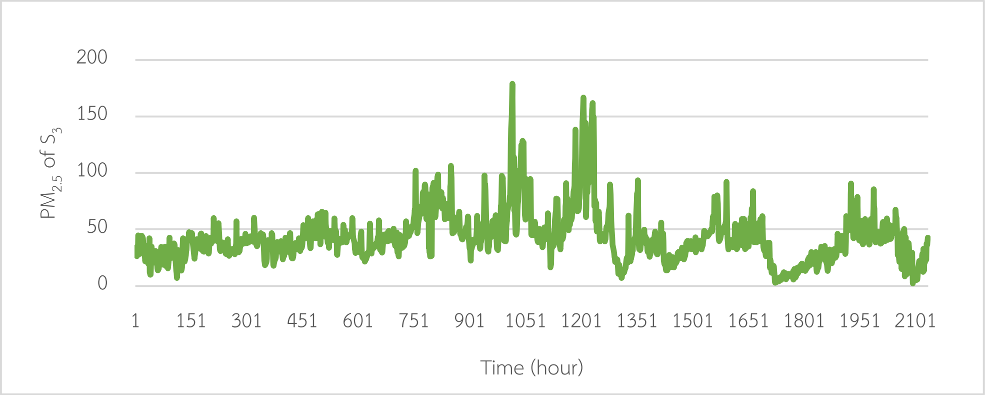
(a) The hourly data of temperature



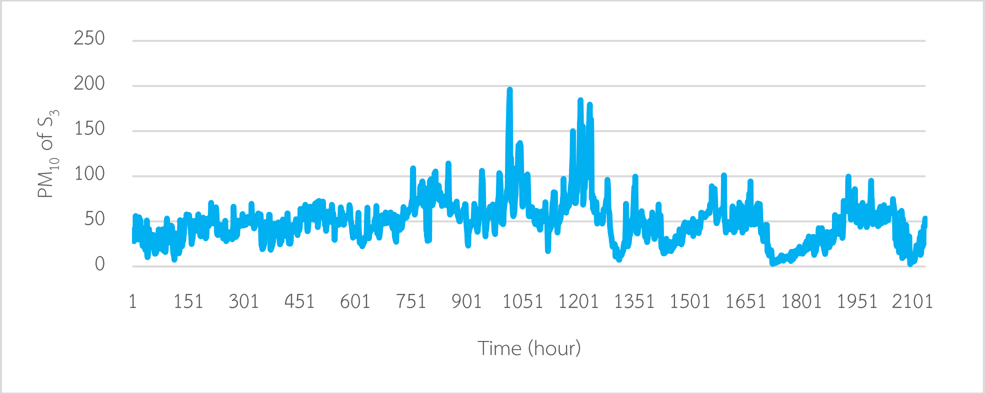
(b) The hourly data of humidity



(c) The hourly data of PM_1

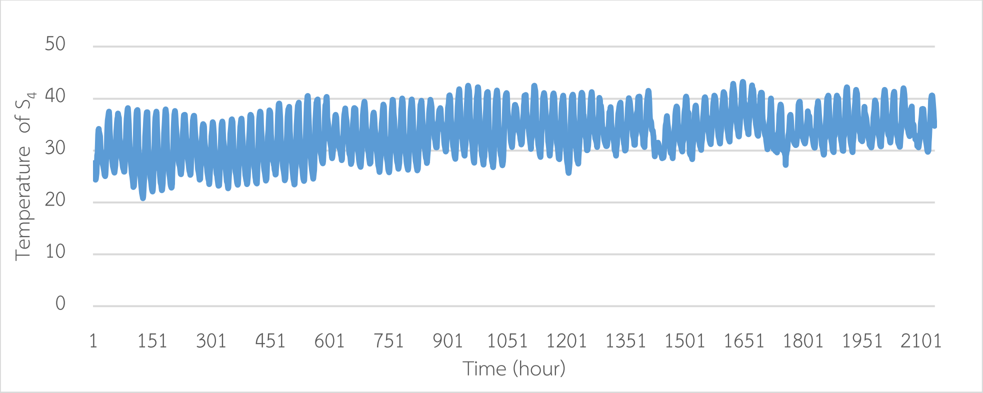


(d) The hourly data of $PM_{2.5}$

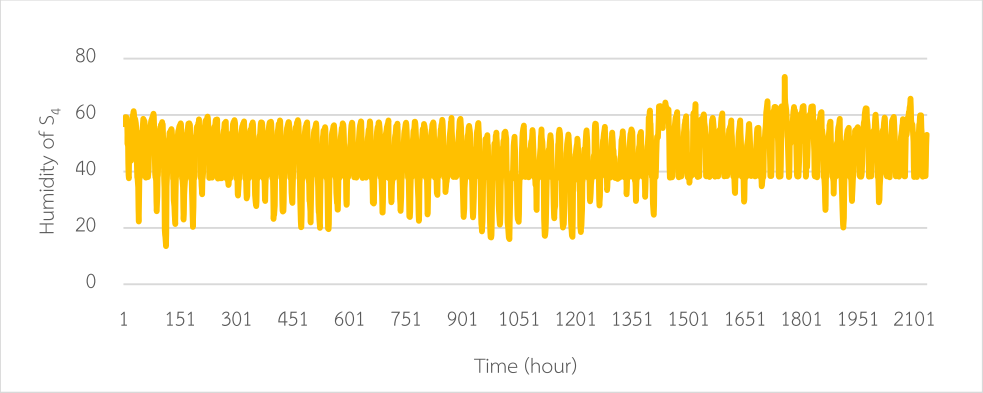


(e) The hourly data of PM_{10}

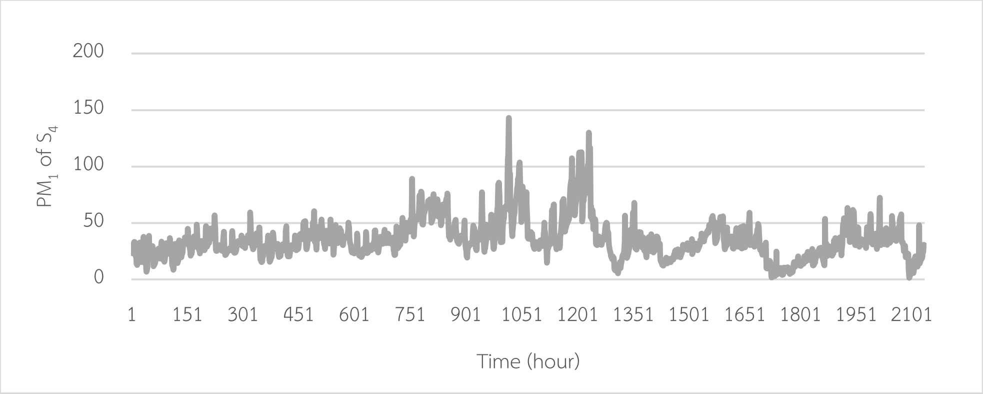
Figure 23 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_3



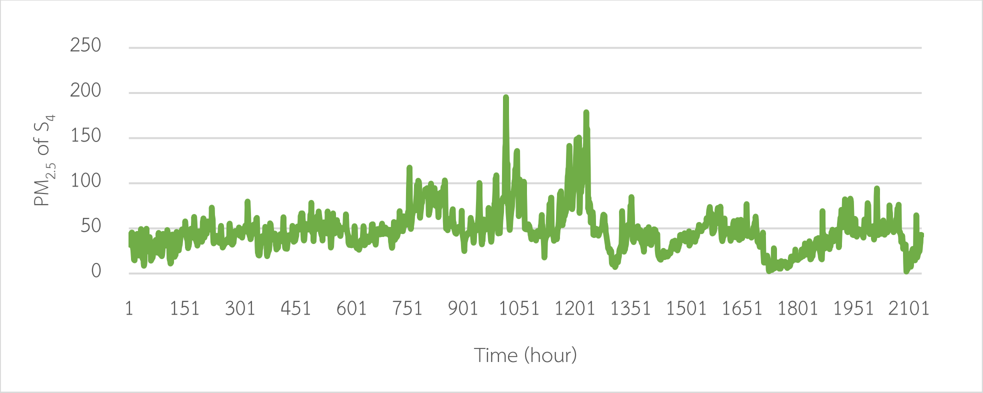
(a) The hourly data of temperature



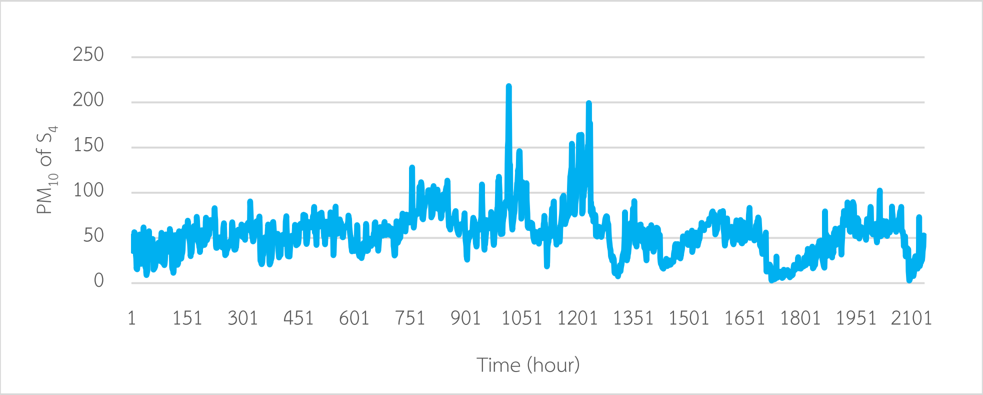
(b) The hourly data of humidity



(c) The hourly data of PM_1

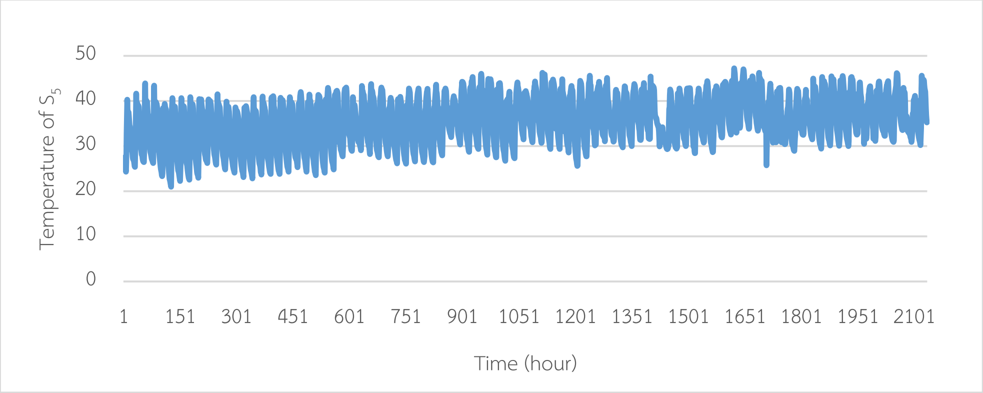


(d) The hourly data of $PM_{2.5}$

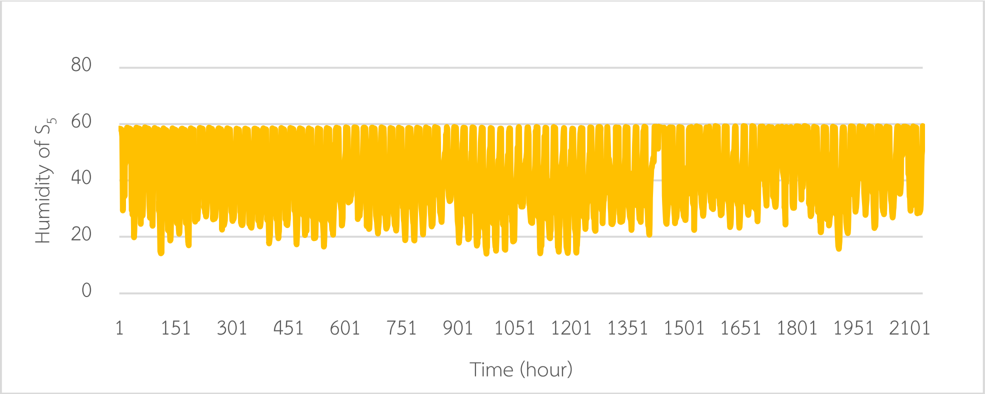


(e) The hourly data of PM_{10}

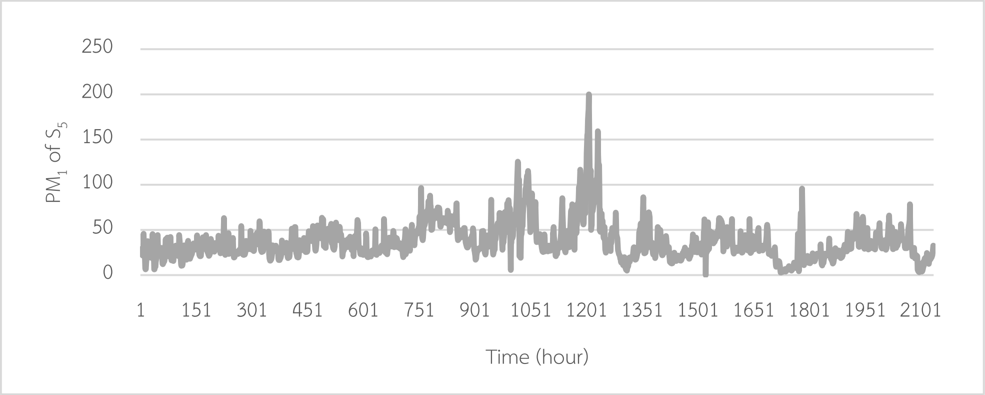
Figure 24 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_4



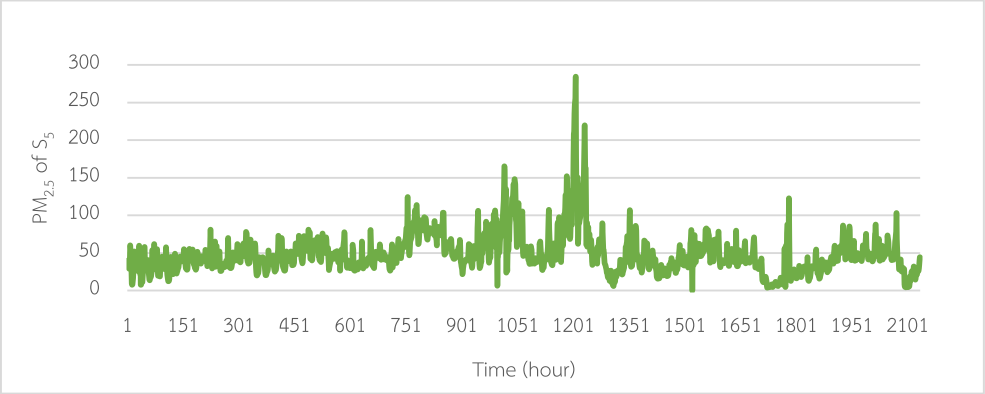
(a) The hourly data of temperature



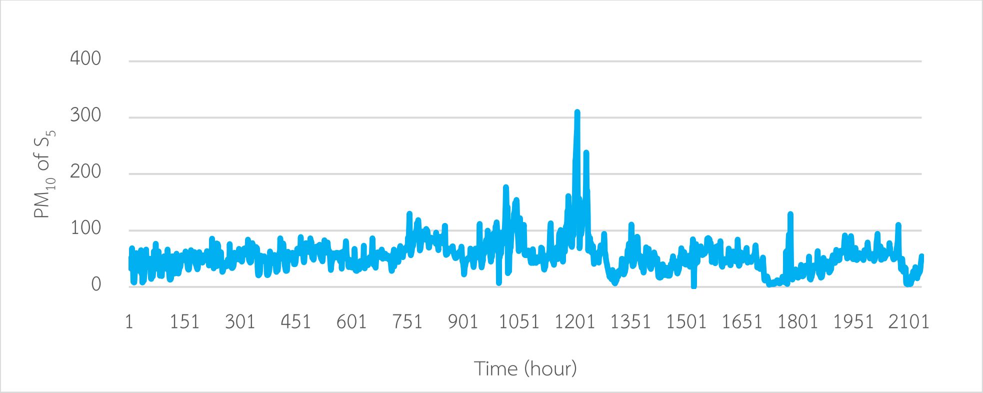
(b) The hourly data of humidity



(c) The hourly data of PM_1

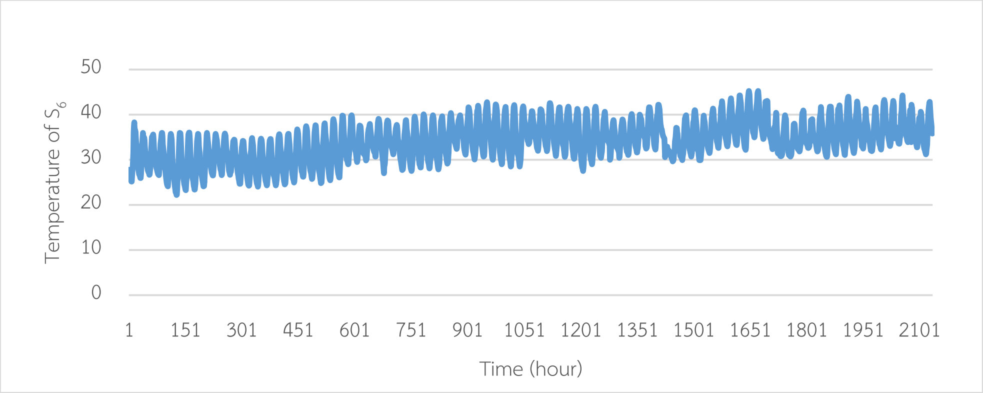


(d) The hourly data of $PM_{2.5}$

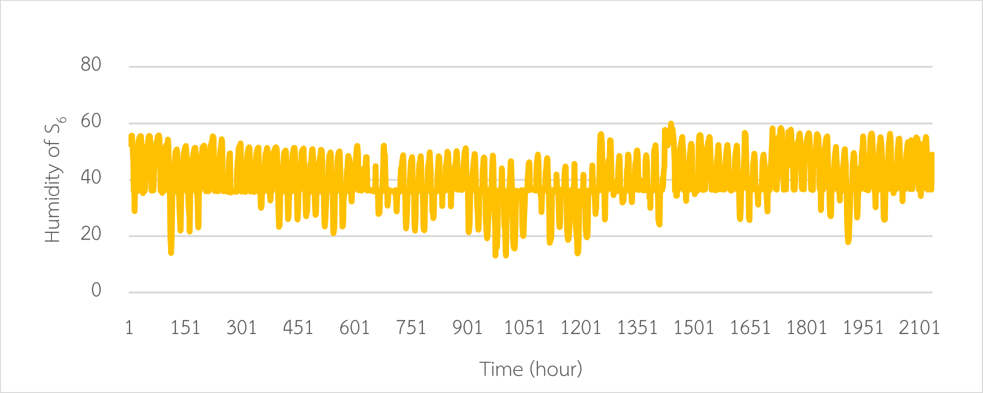


(e) The hourly data of PM_{10}

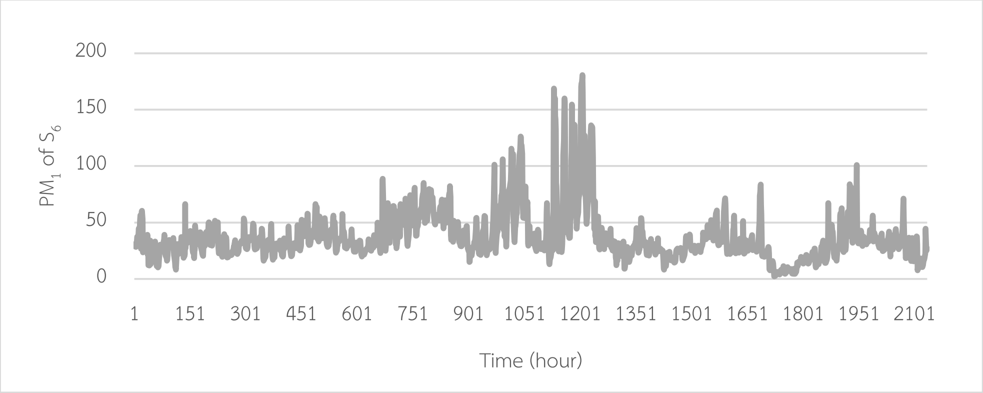
Figure 25 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_5



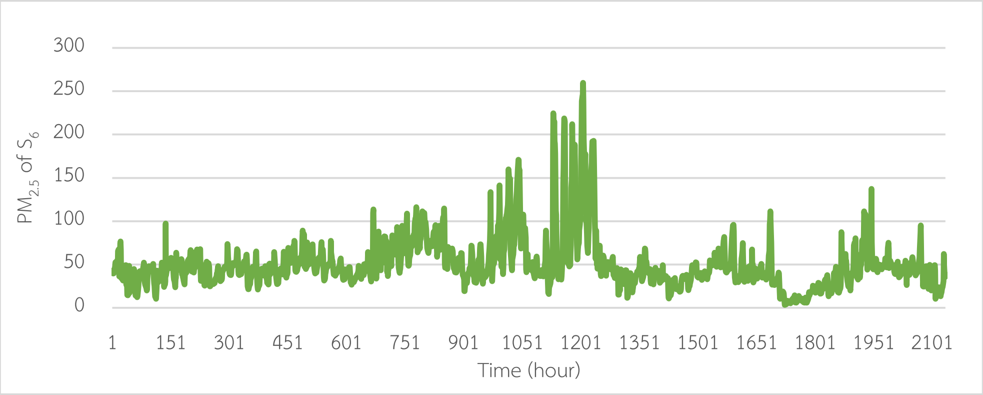
(a) The hourly data of temperature



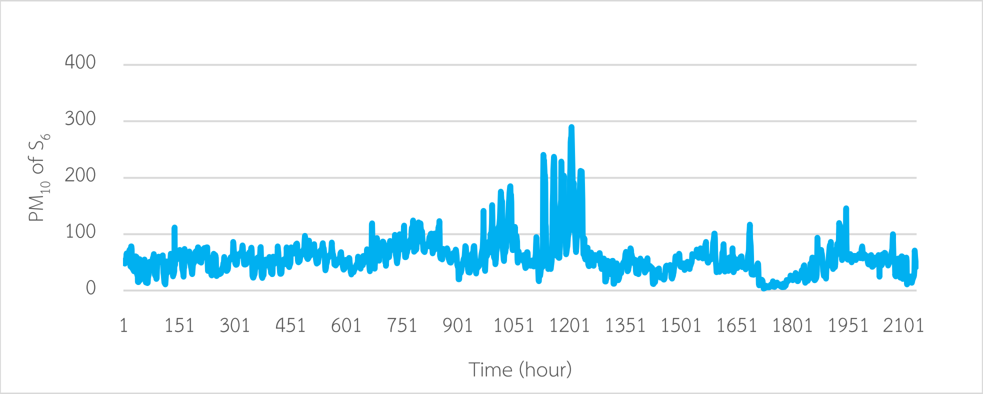
(b) The hourly data of humidity



(c) The hourly data of PM_1



(d) The hourly data of $PM_{2.5}$



(e) The hourly data of PM_{10}

Figure 26 The hourly data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_6

In accordance with the PM regulatory standards from the United States EPA, Thailand’s National Environment Board has determined that the 24-hour standard for PM_{10} should be less than 120 $\mu g/m^3$. Table 8 provides basic statistical data concerning the concentrations of PM_{10} at each of the selected monitoring stations. From these data, it can be seen that some of PM_{10} concentration values recorded at these six monitoring stations are higher than the permissible levels. The mean concentrations for six monitoring stations show little variation with a narrow band between 45-60 $\mu g/m^3$although the standard deviation at each site vary considerably, falling in the range of 18-26 $\mu g/m^3$. It is apparent that since the variances are high, the distribution of the data is relatively broad.

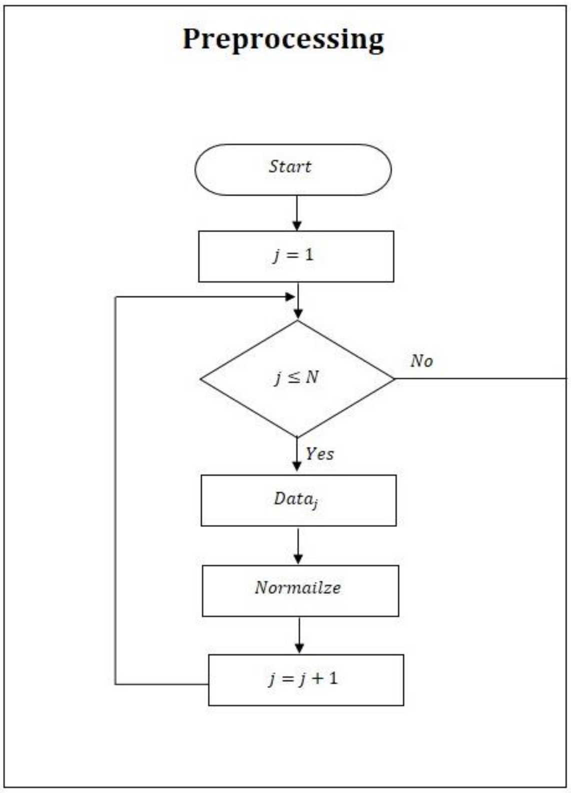
Table 8 The basic statistical values of PM_{10} concentrations for every 24 hours of each monitoring station

Sensor	Minimum	Maximum	Mean	Standard deviation
S_1	6.4043	94.4673	45.4205 (1.9201*)	18.1138
S_2	10.0237	134.8779	59.0781 (2.4878*)	23.4699
S_3	7.4013	118.0505	49.1990 (2.1751*)	20.5199
S_4	8.6913	123.5772	54.7246 (2.2844*)	21.5509
S_5	7.7190	156.5084	54.1556 (2.3812*)	22.4638
S_6	7.9530	162.4158	57.1958 (2.6623*)	25.1162

*Standard error

Part 1: Data preprocessing

The raw data that were collected from the sensor network system consist of data of five features: humidity, temperature, PM_1, $PM_{2.5}$, and PM_{10}. These raw data have to be preprocessed as follows:



Step 1: Each feature has different unit of measurement; therefore, these raw data have to be normalized by using the standard normalization or Z normalization technique [43, 44]. The normalized data are kept in $Data_j$ for $1,..,N$ and N is the number of sensors. After normalization, all data are transformed into the new values with the mean close to zero and the standard deviation close to one. Z normalization is expressed as Equation (4.1).

$$x' = \frac{x - \bar{x}}{s} \tag{4.1}$$

where $x'$is the normalized data, x is the original data, $\bar{x}$ is the mean of the original data and s is the standard deviation of x.

4.2 PM_{10} Concentration Prediction

To predict the concentration of PM_{10} at each monitoring station, a new model based on an integration of the feature selection method, the supervised learning model, and the modified depth-first search algorithm (MDSF) is proposed. Feature selection methods are used to select features for the PM_{10} concentration prediction; whereas, supervised learning models are used for prediction and MDFS seeks to resolve the global-local duality and to improve the prediction accuracy. As there is no explicit premise on which input features affect the prediction of the PM_{10} concentration, thus the feature selection method is used to select only associated features whose data are used as input to the supervised learning models. The prediction process starts from using data from the sensor at the station of interest as the initial input, the process is then repeated by adding data from the sensor at the neighboring station that is selected by MDFS one sensor per iteration. The process repeats until no more station is in the defined radius.

Part 2: PM_{10} concentration prediction

The prediction process combines the feature selection, the supervised learning models, and the modified depth-first-search algorithm together to improve the prediction results. According to section 3.2.1, three feature selection methods are used: forward selection, backward elimination, and GA. Forward selection and backward elimination are straight forward and no parameter setting is needed but GA needs some settings.

The setting of GA in this research starting from using binary encoding for chromosome encoding where each gene in a chromosome represents an individual feature that contains the value of 0 or 1. These binary values are used to indicate whether a feature is “selected” or “not selected” to be included in the prediction. The number of chromosomes in the population is set to 10 and the maximum number of generations is set to 100. Then the multiple linear regression, the multilayer perceptron neural network, and the support vector regression are used as the fitness evaluation function to calculate the fitness scores. The fitness scores are used in the tournament selection method to select the fittest chromosomes. The tournament size (the number of chromosomes for each round of the tournament) is equal to two [45] which represents the number of chromosomes to be chosen as a pair of parents. In terms of crossover, a one-point crossover is used with the probability of 0.5 and the probability of mutation is equal to $\frac{1}{\sigma}$ where $\boldsymbol{\sigma}$ is the total number of features on a chromosome as shown in Table 9. Moreover, there are also some parameter settings for the supervised learning models that are used in this research as shown in Table 10 and all variables used in the prediction process are defined in Table 11. The prediction process is as follows:

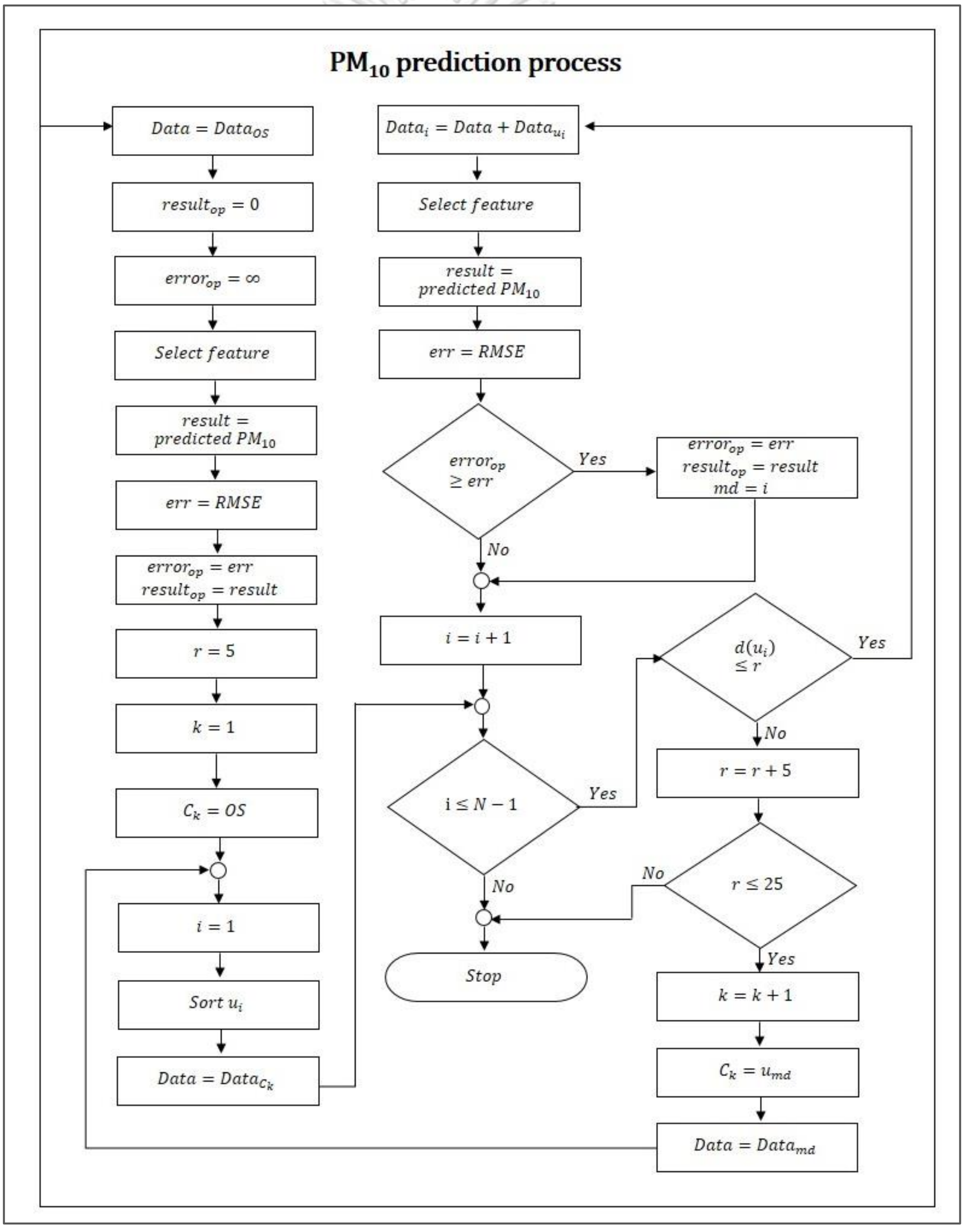

PM_{10} prediction process
$Data = Data_{OS}$
$result_{op} = 0$
$error_{op} = \infty$
Select feature
result = predicted PM_{10}
$err = RMSE$
$error_{op} = err$ $result_{op} = result$
$r = 5$
$k = 1$
$C_k = OS$
$i = 1$
Sort u_i
$Data = Data_{C_k}$
$Data_i = Data + Data_{u_i}$
Select feature
result = predicted PM_{10}
$err = RMSE$
$error_{op} \geq err$
Yes
$error_{op} = err$ $result_{op} = result$ $md = i$
No
$i = i + 1$
$i \leq N - 1$
Yes
No
Stop
$d(u_i) \leq r$
Yes
No
$r = r + 5$
$r \leq 25$
No
Yes
$k = k + 1$
$C_k = u_{md}$
$Data = Data_{md}$

Table 9 Parameter settings for the feature selection methods

Feature selection	**Parameter setting**
Forward selection	-
Backward elimination	-
Genetic algorithm	Population size = 10 Maximum number of iterations = 100 Probability of crossover = 0.5 Probability of mutation = $\frac{1}{\sigma}$ where σ is the total number of features in a chromosome

Table 10 Parameter settings for supervised learning models

Model	**Parameter setting**
Multiple linear regression	-
Multilayer perceptron neural network	$\delta = 0.0001$ Learning rate = 0.01 Maximum number of iterations = 200 Number of hidden layers = 1 Number of hidden nodes $= \left\lceil \frac{(k+1)}{2} \right\rceil + 1$ where k is the total number of input nodes
Support vector regression	$C = 0$ $\varepsilon = 0.0001$ Kernel type = dot product (linear) Maximum number of iterations = 100000

Table 11 The definitions of variables for the PM_{10} concentration prediction

$Data_{OS}$	The matrix containing data of the observed station (OS) where PM_{10} is predicted
C_k	The station of interest during each iteration for $k = 1, 2, \dots, N$
u_i	The neighboring station for $i = 1, \dots, N-1$
$d(u_i)$	The distance between C_k and u_i
$Data_{C_k}$	The matrix containing data of five features from a sensor at station C_k
$Data_{u_i}$	The matrix containing data of five features from a sensor at station u_i
N	The total number of sensors

Step 2.1 PM_{10} prediction starts from using only data from the sensor at the observed station; i.e., $Data = Data_{c_1} = Data_{OS}$. Features that are related to the PM_{10} prediction are selected from $Data_{c_1}$ using feature selection algorithms. At each iteration, data of the selected features stored in $Data_{c_1}$ are transmitted to supervised learning models to predict the concentration of PM_{10}. Afterwards, the prediction results and the root mean square errors are saved in the $result$ and err, respectively. Besides, the result with the minimum root mean square error is stored as the initial optimal result, $result_{op}$, and the root mean square error is stored as the initial optimal error, $error_{op}$.

Step 2.2 Next, the data from the sensor at a neighboring station located within the determined radius is added to $Data$. The neighboring sensor can be found by sorting all stations according to the distance $d(u_i)$ between each station u_i and station C_k for $i = 1, .., N-1$. At station u_i, if $d(u_i)$ is shorter than the radius r from C_k then u_i is selected as the neighboring station. The radius r is measured in kilometers and starts from $r = 5$ based on the actual distances which can be divided into four rings based on the location of monitoring stations. From Table 7, most of the monitoring stations have their closest neighbors within 5 kilometers but some have their closest neighbors within 10 kilometers. Thus, the initial radius is set at 5 and increased by 5 if no neighbor is found within the determined radius. An example of the rings around monitoring station S_2 starting from $r = 5$ to $r = 25$ is shown in Figure 20.

Step 2.3 The data of the neighboring sensor $(Data_{u_i})$ is integrated with the data $(Data)$ to form a new data set, $Data_i$.

Step 2.4 With the new data $Data_i$, the same feature selection algorithms as in step 2.1 are used to select the features that are associated with the PM_{10} prediction. As in step 2.1, the data of the selected features are forwarded to supervised learning models and the prediction results are stored, $result$, and RMSE from the predicts are calculated and stored in err.

Step 2.5 The current $error_{op}$ is compared to RMSEs stored in err. If any err is less than $error_{op}$ then the $error_{op}$ value is set to equal to (err) and the $result_{op}$ value is set to $result$. The index of the neighboring station with the optimal result is kept in md.

Step 2.6 Considering the next neighboring station, u_i, for $i = i + 1$. If the distance between C_k and u_i, $d(u_i)$, is shorter than the determined radius then the proposed method goes back to step 2.3. If the distance between C_k and u_i, $d(u_i)$, is longer than the radius then the radius will be increased by 5 and goes to step 2.7.

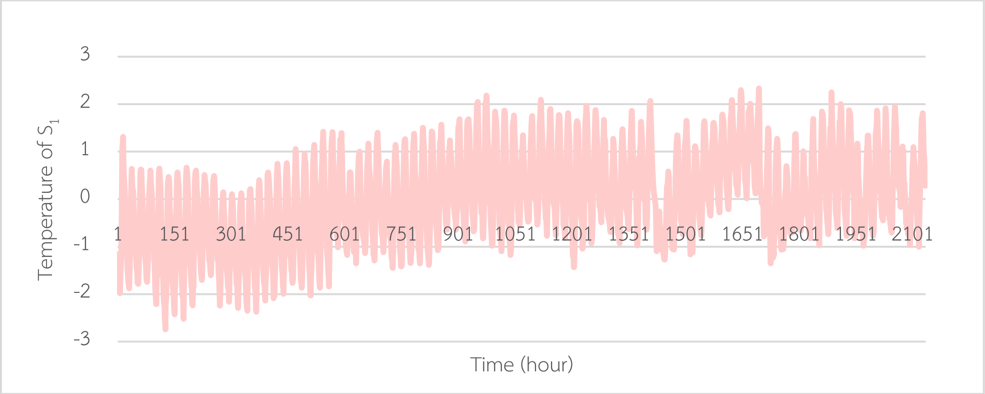
Step 2.7 If the radius is less than 25 then MDFS is applied to this step by shifting the station of interest C_k to the neighboring station that is kept in md and setting $Data = Data_{md}$. The process then goes back to step 2.3. If the radius is larger than 25, the final result of the prediction is obtained from the value that is stored in $result_{op}$.

CHAPTER 5 RESULTS

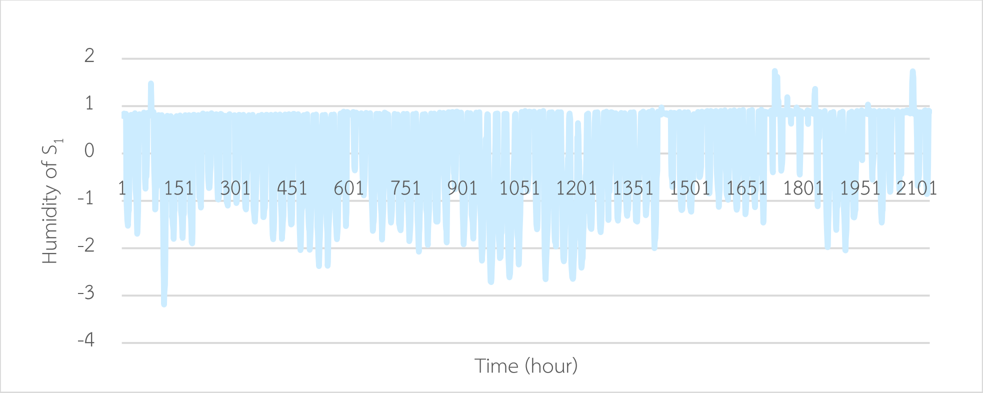
In this chapter, the first section presents the data from six monitoring stations that were normalized in the preprocessing step and the experimental results are presented in the following section.

5.1 The normalized data

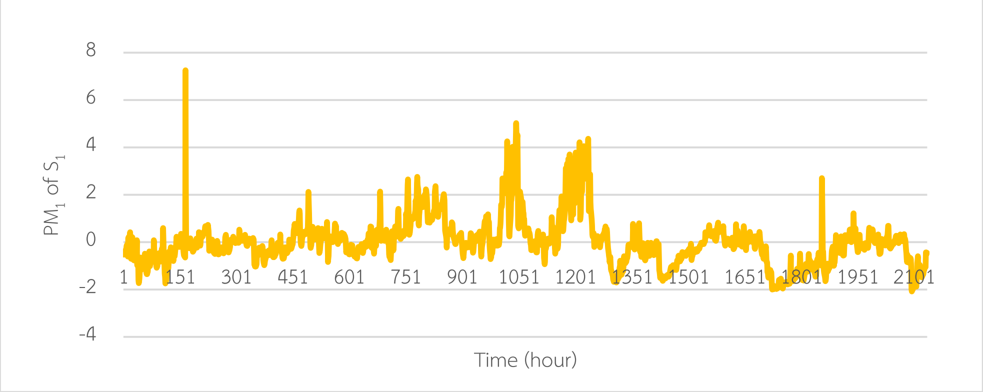
For our case study on the dataset obtained from six monitoring stations in Nan province for the period from the start of February 2017 to the end of April 2017, were normalized using Z normalization. Figure 27-32 shows the normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} for the sensors at stations S_1, S_2, S_3, S_4, S_5, and S_6, respectively. After normalization, it can be noticed that all of the data have the means equal to zero and the standard deviations equal to one.



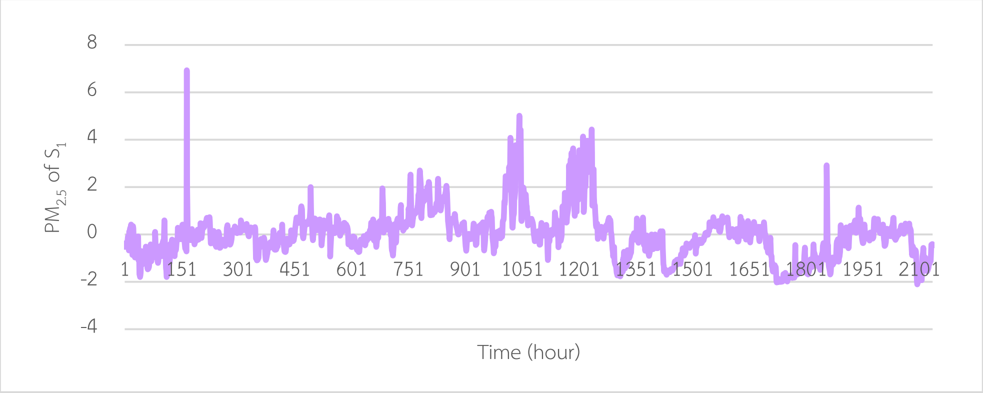
(a) The normalized data of temperature



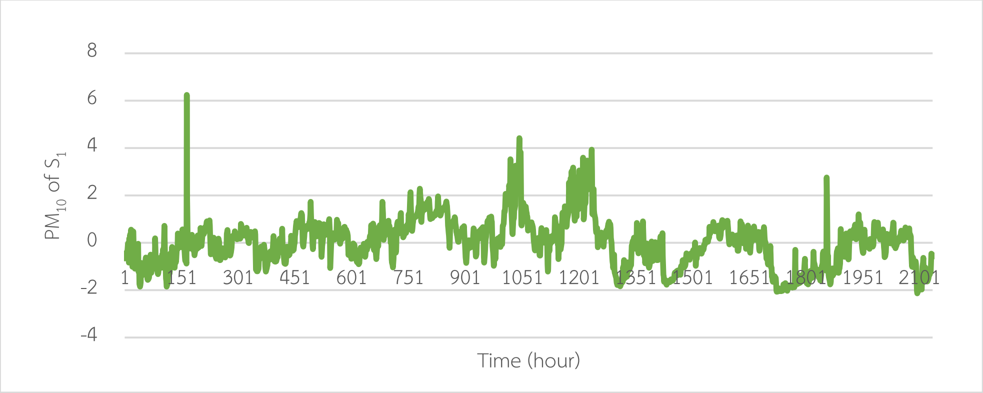
(b) The normalized data of humidity



(c) The normalized data of PM_1

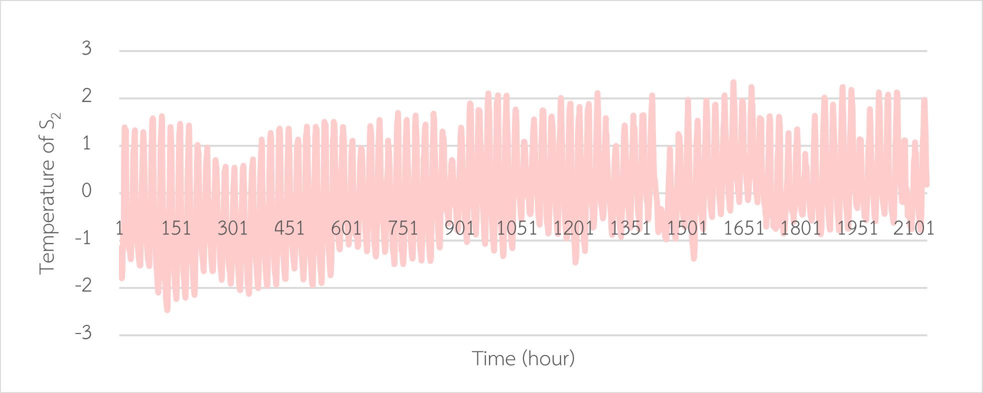


(d) The normalized data of $PM_{2.5}$

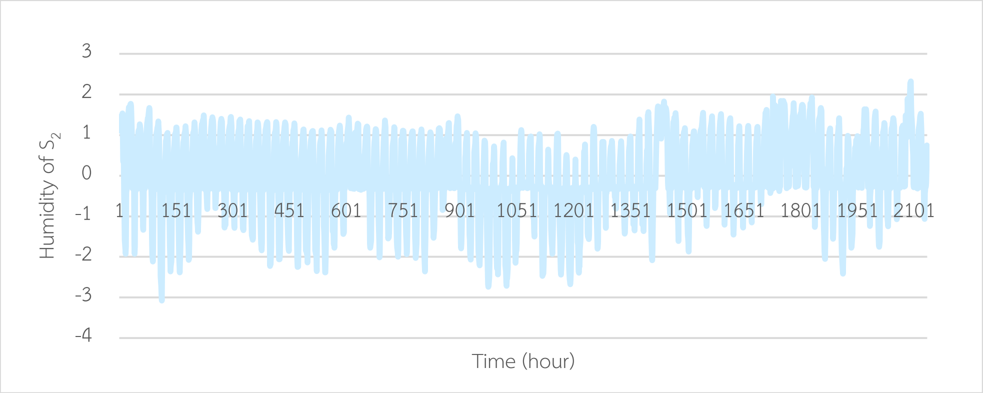


(e) The normalized data of PM_{10}

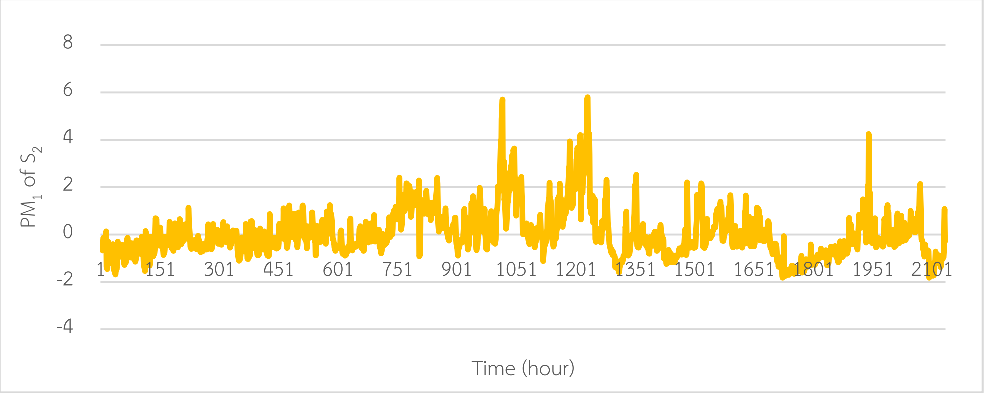
Figure 27 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_1



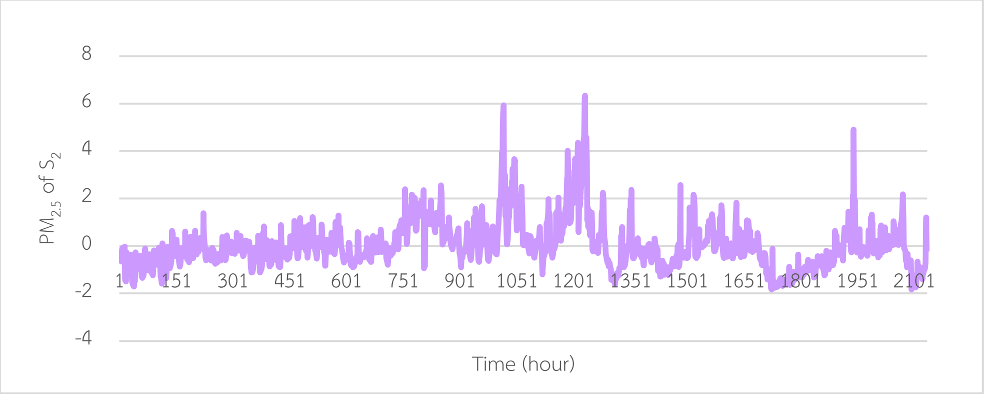
(a) The normalized data of temperature



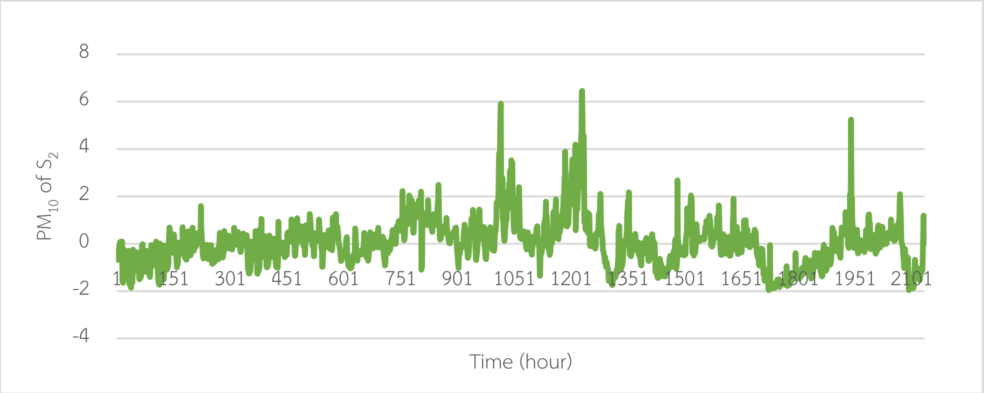
(b) The normalized data of humidity



(c) The normalized data of PM_1

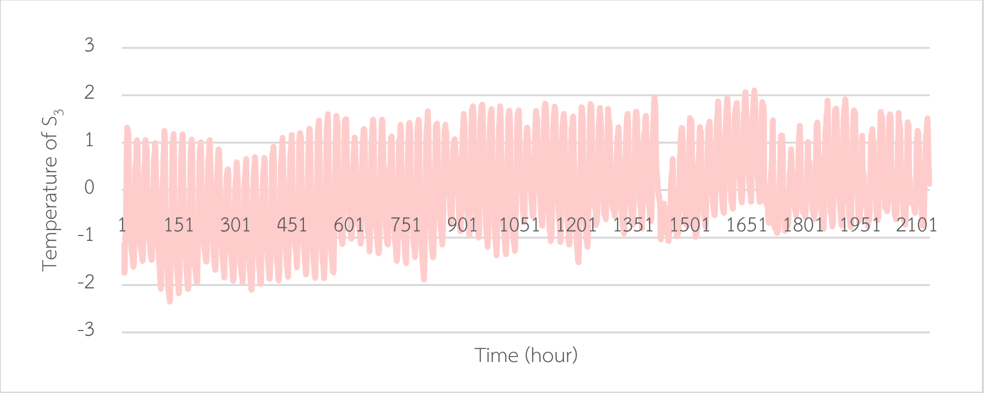


(d) The normalized data of $PM_{2.5}$

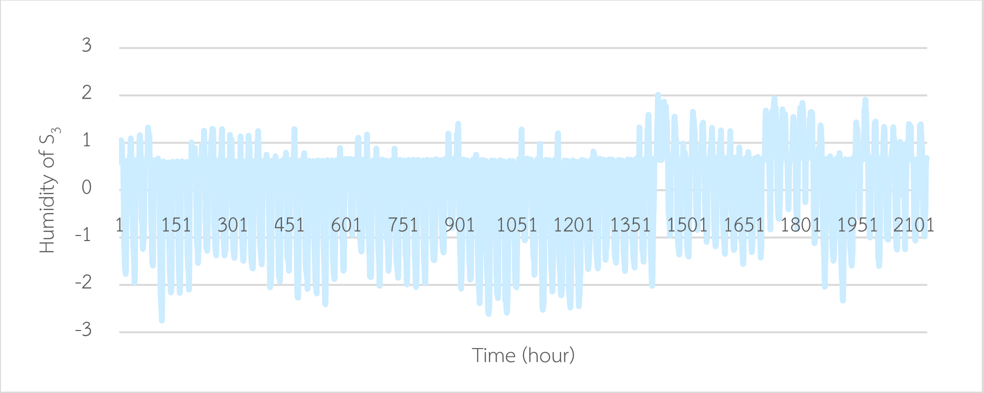


(e) The normalized data of PM_{10}

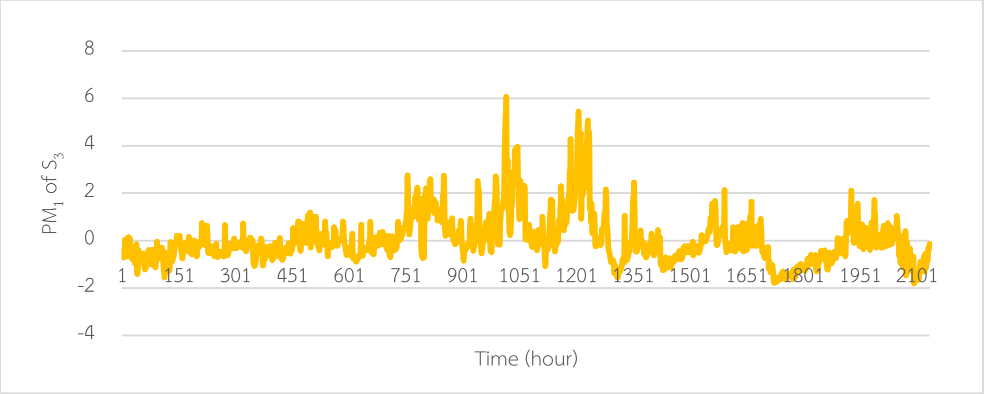
Figure 28 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_2



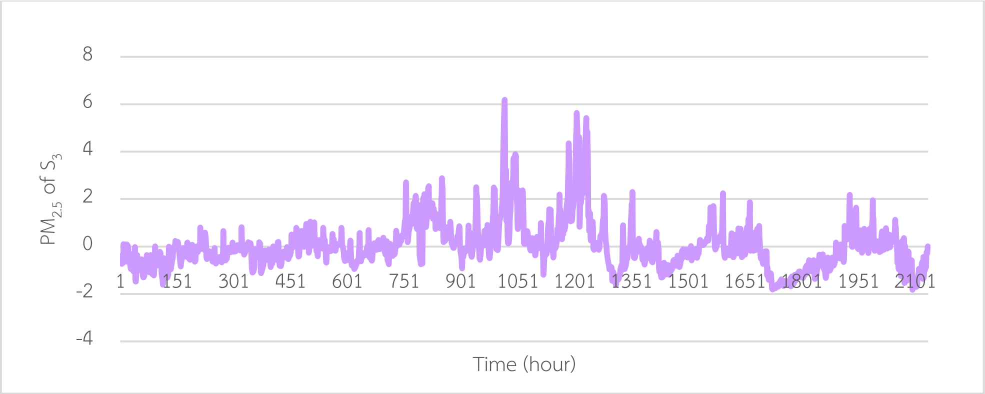
(a) The normalized data of temperature



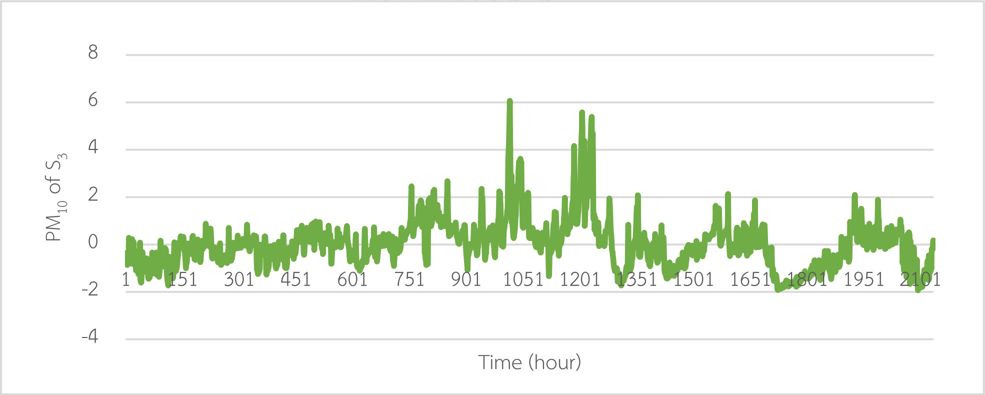
(b) The normalized data of humidity



(c) The normalized data of PM_1

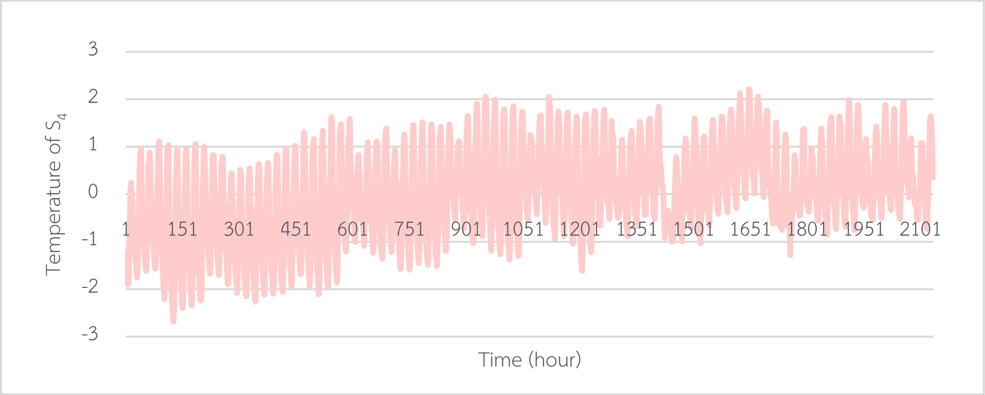


(d) The normalized data of $PM_{2.5}$

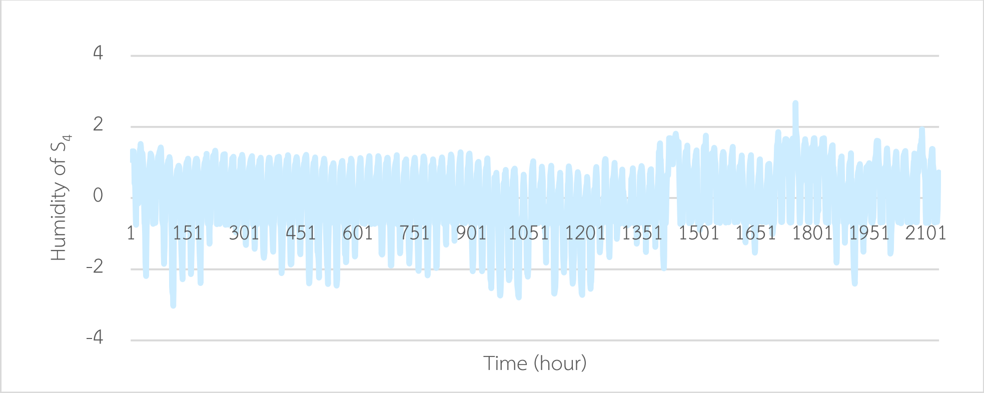


(e) The normalized data of PM_{10}

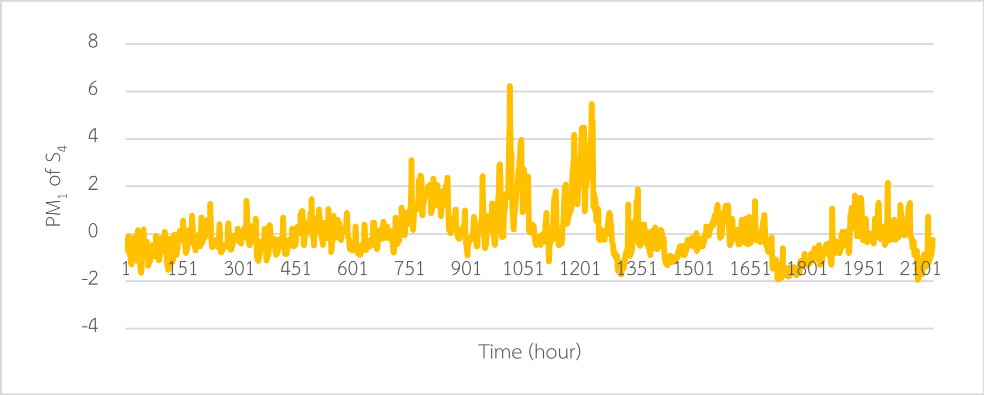
Figure 29 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_3



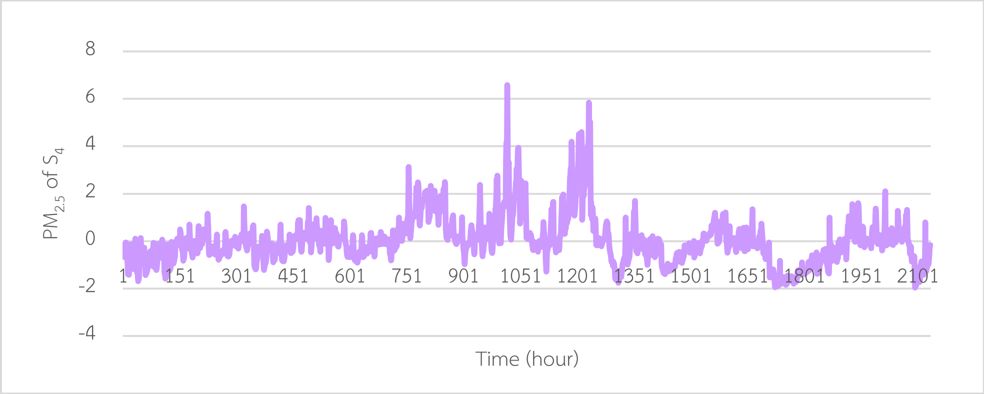
(a) The normalized data of temperature



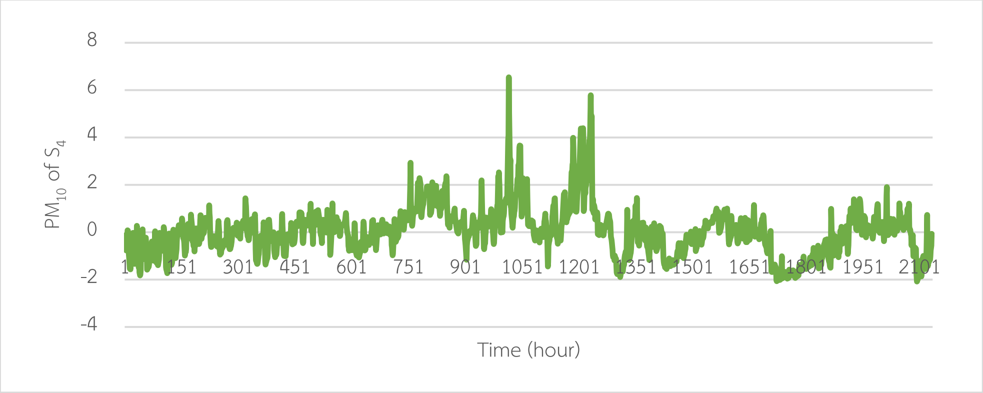
(b) The normalized data of humidity



(c) The normalized data of PM_1

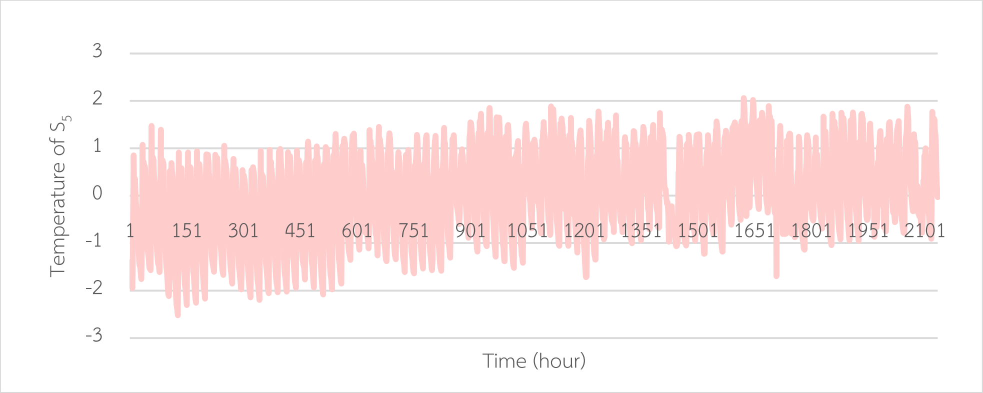


(d) The normalized data of $PM_{2.5}$

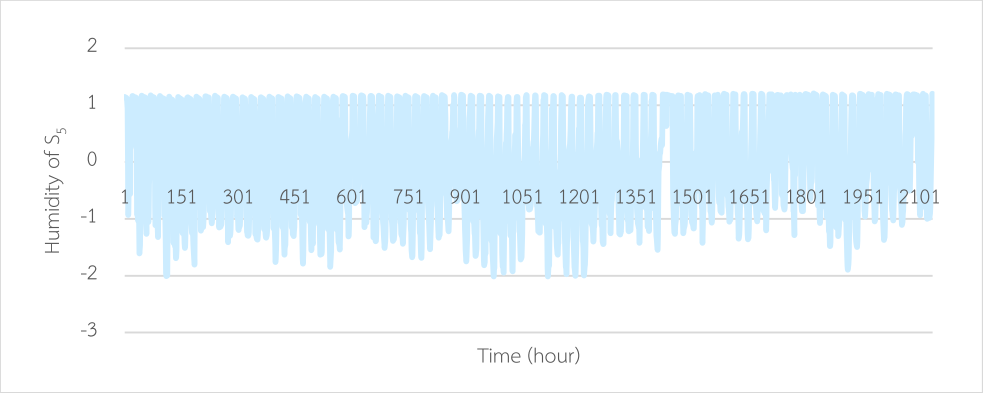


(e) The normalized data of PM_{10}

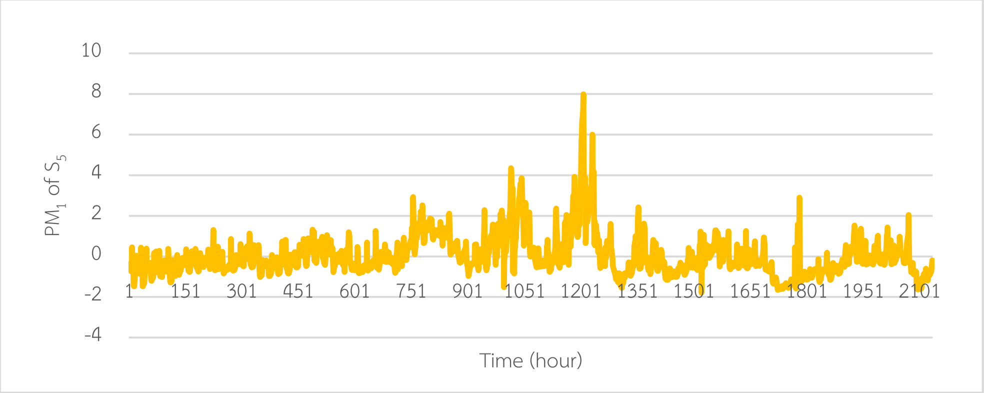
Figure 30 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_4



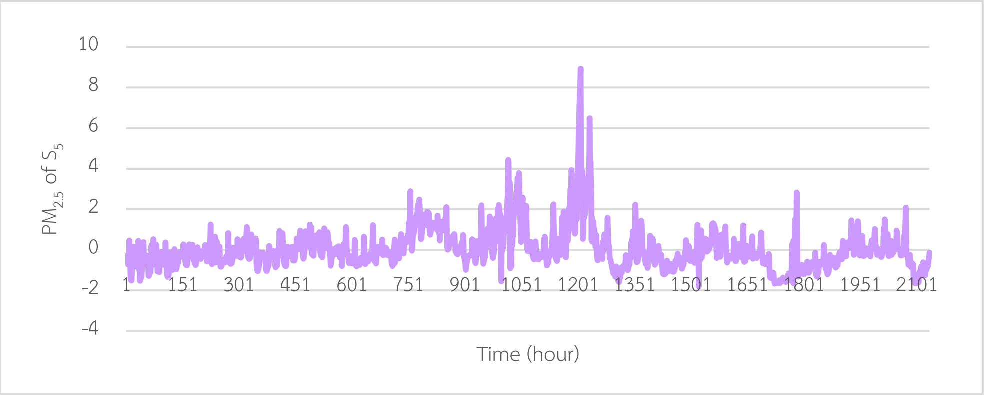
(a) The normalized data of temperature



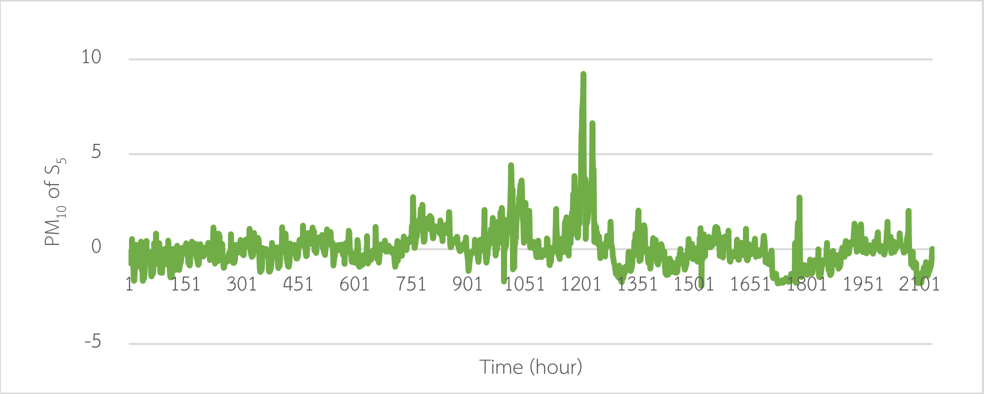
(b) The normalized data of humidity



(c) The normalized data of PM_1

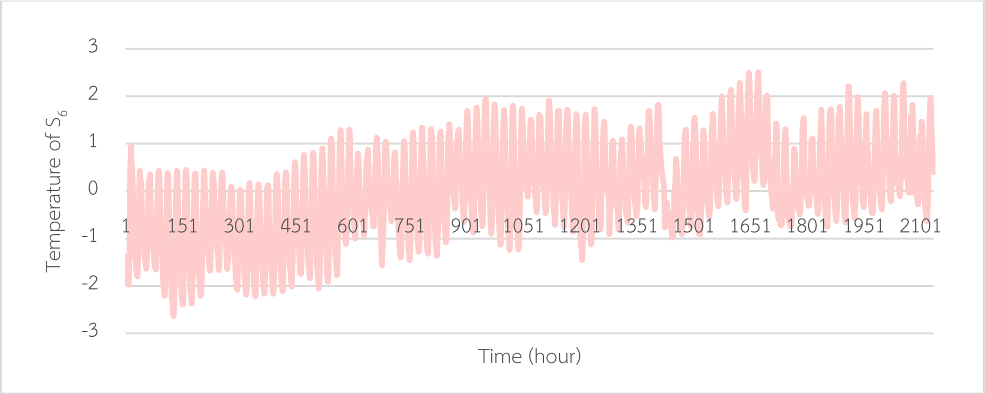


(d) The normalized data of $PM_{2.5}$

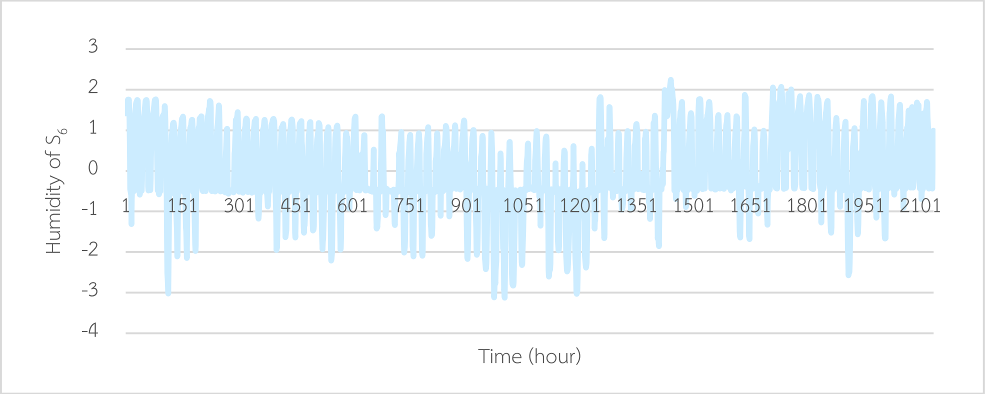


(e) The normalized data of PM_{10}

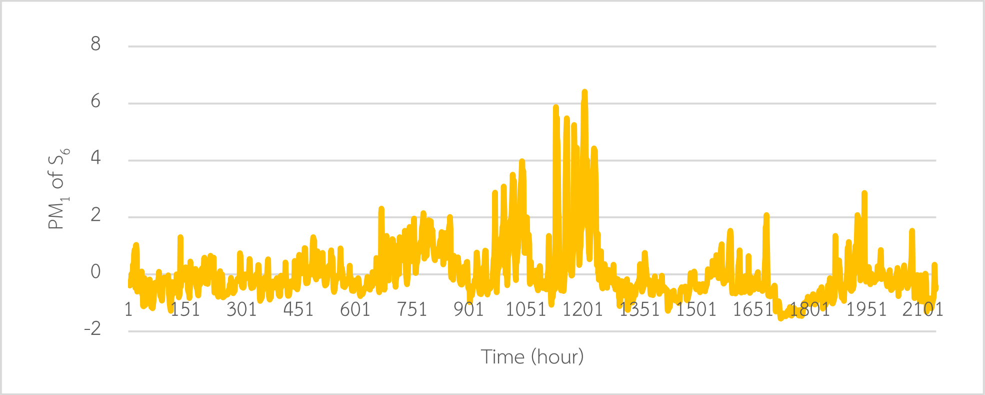
Figure 31 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_5



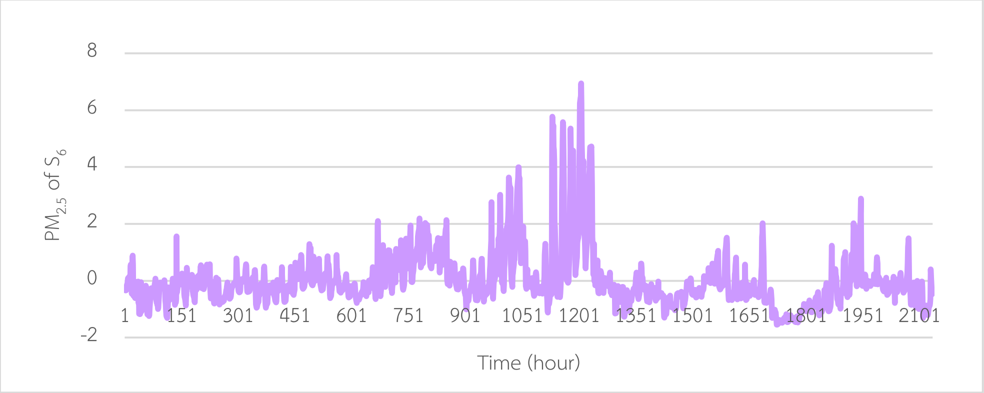
(a) The normalized data of temperature



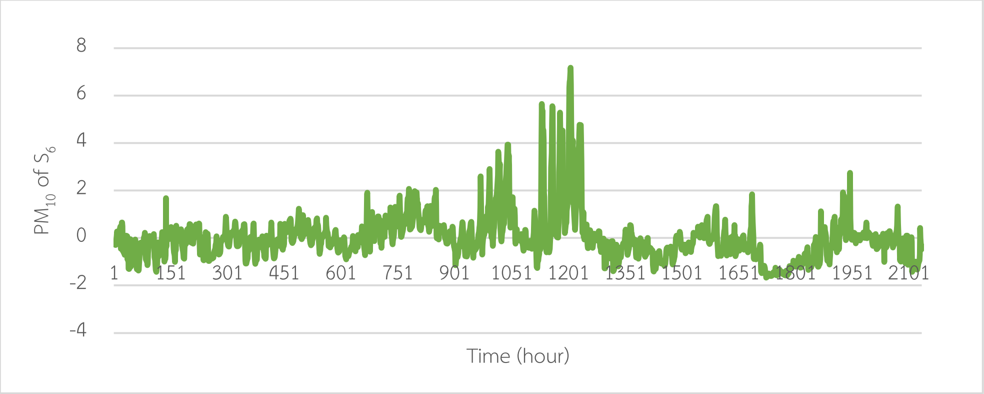
(b) The normalized data of humidity



(c) The normalized data of PM_1



(d) The normalized data of $PM_{2.5}$



(e) The normalized data of PM_{10}

Figure 32 The normalized data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} from sensor S_6

5.2 The experimental results for PM_{10} concentration prediction

For our case study on the dataset obtained from six monitoring stations, 70 percent of data in the chronological order were used for training and another 30 percent of data were used for testing; i.e., the split test. In order to evaluate the precision of the prediction results, four standard evaluation criteria; i.e. RMSE, Pearson correlation coefficient (R^2), MAE, and MAPE, were used. The experiments were performed in three different phases on the same dataset. The first two phases used only data from a sensor at the observed station to predict PM_{10} of that particular station. The first phase predicts PM_{10} concentration using three different supervised learning models; i.e. the multiple linear regression (MLR), the multilayer perceptron neural network (MLP), and the support vector regression (SVR). The second phase applied three different feature selection techniques; i.e., forward selection (FS), backward elimination (BE), and genetic algorithm (GA) to select only correlated features from data and then used three different supervised learning models as in the first phase to predict PM_{10} from these selected features. The third phase implemented the proposed method by adding more data from the neighboring sensors that were selected by MDFS to the dataset, used three different feature selection techniques to select only correlated features from the new dataset, and then used three different supervised learning models.

In the first phase, the experiments used data from the sensor at the observed station to predict PM_{10} concentrations for the next hour of that particular sensor by three different supervised learning models without applying any feature selection technique. Table 12 shows RMSE, MAE, MAPE, and R^2 of prediction results obtaining from MLR, MLP, and SVR, and it can be seen that SVR yields the best results.

Table 12 The PM_{10} concentration prediction results of each monitoring station using three models.

		Observed stations					
		S_1	S_2	S_3	S_4	S_5	S_6
MLR	RMSE	7.129	11.116	8.737	8.370	10.221	9.781
	MAE	5.422	7.796	6.851	6.537	7.347	7.541
	MAPE (%)	29.14	27.27	30.66	27.14	29.20	26.37
	R^2	0.886	0.816	0.844	0.863	0.760	0.819
MLP	RMSE	5.955	10.007	8.213	7.027	9.648	9.092
	MAE	3.736	5.775	5.837	4.821	6.566	6.585
	MAPE (%)	15.15	14.92	17.82	14.65	21.68	19.46
	R^2	0.897	0.849	0.871	0.901	0.795	0.855
SVR	RMSE	5.703	10.211	7.494	6.882	9.217	8.716
	MAE	3.192	5.996	4.774	4.497	5.774	5.910
	MAPE (%)	11.85	17.21	16.47	15.57	20.58	19.33
	R^2	0.900	0.841	0.872	0.898	0.804	0.853

In the second phase, we integrated three feature selection techniques with three supervised learning models for the total of nine different combinations of methods to predict the PM_{10} concentration. Table 13 shows the optimal combination of methods for each observed station. The prediction results for the next hour of each observed station when feature selection techniques were applied to data before they were forwarded to three supervised learning models have lower RMSE values than those of the prediction results obtained from the first phase. At the observed stations S_1 and S_2, the optimal prediction results were obtained from using GA with MLP; while at the observed station S_3, the optimal prediction result was obtained from using FS or BE or GA with MLR. At the observed stations S_4, S_5, and S_6, the optimal prediction results were obtained from using FS or BE or GA with SVR.

Table 13 The PM_{10} concentration prediction results for the next hour of each observed station using feature selection and supervised learning models.

Observed stations	The PM_{10} concentration prediction results from the second phase					
	Feature selection	Supervised learning model	RMSE	MAE	MAPE (%)	R^2
S_1	GA	MLP	5.610	3.139	13.22	0.902
S_2	GA	MLP	9.898	5.908	17.32	0.850
S_3	FS/BE/GA	MLR	7.369	4.631	17.55	0.877
S_4	FS/BE/GA	SVR	6.620	4.048	12.94	0.905
S_5	FS/BE/GA	SVR	8.950	5.176	16.45	0.817
S_6	FS/BE/GA	SVR	8.439	5.306	15.87	0.862

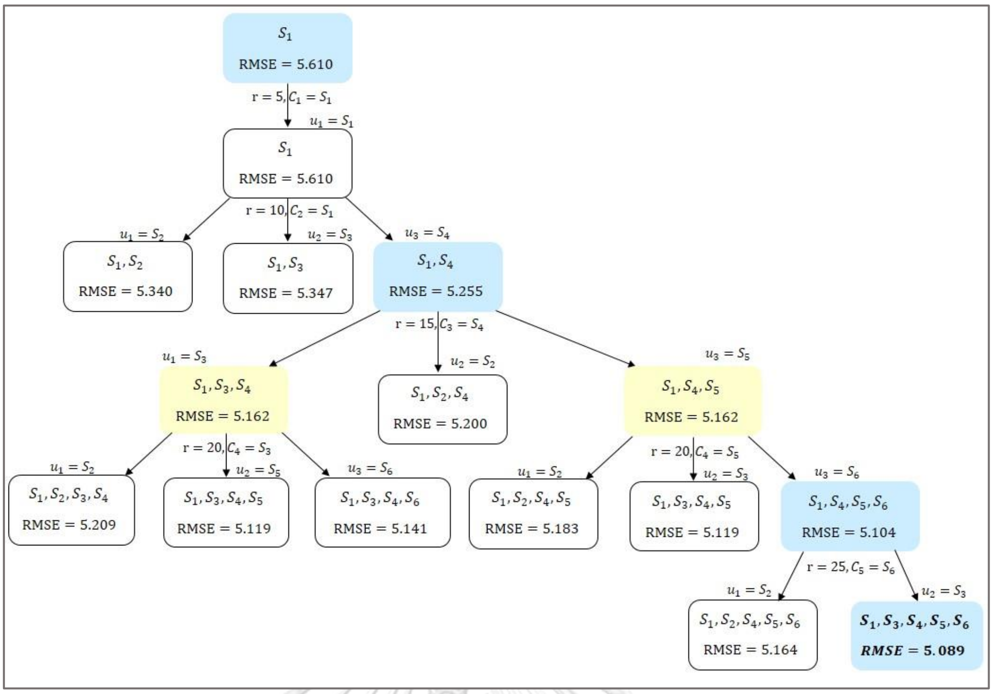
For the third phase, we propose to use data from the neighboring sensors to predict the PM_{10} concentration of the observed station by using MDFS to select the neighboring sensors. Every time the neighboring sensor was selected, its data were added to the existing data, and then feature selection techniques were used to select the correlated features and finally the supervised learning models were used to predict the PM_{10} concentration. Table 14 shows the prediction results which can be noticed that RMSE of all results are less than those of the first two phases. At the observed stations $S_1, S_2, S_3,$ and S_4, the optimal prediction results were obtained from using GA with MLP; while the optimal prediction result at the observed station S_5 was obtained from using FS with MLR. At the observed station S_6, the optimal prediction result was obtained from FS with SVR. The experiments in the third phase used the proposed MDFS algorithm to select neighboring sensors whose data may influence the PM_{10} prediction of the observed station with no need of wind direction and wind speed data.

Table 14 The PM_{10} concentration prediction results for the next hour of each observed station using MDFS algorithm.

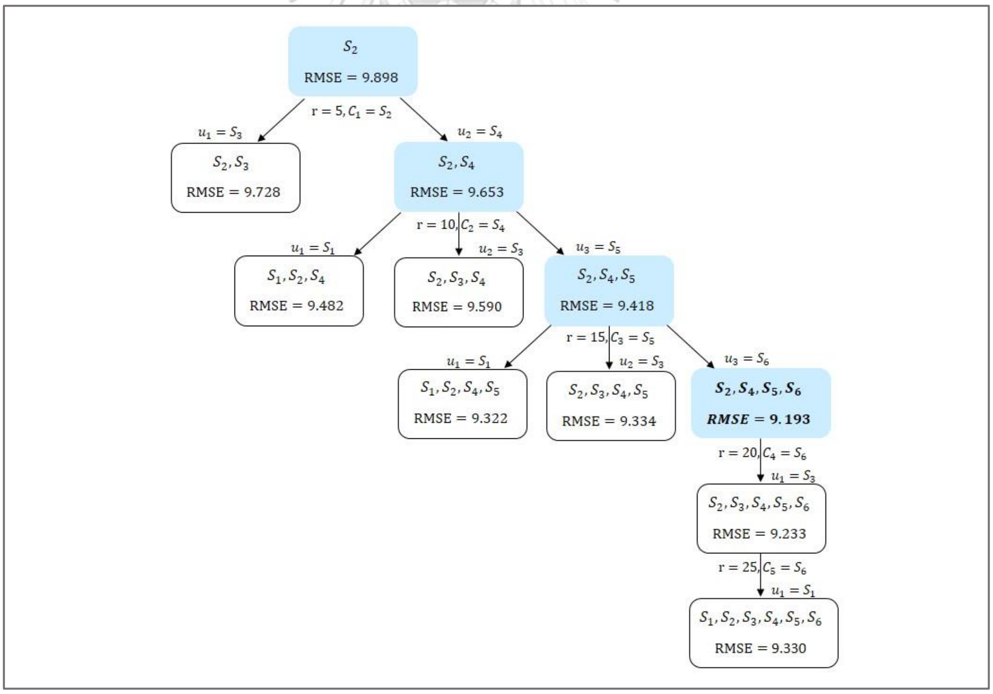
Observed stations	The PM_{10} concentration prediction results from the third phase					
	Feature selection	Supervised learning model	RMSE	MAE	MAPE (%)	R^2
S_1	GA	MLP	**5.089**	2.963	12.22	0.919
S_2	GA	MLP	**9.193**	6.163	20.77	0.871
S_3	GA	MLP	**6.846**	4.791	19.15	0.893
S_4	GA	MLP	**6.152**	4.129	15.62	0.918
S_5	FS	MLR	**8.385**	4.936	15.56	0.841
S_6	FS	SVR	**7.883**	5.021	14.90	0.880

Figure 33 shows the process of selecting the neighboring stations using the MDFS algorithm as presented in Figure 17 for sensors S_1, S_2, S_3, S_4, S_5, and S_6, respectively. Suppose we would like to predict the PM_{10} concentration at monitoring station S_1 as shown in Figure 33(a), the proposed method starts the prediction from setting S_1 as the observed station and using only data obtained from the sensor at station S_1. Next, the sensors at the neighboring stations whose locations are within the specified radius are selected one at a time and its data are added to the existing data. The neighboring station whose additional data yields the lowest RMSE will be selected as the new station of interest according to MDFS which is S_4 from our example and the optimal prediction result with the lowest RMSE is kept at this step. Once S_4 becomes the station of interest, the radius is increased by 10 because there is no neighboring station within the next 5 kilometers and the proposed method is repeated by selecting the neighboring sensors whose locations are within the specified radius one at a time and adding its data to the existing data. The neighboring station whose additional data yields the lowest RMSE will be selected as the new station of interest and the optimal prediction result with the lowest RMSE is kept. The process is repeated until there is no more neighboring station within the specified radius or the radius is greater than 25 kilometers. According to Figure 33(a), it can be seen that the optimal result from the proposed method for PM_{10} concentration prediction of the observed station, S_1, used data from only four stations S_1, S_3, S_4, S_5, S_6 with the lowest RMSE equals to 5.089.

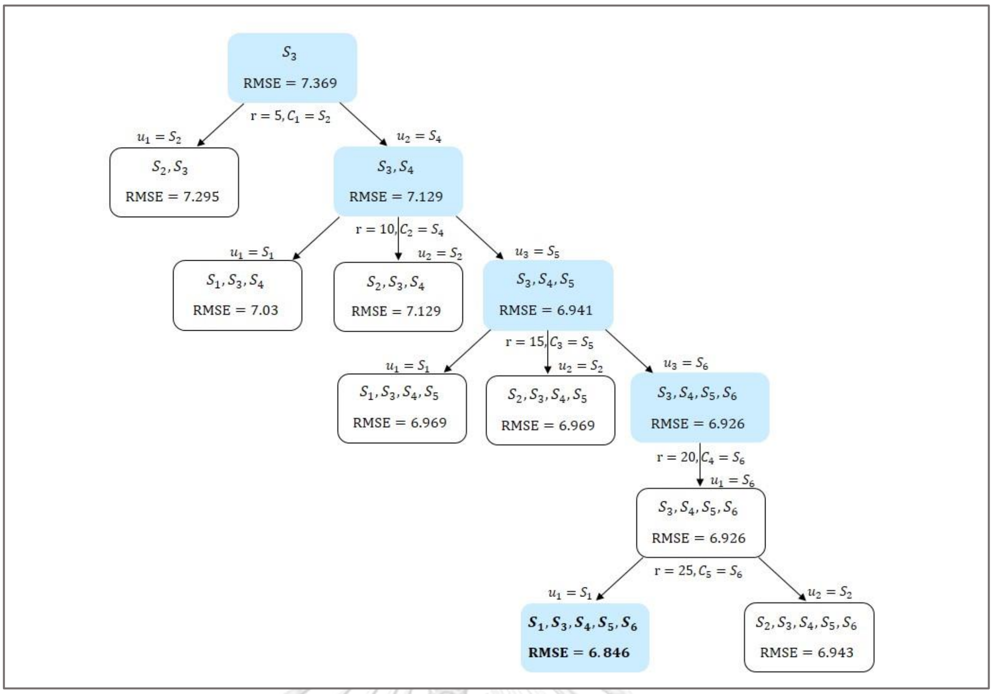
Likewise, suppose we would like to predict PM_{10} concentration at monitoring station S_2 as shown in Figure 33(b), the proposed method starts the prediction from setting S_2 as the observed station and using only data obtained from the sensor at station S_2. Then the sensors at the neighboring stations whose locations are within the specified radius are selected one at a time and its data are added to the existing data. The neighboring station whose additional data yields the lowest RMSE will be selected as the new station of interest according to MDFS which is S_4 from our example and the optimal prediction result with the lowest RMSE is kept at this step. Once S_4 becomes the station of interest, the radius is increased by 5 and the proposed method is repeated by selecting the neighboring sensors whose locations are within the specified radius one at a time and adding its data to the existing data. The neighboring station whose additional data yields the lowest RMSE will be selected as the new station of interest and the optimal prediction result with the lowest RMSE is kept. The process is repeated until there is no more neighboring station within the specified radius or the radius is greater than 25 kilometers. According to Figure 33(b), it can be seen that the optimal result from the proposed method for PM_{10} concentration prediction of the monitoring station, S_2, used data from only four stations S_2, S_4, S_5, S_6 with the lowest RMSE equals to 9.193.


S_1
RMSE = 5.610
r = 5, $C_1 = S_1$
$u_1 = S_1$
S_1
RMSE = 5.610
r = 10, $C_2 = S_1$
$u_1 = S_2$
S_1, S_2
RMSE = 5.340
$u_2 = S_3$
S_1, S_3
RMSE = 5.347
$u_3 = S_4$
S_1, S_4
RMSE = 5.255
r = 15, $C_3 = S_4$
$u_1 = S_3$
S_1, S_3, S_4
RMSE = 5.162
$u_2 = S_2$
S_1, S_2, S_4
RMSE = 5.200
$u_3 = S_5$
S_1, S_4, S_5
RMSE = 5.162
r = 20, $C_4 = S_3$
$u_1 = S_2$
S_1, S_2, S_3, S_4
RMSE = 5.209
$u_2 = S_5$
S_1, S_3, S_4, S_5
RMSE = 5.119
$u_3 = S_6$
S_1, S_3, S_4, S_6
RMSE = 5.141
r = 20, $C_4 = S_5$
$u_1 = S_2$
S_1, S_2, S_4, S_5
RMSE = 5.183
$u_2 = S_3$
S_1, S_3, S_4, S_5
RMSE = 5.119
$u_3 = S_6$
S_1, S_4, S_5, S_6
RMSE = 5.104
r = 25, $C_5 = S_6$
$u_1 = S_2$
S_1, S_2, S_4, S_5, S_6
RMSE = 5.164
$u_2 = S_3$
$\boldsymbol{S_1, S_3, S_4, S_5, S_6}$
$\boldsymbol{RMSE = 5.089}$

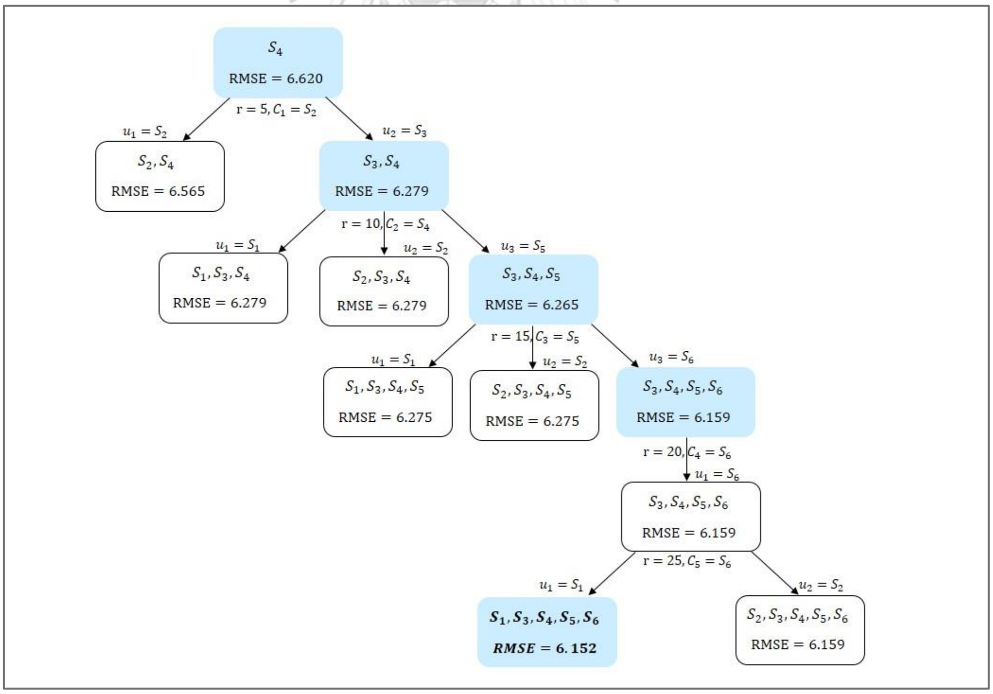
(a) Sensor of S_1


S_2
RMSE = 9.898
r = 5, $C_1 = S_2$
$u_1 = S_3$
S_2, S_3
RMSE = 9.728
$u_2 = S_4$
S_2, S_4
RMSE = 9.653
r = 10, $C_2 = S_4$
$u_1 = S_1$
S_1, S_2, S_4
RMSE = 9.482
$u_2 = S_3$
S_2, S_3, S_4
RMSE = 9.590
$u_3 = S_5$
S_2, S_4, S_5
RMSE = 9.418
r = 15, $C_3 = S_5$
$u_1 = S_1$
S_1, S_2, S_4, S_5
RMSE = 9.322
$u_2 = S_3$
S_2, S_3, S_4, S_5
RMSE = 9.334
$u_3 = S_6$
$\boldsymbol{S_2, S_4, S_5, S_6}$
$\boldsymbol{RMSE = 9.193}$
r = 20, $C_4 = S_6$
$u_1 = S_3$
S_2, S_3, S_4, S_5, S_6
RMSE = 9.233
r = 25, $C_5 = S_6$
$u_1 = S_1$
$S_1, S_2, S_3, S_4, S_5, S_6$
RMSE = 9.330

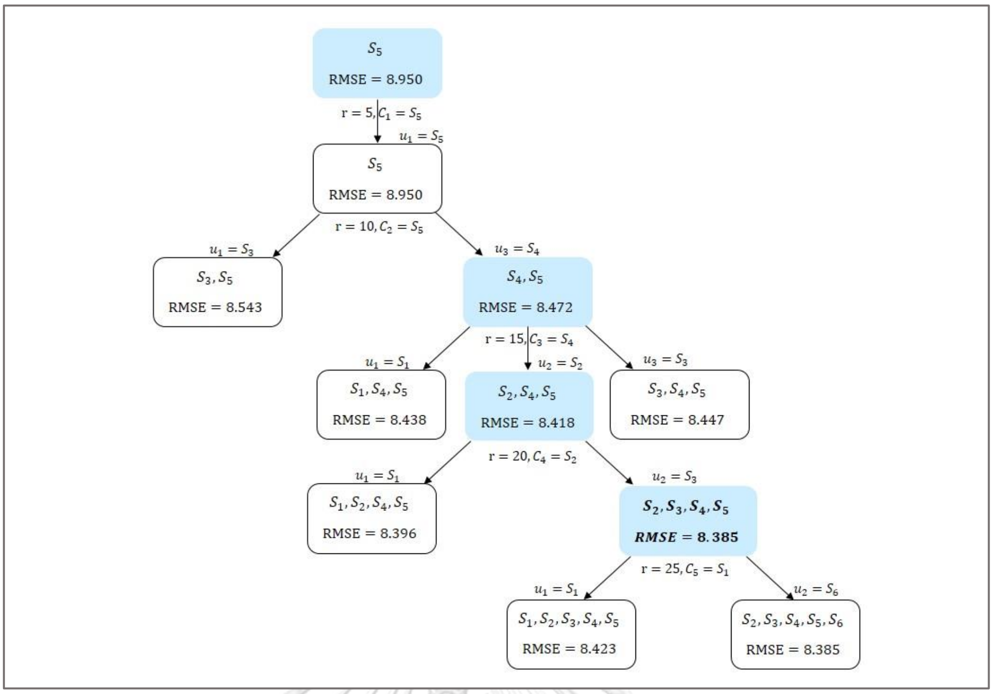
(b) Sensor of S_2


S_3
RMSE = 7.369
r = 5, $C_1 = S_2$
$u_1 = S_2$
S_2, S_3
RMSE = 7.295
$u_2 = S_4$
S_3, S_4
RMSE = 7.129
r = 10, $C_2 = S_4$
$u_1 = S_1$
S_1, S_3, S_4
RMSE = 7.03
$u_2 = S_2$
S_2, S_3, S_4
RMSE = 7.129
$u_3 = S_5$
S_3, S_4, S_5
RMSE = 6.941
r = 15, $C_3 = S_5$
$u_1 = S_1$
S_1, S_3, S_4, S_5
RMSE = 6.969
$u_2 = S_2$
S_2, S_3, S_4, S_5
RMSE = 6.969
$u_3 = S_6$
S_3, S_4, S_5, S_6
RMSE = 6.926
r = 20, $C_4 = S_6$
$u_1 = S_6$
S_3, S_4, S_5, S_6
RMSE = 6.926
r = 25, $C_5 = S_6$
$u_1 = S_1$
$\boldsymbol{S_1, S_3, S_4, S_5, S_6}$
RMSE = 6.846
$u_2 = S_2$
S_2, S_3, S_4, S_5, S_6
RMSE = 6.943

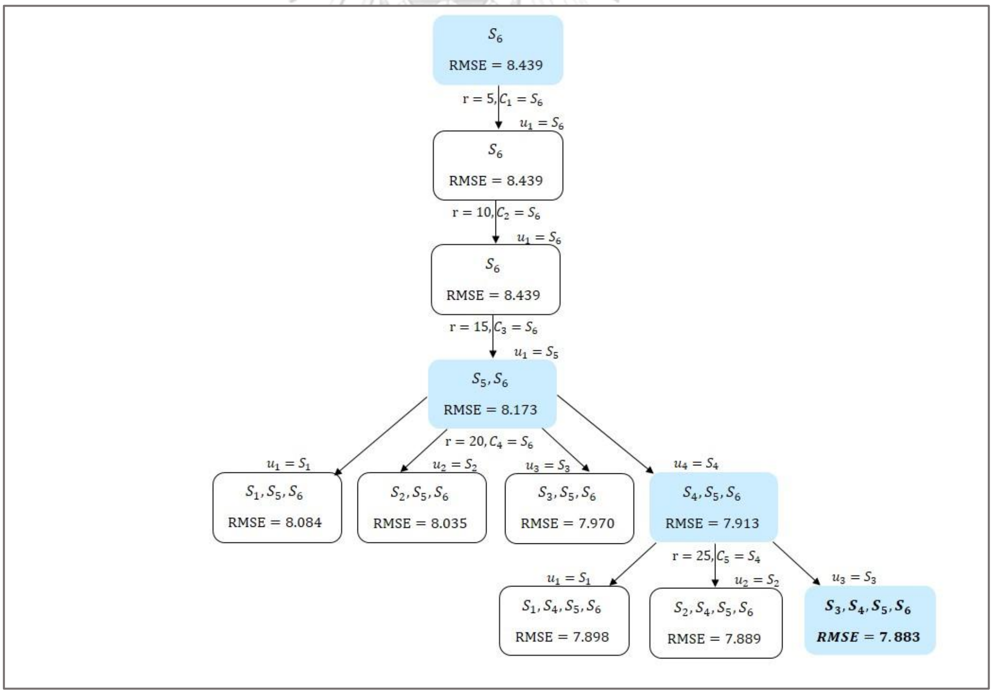
(c) Sensor of S_3


S_4
RMSE = 6.620
r = 5, $C_1 = S_2$
$u_1 = S_2$
S_2, S_4
RMSE = 6.565
$u_2 = S_3$
S_3, S_4
RMSE = 6.279
r = 10, $C_2 = S_4$
$u_1 = S_1$
S_1, S_3, S_4
RMSE = 6.279
$u_2 = S_2$
S_2, S_3, S_4
RMSE = 6.279
$u_3 = S_5$
S_3, S_4, S_5
RMSE = 6.265
r = 15, $C_3 = S_5$
$u_1 = S_1$
S_1, S_3, S_4, S_5
RMSE = 6.275
$u_2 = S_2$
S_2, S_3, S_4, S_5
RMSE = 6.275
$u_3 = S_6$
S_3, S_4, S_5, S_6
RMSE = 6.159
r = 20, $C_4 = S_6$
$u_1 = S_6$
S_3, S_4, S_5, S_6
RMSE = 6.159
r = 25, $C_5 = S_6$
$u_1 = S_1$
$\boldsymbol{S_1, S_3, S_4, S_5, S_6}$
RMSE = 6.152
$u_2 = S_2$
S_2, S_3, S_4, S_5, S_6
RMSE = 6.159

(d) Sensor of S_4


S_5
RMSE = 8.950
r = 5, $C_1 = S_5$
$u_1 = S_5$
S_5
RMSE = 8.950
r = 10, $C_2 = S_5$
$u_1 = S_3$
S_3, S_5
RMSE = 8.543
$u_3 = S_4$
S_4, S_5
RMSE = 8.472
r = 15, $C_3 = S_4$
$u_1 = S_1$
S_1, S_4, S_5
RMSE = 8.438
$u_2 = S_2$
S_2, S_4, S_5
RMSE = 8.418
$u_3 = S_3$
S_3, S_4, S_5
RMSE = 8.447
r = 20, $C_4 = S_2$
$u_1 = S_1$
S_1, S_2, S_4, S_5
RMSE = 8.396
$u_2 = S_3$
S_2, S_3, S_4, S_5
RMSE = 8.385
r = 25, $C_5 = S_1$
$u_1 = S_1$
S_1, S_2, S_3, S_4, S_5
RMSE = 8.423
$u_2 = S_6$
S_2, S_3, S_4, S_5, S_6
RMSE = 8.385

(e) Sensor of S_5


S_6
RMSE = 8.439
r = 5, $C_1 = S_6$
$u_1 = S_6$
S_6
RMSE = 8.439
r = 10, $C_2 = S_6$
$u_1 = S_6$
S_6
RMSE = 8.439
r = 15, $C_3 = S_6$
$u_1 = S_5$
S_5, S_6
RMSE = 8.173
r = 20, $C_4 = S_6$
$u_1 = S_1$
S_1, S_5, S_6
RMSE = 8.084
$u_2 = S_2$
S_2, S_5, S_6
RMSE = 8.035
$u_3 = S_3$
S_3, S_5, S_6
RMSE = 7.970
$u_4 = S_4$
S_4, S_5, S_6
RMSE = 7.913
r = 25, $C_5 = S_4$
$u_1 = S_1$
S_1, S_4, S_5, S_6
RMSE = 7.898
$u_2 = S_2$
S_2, S_4, S_5, S_6
RMSE = 7.889
$u_3 = S_3$
S_3, S_4, S_5, S_6
RMSE = 7.883

(f) Sensor of S_6

Figure 33 The process of selecting data from the neighboring stations using the MDFS algorithm for PM_{10} concentration prediction of each monitoring station

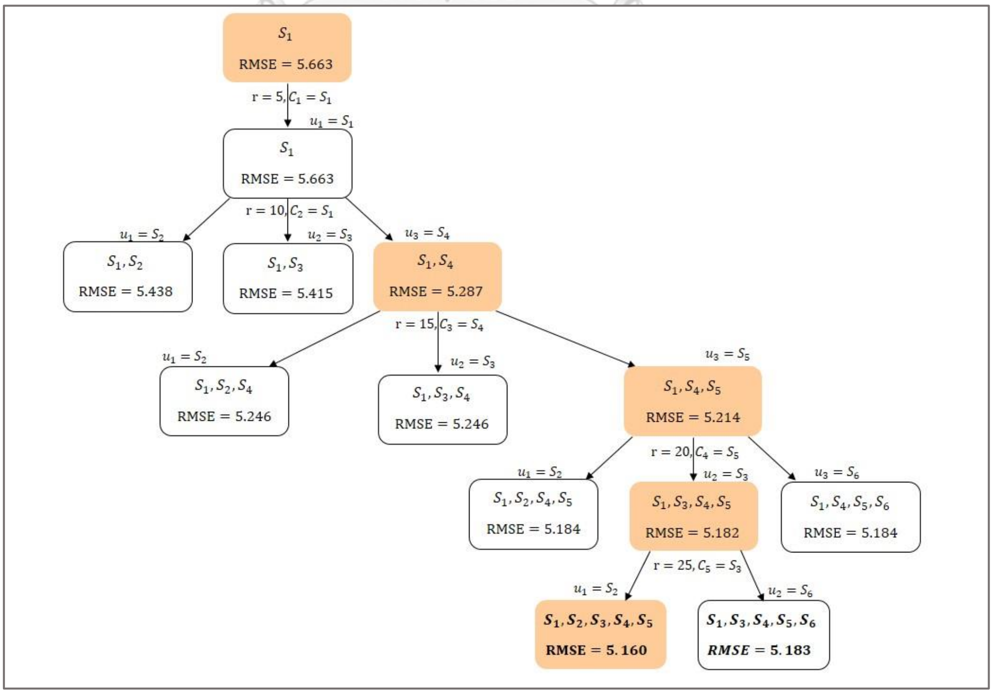
For the last phase, from Figure 21-26 of each sensor, the hourly data of PM_1, $PM_{2.5}$, and PM_{10} occasionally have extreme values at some certain times and we would like to test whether the normalization has any effect on the prediction; therefore, we test the proposed PM_{10} concentration prediction model on the actual data using three different feature selection techniques; i.e., forward selection (FS), backward elimination (BE), and genetic algorithm (GA), and two supervised learning models; i.e. the multiple linear regression (MLR), and the support vector regression (SVR). Table 15 shows the prediction results which can be noticed that RMSE of all the results are higher than those of the normalized data in the third phases. The prediction results from the actual data for the next hour of each observed station were obtained from using FS or BE or GA with MLR or SVR. Figure 34 shows the process of selecting the neighboring stations using the MDFS algorithm as presented in part 2 of Figure 17 for sensors S_1, S_2, S_3, S_4, S_5, and S_6, respectively.

Suppose we would like to predict PM_{10} concentration at monitoring station S_2 as shown in Figure 34(b), the prediction starts from setting S_2 as the observed station and using only data obtained from the sensor at station S_2. Then the sensors at the neighboring stations whose locations are within the specified radius are selected one at a time and its data are added to the existing data. The neighboring station whose additional data yields the lowest RMSE will be selected as the new station of interest according to the MDFS which is S_4 from our example, and the optimal prediction result with the lowest RMSE is kept at this step. Once S_4 becomes the station of interest, the radius is increased by 5 and the proposed method is repeated by selecting the neighboring sensors whose locations are within the specified radius one at a time and adding its data to the existing data. The neighboring station whose additional data yields the lowest RMSE will be selected as the new station of interest and the optimal prediction result with the lowest RMSE is kept. The process is repeated until there is no more neighboring station within the specified radius or the radius is greater than 25 kilometers. According to Figure 34(b), it can be seen that the optimal result for PM_{10}

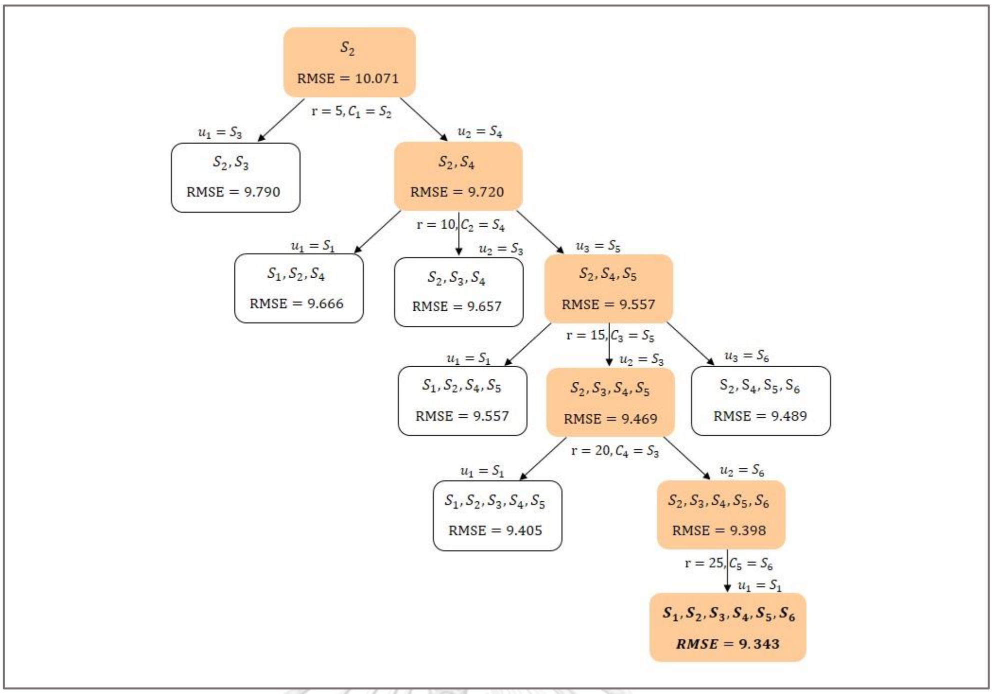
concentration prediction of the monitoring station, S_2, used data from only four stations S_1, S_3, S_4, S_5, S_6 with the lowest RMSE equals to 9.343.

Table 15 The PM_{10} concentration prediction results for the next hour of each observed station using MDFS algorithm with the actual data.

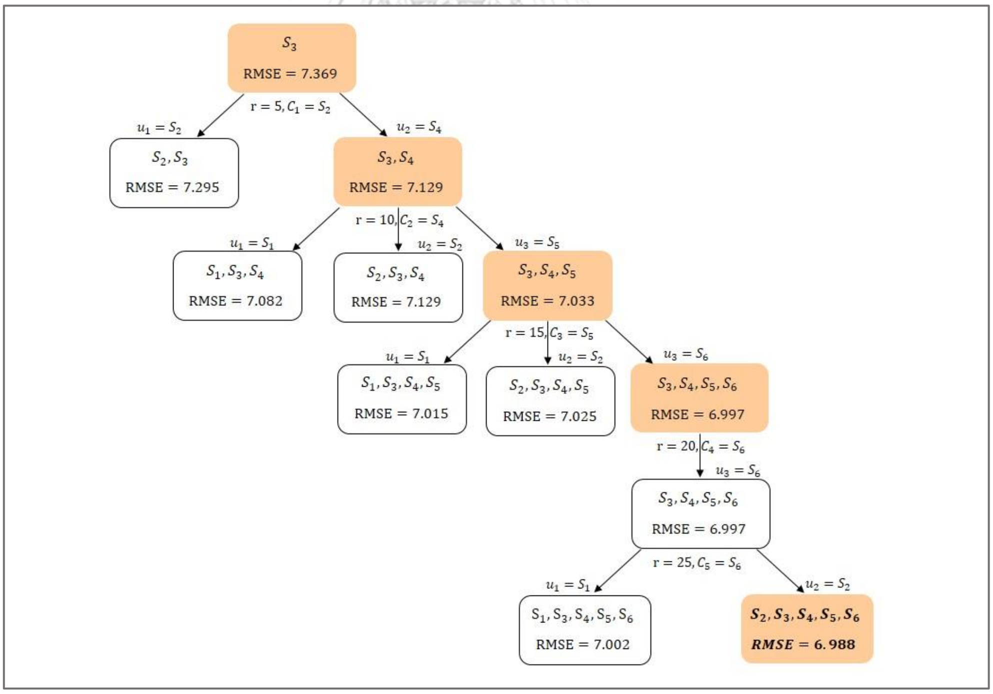
Observed stations	The PM_{10} concentration prediction results from the last phase					
	Feature selection	Supervised learning model	RMSE	MAE	MAPE (%)	R^2
S_1	BW	SVR	5.160	3.031	11.79	0.917
S_2	BW	MLR	9.343	5.741	17.64	0.867
S_3	FS	MLR	6.988	4.473	16.28	0.888
S_4	GA	SVR	6.154	3.956	14.16	0.918
S_5	GA	MLR	8.375	4.935	15.59	0.840
S_6	GA	SVR	7.864	4.966	14.79	0.881


S_1
RMSE = 5.663
r = 5, $C_1 = S_1$
$u_1 = S_1$
S_1
RMSE = 5.663
r = 10, $C_2 = S_1$
$u_1 = S_2$
$u_2 = S_3$
$u_3 = S_4$
S_1, S_2
RMSE = 5.438
S_1, S_3
RMSE = 5.415
S_1, S_4
RMSE = 5.287
r = 15, $C_3 = S_4$
$u_1 = S_2$
$u_2 = S_3$
$u_3 = S_5$
S_1, S_2, S_4
RMSE = 5.246
S_1, S_3, S_4
RMSE = 5.246
S_1, S_4, S_5
RMSE = 5.214
r = 20, $C_4 = S_5$
$u_1 = S_2$
$u_2 = S_3$
$u_3 = S_6$
S_1, S_2, S_4, S_5
RMSE = 5.184
S_1, S_3, S_4, S_5
RMSE = 5.182
S_1, S_4, S_5, S_6
RMSE = 5.184
r = 25, $C_5 = S_3$
$u_1 = S_2$
$u_2 = S_6$
S_1, S_2, S_3, S_4, S_5
RMSE = 5.160
S_1, S_3, S_4, S_5, S_6
RMSE = 5.183

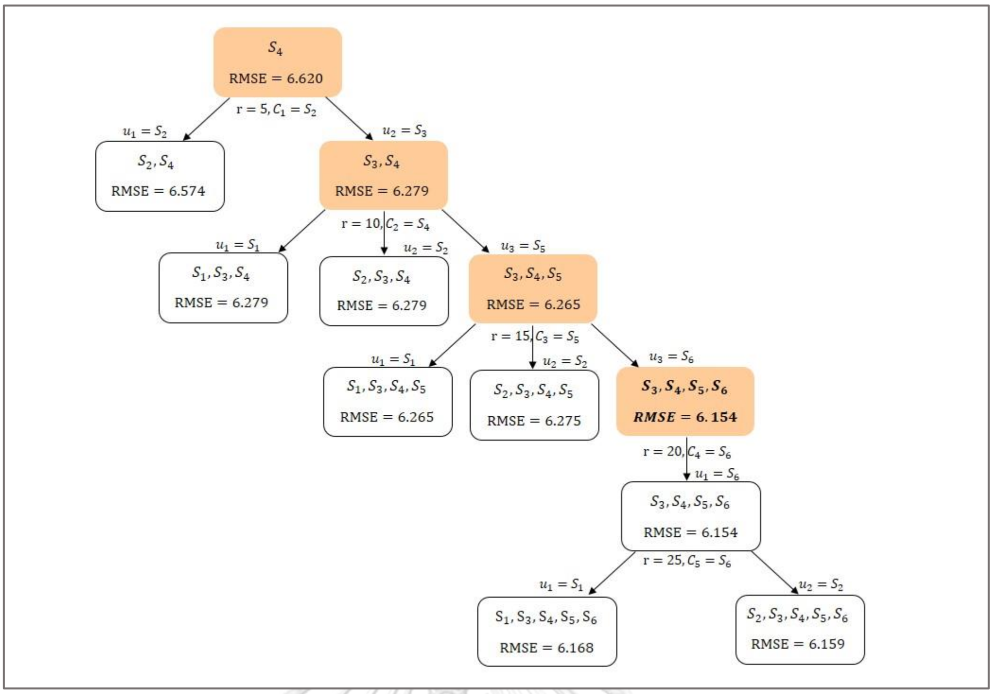
(a) Sensor of S_1


S_2
RMSE = 10.071
r = 5, $C_1 = S_2$
$u_1 = S_3$
$u_2 = S_4$
S_2, S_3
RMSE = 9.790
S_2, S_4
RMSE = 9.720
r = 10, $C_2 = S_4$
$u_1 = S_1$
$u_2 = S_3$
$u_3 = S_5$
S_1, S_2, S_4
RMSE = 9.666
S_2, S_3, S_4
RMSE = 9.657
S_2, S_4, S_5
RMSE = 9.557
r = 15, $C_3 = S_5$
$u_1 = S_1$
$u_2 = S_3$
$u_3 = S_6$
S_1, S_2, S_4, S_5
RMSE = 9.557
S_2, S_3, S_4, S_5
RMSE = 9.469
S_2, S_4, S_5, S_6
RMSE = 9.489
r = 20, $C_4 = S_3$
$u_1 = S_1$
$u_2 = S_6$
S_1, S_2, S_3, S_4, S_5
RMSE = 9.405
S_2, S_3, S_4, S_5, S_6
RMSE = 9.398
r = 25, $C_5 = S_6$
$u_1 = S_1$
$\boldsymbol{S_1, S_2, S_3, S_4, S_5, S_6}$
RMSE = 9.343

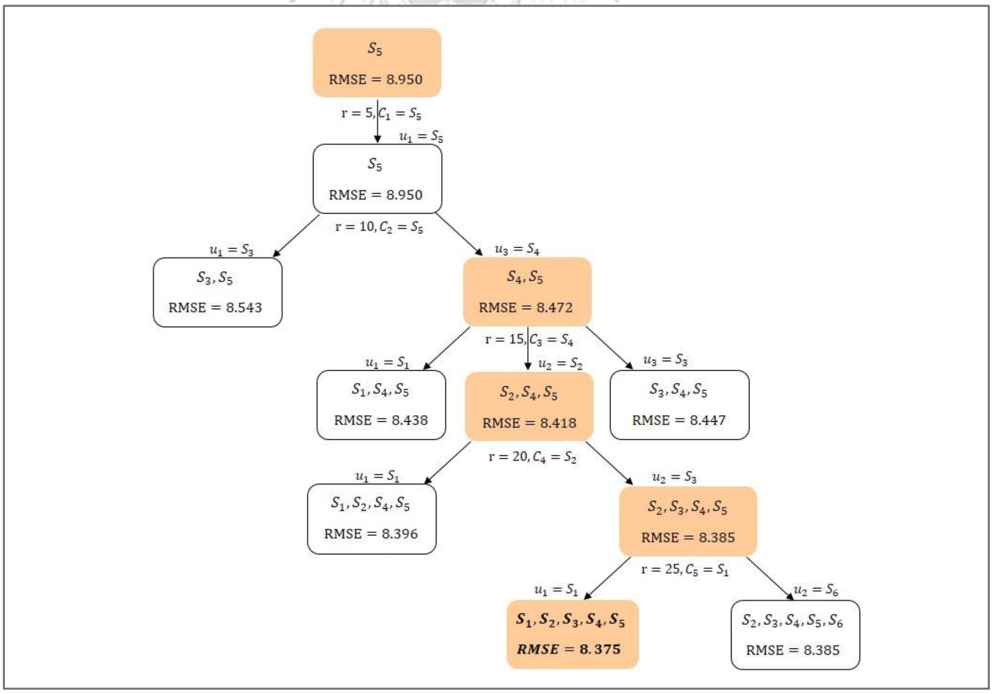
(b) Sensor of S_2


S_3
RMSE = 7.369
r = 5, $C_1 = S_2$
$u_1 = S_2$
$u_2 = S_4$
S_2, S_3
RMSE = 7.295
S_3, S_4
RMSE = 7.129
r = 10, $C_2 = S_4$
$u_1 = S_1$
$u_2 = S_2$
$u_3 = S_5$
S_1, S_3, S_4
RMSE = 7.082
S_2, S_3, S_4
RMSE = 7.129
S_3, S_4, S_5
RMSE = 7.033
r = 15, $C_3 = S_5$
$u_1 = S_1$
$u_2 = S_2$
$u_3 = S_6$
S_1, S_3, S_4, S_5
RMSE = 7.015
S_2, S_3, S_4, S_5
RMSE = 7.025
S_3, S_4, S_5, S_6
RMSE = 6.997
r = 20, $C_4 = S_6$
$u_3 = S_6$
S_3, S_4, S_5, S_6
RMSE = 6.997
r = 25, $C_5 = S_6$
$u_1 = S_1$
$u_2 = S_2$
S_1, S_3, S_4, S_5, S_6
RMSE = 7.002
$\boldsymbol{S_2, S_3, S_4, S_5, S_6}$
RMSE = 6.988

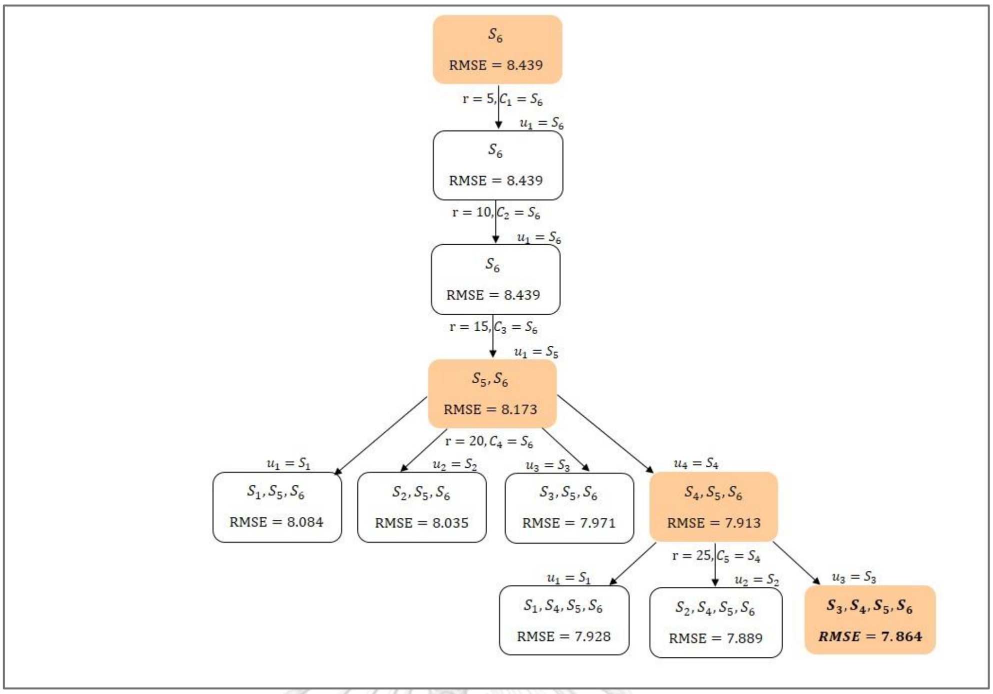
(c) Sensor of S_3


S_4
RMSE = 6.620
r = 5, $C_1 = S_2$
$u_1 = S_2$
S_2, S_4
RMSE = 6.574
$u_2 = S_3$
S_3, S_4
RMSE = 6.279
r = 10, $C_2 = S_4$
$u_1 = S_1$
S_1, S_3, S_4
RMSE = 6.279
$u_2 = S_2$
S_2, S_3, S_4
RMSE = 6.279
$u_3 = S_5$
S_3, S_4, S_5
RMSE = 6.265
r = 15, $C_3 = S_5$
$u_1 = S_1$
S_1, S_3, S_4, S_5
RMSE = 6.265
$u_2 = S_2$
S_2, S_3, S_4, S_5
RMSE = 6.275
$u_3 = S_6$
S_3, S_4, S_5, S_6
RMSE = 6.154
r = 20, $C_4 = S_6$
$u_1 = S_6$
S_3, S_4, S_5, S_6
RMSE = 6.154
r = 25, $C_5 = S_6$
$u_1 = S_1$
S_1, S_3, S_4, S_5, S_6
RMSE = 6.168
$u_2 = S_2$
S_2, S_3, S_4, S_5, S_6
RMSE = 6.159

(d) Sensor of S_4


S_5
RMSE = 8.950
r = 5, $C_1 = S_5$
$u_1 = S_5$
S_5
RMSE = 8.950
r = 10, $C_2 = S_5$
$u_1 = S_3$
S_3, S_5
RMSE = 8.543
$u_3 = S_4$
S_4, S_5
RMSE = 8.472
r = 15, $C_3 = S_4$
$u_1 = S_1$
S_1, S_4, S_5
RMSE = 8.438
$u_2 = S_2$
S_2, S_4, S_5
RMSE = 8.418
$u_3 = S_3$
S_3, S_4, S_5
RMSE = 8.447
r = 20, $C_4 = S_2$
$u_1 = S_1$
S_1, S_2, S_4, S_5
RMSE = 8.396
$u_2 = S_3$
S_2, S_3, S_4, S_5
RMSE = 8.385
r = 25, $C_5 = S_1$
$u_1 = S_1$
S_1, S_2, S_3, S_4, S_5
RMSE = 8.375
$u_2 = S_6$
S_2, S_3, S_4, S_5, S_6
RMSE = 8.385

(e) Sensor of S_5


S_6
RMSE = 8.439
r = 5, $C_1 = S_6$
$u_1 = S_6$
S_6
RMSE = 8.439
r = 10, $C_2 = S_6$
$u_1 = S_6$
S_6
RMSE = 8.439
r = 15, $C_3 = S_6$
$u_1 = S_5$
S_5, S_6
RMSE = 8.173
r = 20, $C_4 = S_6$
$u_1 = S_1$
S_1, S_5, S_6
RMSE = 8.084
$u_2 = S_2$
S_2, S_5, S_6
RMSE = 8.035
$u_3 = S_3$
S_3, S_5, S_6
RMSE = 7.971
$u_4 = S_4$
S_4, S_5, S_6
RMSE = 7.913
r = 25, $C_5 = S_4$
$u_1 = S_1$
S_1, S_4, S_5, S_6
RMSE = 7.928
$u_2 = S_2$
S_2, S_4, S_5, S_6
RMSE = 7.889
$u_3 = S_3$
S_3, S_4, S_5, S_6
RMSE = 7.864

(f) Sensor of S_6

Figure 34 The process of selecting data from the neighboring stations using the MDFS algorithm with the actual data for PM$_{10}$ concentration prediction of each monitoring station

CHAPTER 6 DISCUSSION AND CONCLUSION

In this research, the data of temperature, humidity, PM_1, $PM_{2.5}$, and PM_{10} in Nan Province of Thailand, which were collected every second from the sensor network system over a three-month period, were used to evaluate the proposed PM_{10} concentration prediction model. Most of the farmers burn their cornfields after harvesting during the period of three months in February, March, and April, which causes the PM_{10} concentration to be over the standard level. An initial set of 2136 data for each sensor where 70 percent of the data were used to train the model and 30 percent of the data were used to test the model. The optimal prediction was selected from the prediction results with the lowest RMSE. From the first phase when data of all features at the observed station were used, the PM_{10} concentration prediction results show that SVR yields the optimal predictions for all stations. In the second phase when feature selections were used to select only relevant features before passing data of the selected features to the supervised learning models, it can be noticed that the prediction results from all three models were improved. From RMSE in Table 13, the results from using GA with all three models are better for all six stations; however, FS and BE can also be used depending on the data and the location of each station. From the first two phases, the prediction used only data from the observed station. In fact, there are other factors that affect the PM_{10} concentration at each location; such as the environment of the nearby areas; thus in the third phase, we propose to include the air pollutant and air quality measurements of the neighboring stations, which were selected by the MDFS algorithm, into the prediction. From the experimental results, it can be noticed that GA is the most suitable feature selection method and MLP is the most suitable model to be used for the PM_{10} concentration prediction. Nonetheless, Table 14 reveals that FS is more suitable for station S_5 and S_6 than GA which is similar to the results in Table 13 where all three feature selections can be used for stations S_5 and S_6. The reason that different feature selection methods and supervised learning models were used for stations S_5 and S_6 is because these two stations are further away from other stations which decrease the correlation between the observed stations and their neighboring stations and weaken the effect of features from their neighboring stations. In the last phase, the actual data were used as input for the MDFS algorithm to predict PM_{10} concentration. From the experimental results, it can be noticed that GA with MLR or SVR are the most suitable models to be used for the PM_{10} concentration prediction. Nonetheless, Table 15 reveals that BW is more suitable for station S_1 and S_2, and FW is more suitable for station S_3 with MLR or SVR for PM_{10} concentration prediction which are different from the results in Table 13 and Table 14. The experimental results of the last phase are not as good as the PM_{10} concentration prediction results from the third phase; however, the actual data can also be used with the MDFS algorithm for PM_{10} concentration prediction at each monitoring station.

From the experimental results for PM_{10} concentration prediction, we can conclude that the proposed method with normalized data yields the optimal prediction when comparing to the others. In addition, the proposed method can predict the concentration of PM_{10} one hour ahead of any observed station without wind direction, wind speed, and other air quality data. Table 14 shows that the prediction results are better when using MLP with data of the GA-selected features from the sensor at the observed station and the sensors at its neighboring stations that were selected by the MDFS algorithm.

However, the proposed method still has some limitations. Firstly, the proposed method was developed based on the fact that the data are limited to only six sensors. We can use only six sensors out of 14 sensors from the sensor network system because other eight sensors have a lot of missing data. Secondly, the proposed method was developed using only two types of air quality and three types of air pollutant from the online sensor network system which monitors air pollution by assessing haze levels of Dr. Garavig Tanaksaranond at Department of Survey Engineering, Faculty of Engineering, Chulalongkorn University. In fact, there are more features that may improve the prediction precision, such as meteorological data and structural data of air quality which are not available. Finally, in order to make the predictions more accurate and usable in air pollution management, more features and more data may be required for a large-scale environmental system.

The proposed PM_{10} concentration prediction method can be used as the prediction framework for other air pollutants. Moreover, the proposed method can determine the relationship between features which can be used for particulate matter concentration management. Therefore, the framework that can be derived from the proposed method can be used for managing the environment and maintaining compliance with particulate matter concentration management regulations and policy.

As a conclusion, the novel method for predicting particulate matter concentration in Nan Province of Thailand that integrates the feature selection method, the supervised learning model, and the modified depth-first search algorithm is proposed. The root mean square error and the Pearson correlation coefficient clearly indicate that the proposed method yields the optimal prediction results as overall RMSE is less than 9.193 and overall R is above 0.841.

For future work, three main research topics will be covered. Firstly, the problem on missing data will be solved because particulate matter concentration prediction models always require complete data which are quite impossible in the real situation. Thus, these missing data can be filled by synthesizing the data to make data of eight sensors complete for the prediction for one-month period. Secondly, from the spread of the particulate matter which affects the whole world, we are interested in studying the associations between different sizes of particulate matters. Finally, if there is a large amount of data related to the environment then deep learning will be considered as a new method for particulate matter prediction.

REFERENCES

1. Friedlander, S.K. and P. Friedlander, *Smoke, Dust, and Haze: Fundamentals of Aerosol Dynamics*. Topics in chemical engineering. 2000: Oxford University Press.
2. Turner, M.C., et al., *Long-term ambient fine particulate matter air pollution and lung cancer in a large cohort of never-smokers.* American Journal of Respiratory and Critical Care Medicine, 2001. **184**(12): p. 1374-1381.
3. Dockery, D.W. and C.A. Pope, *Acute Respiratory Effects of Particulate Air Pollution.* Annual Review of Public Health, 1994. **15**(1): p. 107-132.
4. Kingdom, U., et al., *Air Quality Air Quality.* 2008.
5. Rom, W.N. and S.B. Markowitz, *Environmental and Occupational Medicine*. Environmental and Occupational Medicine. 2007: Wolters Kluwer/Lippincott Williams & Wilkins.
6. Reddy, S.V.J., *Climate Change: Myths and Realities*. Jeevananda Reddy.
7. Lin, K.-P., P.-F. Pai, and S.-L. Yang, *Forecasting concentrations of air pollutants by logarithm support vector regression with immune algorithms.* Applied Mathematics and Computation, 2011. **217**(12): p. 5318-5327.
8. Zhao, C.-X., et al., *Temporal and spatial distribution of PM2.5 and PM10 pollution status and the correlation of particulate matters and meteorological factors during winter and spring in Beijing.* Huan jing ke xue= Huanjing kexue, 2014. **35**(2): p. 418-427.
9. Li, H., et al., *Particulate matters pollution characteristic and the correlation between PM (PM 2.5, PM 10) and meteorological factors during the summer in Shijiazhuang.* Journal of Environmental Protection, 2015. **6**(05): p. 457.
10. Auder, B., et al., *Sequential aggregation of heterogeneous experts for PM10 forecasting.* Atmospheric Pollution Research, 2016. **7**(6): p. 1101-1109.
11. Hernandez, G., et al., *Temperature and humidity effects on particulate matter concentrations in a sub-tropical climate during winter.* 2017.
12. Franceschi, F., M. Cobo, and M. Figueredo, *Discovering relationships and forecasting PM10 and PM2.5 concentrations in Bogotá Colombia, using Artificial*

Neural Networks, Principal Component Analysis, and k-means clustering. Atmospheric Pollution Research, 2018. **9**(5): p. 912-922.

13. Garcia Nieto, P.J., et al., *PM10 concentration forecasting in the metropolitan area of Oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study.* Sci Total Environ, 2018. **621**: p. 753-761.
14. Shang, Z., et al., *A novel model for hourly PM2.5 concentration prediction based on CART and EELM.* Sci Total Environ, 2019. **651**(Pt 2): p. 3043-3052.
15. Maier, H.R. and G.C. Dandy, *Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications.* Environmental Modelling & Software, 2000. **15**(1): p. 101-124.
16. R., H.R., *A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression.* Biometrics, 1976. **32**(1): p. 49.
17. Draper, N.R. and H. Smith, *Applied Regression Analysis*. Wiley Series in Probability and Statistics. 2014: Wiley.
18. Darwin, C., *The origin of species by means of natural selection : or the preservation of favored races in the struggle for life, and the descent of man and selection in relation to sex.* The Modern library of the world's best books. 1936, New York: Modern library.
19. Darwin, C., *On the origin of species : by means of natural selection*. 2009, Floating Press: Waiheke Island.
20. Holland, J., *Adaptation in natural and artificial systems : an introductory analysis with application to biology.* Control and artificial intelligence, 1975.
21. Goldberg, D.E. and J.H. Holland, *Genetic algorithms and machine learning.* 1988.
22. Sarmady, S., *An investigation on genetic algorithm parameters.* School of Computer Science, Universiti Sains Malaysia, 2007. **126**.
23. De Jong, K.A. and W.M. Spears, *A formal analysis of the role of multi-point crossover in genetic algorithms.* Annals of mathematics and Artificial intelligence, 1992. **5**(1): p. 1-26.
24. Hassanat, A., et al., *Choosing Mutation and Crossover Ratios for Genetic Algorithms—A Review with a New Dynamic Approach.* Information, 2019. **10**(12).

25. Brownlee, J., *Master Machine Learning Algorithms: Discover How They Work and Implement Them From Scratch*. 2016: Jason Brownlee.
26. Chaloulakou, A., G. Grivas, and N. Spyrellis, *Neural network and multiple regression models for PM10 prediction in Athens: a comparative assessment.* J Air Waste Manag Assoc, 2003. **53**(10): p. 1183-90.
27. Legendre, A.M., *Nouvelles méthodes pour la détermination des orbites des comètes*. 1805: F. Didot.
28. Voukantsis, D., et al., *Intercomparison of air quality data using principal component analysis, and forecasting of PM(1)(0) and PM(2).(5) concentrations using artificial neural networks, in Thessaloniki and Helsinki.* Sci Total Environ, 2011. **409**(7): p. 1266-76.
29. Buscema, M., *Back propagation neural networks.* Subst Use Misuse, 1998. **33**(2): p. 233-70.
30. Cortes, C. and V. Vapnik, *Support-vector networks.* Machine learning, 1995. **20**(3): p. 273-297.
31. Mangasarian, O.L., *Nonlinear programming*. McGraw-Hill series in systems science. 1969, New York: McGraw-Hill.
32. McCormick, G.P., *Nonlinear programming : theory, algorithms, and applications*. 1983, New York: Wiley.
33. Vanderbei, R.J., *Book Review: Linear Programming: A Modern Integrated Analysis.* Interfaces, 1997. **27**(2): p. 120-122.
34. Even, S. and G. Even, *Graph Algorithms*. 2011: Cambridge University Press.
35. Wang, H. and J.Z. B, *and k -Center Problems in the Plane.* 2014. **3**: p. 3-14.
36. Wasserman, L., *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. 2013: Springer New York.
37. Freedman, D., R. Pisani, and R. Purves, *Statistics: Fourth International Student Edition*. International student edition. 2007: W.W. Norton & Company.
38. Feng, Y., et al., *Ozone concentration forecast method based on genetic algorithm optimized back propagation neural networks and support vector machine data classification.* Atmospheric Environment, 2011. **45**(11): p. 1979-1985.

39. Chai, T. and R.R. Draxler, *Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature.* Geoscientific Model Development, 2014. **7**(3): p. 1247-1250.
40. Sammut, C. and G.I. Webb, *Mean Absolute Error BT - Encyclopedia of Machine Learning*. 2010, Springer US: Boston, MA. p. 652.
41. Swamidass, P.M., *MAPE (mean absolute percentage error) Mean Absolute Percentage ErroR (MAPE) BT - Encyclopedia of Production and Manufacturing Management*. 2000, Springer US: Boston, MA. p. 462-462.
42. Chunitiphisan, S., et al., *Particulate Matter Monitoring Using Inexpensive Sensors and Internet GIS: A Case Study in Nan, Thailand.* Engineering Journal, 2018. **22**(2): p. 25-37.
43. Lee, J. and J.H. Maunsell, *A normalization model of attentional modulation of single unit responses.* PLoS One, 2009. **4**(2): p. e4651.
44. Reynolds, J.H. and D.J. Heeger, *The normalization model of attention.* Neuron, 2009. **61**(2): p. 168-85.
45. Goldberg, D.E. and K. Deb, *A comparative analysis of selection schemes used in genetic algorithms*, in *Foundations of genetic algorithms*. 1991. p. 69-93.

APPENDIX

Observed station: $\boldsymbol{S_1}$

For $r = 5, C_1 = S_1$:

Where $u_1 = S_1$

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	1	0	0	1	0	1	0	1
Humidity of S_1	1	1	1	0	0	0	1	0	1
PM_1 of S_1	0	0	0	0	1	1	0	1	0
$PM_{2.5}$ of S_1	0	1	0	1	0	0	0	0	0
PM_{10} of S_1	1	1	1	1	1	1	1	1	1
RMSE	5.693	5.693	5.693	5.652	5.618	5.61	5.663	5.664	5.663
R	0.901	0.9	0.901	0.901	0.902	0.902	0.901	0.901	0.901

For $r = 10, C_2 = S_1$:

Where $u_1 = S_2$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	1	1	0	1	0	0	1	1
Humidity of S_1	1	0	0	0	1	1	1	1	1
PM_1 of S_1	0	0	0	1	1	1	1	0	0
$PM_{2.5}$ of S_1	0	0	1	0	1	1	0	0	0
PM_{10} of S_1	1	1	1	1	1	1	1	1	1
Temperature of S_2	0	1	0	0	0	0	0	1	1
Humidity of S_2	0	1	1	0	0	0	0	1	1
PM_1 of S_2	0	1	0	1	1	1	1	0	0
$PM_{2.5}$ of S_2	0	1	0	0	1	0	1	0	0
PM_{10} of S_2	1	1	1	0	1	1	1	1	1
RMSE	5.481	5.445	5.461	5.368	5.408	5.34	5.453	5.438	5.438
R	0.909	0.908	0.908	0.911	0.91	0.912	0.908	0.908	0.908

Where $u_2 = S_3$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	0	0	0	1	0	1	1	1
Humidity of S_1	1	0	0	0	1	1	1	0	0
PM_1 of S_1	0	0	0	1	1	1	0	0	0
$PM_{2.5}$ of S_1	0	0	1	0	1	0	0	0	0
PM_{10} of S_1	1	1	1	1	1	1	1	1	1
Temperature of S_3	0	1	1	0	0	0	1	1	1
Humidity of S_3	0	1	1	0	0	0	1	1	1
PM_1 of S_3	0	0	0	1	1	0	1	1	1
$PM_{2.5}$ of S_3	1	0	0	0	1	1	1	1	1
PM_{10} of S_3	1	1	1	0	1	1	1	1	1
RMSE	5.452	5.433	5.433	5.395	5.444	5.347	5.423	5.415	5.415
R	0.91	0.909	0.909	0.909	0.909	0.911	0.909	0.909	0.909

Where $u_3 = S_4$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	1	1	0	1	1	1	1	1
Humidity of S_1	1	0	0	0	1	1	1	1	1
PM_1 of S_1	0	0	0	1	1	1	0	0	0
$PM_{2.5}$ of S_1	0	0	0	0	1	1	0	0	0
PM_{10} of S_1	1	1	1	1	1	1	1	1	1
Temperature of S_4	1	1	1	0	0	0	1	1	1
Humidity of S_4	1	1	1	0	0	0	1	1	1
PM_1 of S_4	0	0	0	1	1	1	1	1	1
$PM_{2.5}$ of S_4	1	1	1	0	1	1	1	1	1
PM_{10} of S_4	1	1	1	0	1	1	1	1	1
RMSE	5.309	5.287	5.287	5.32	5.292	5.255	5.303	5.303	5.303
R	0.913	0.913	0.913	0.912	0.914	0.914	0.913	0.913	0.913

For $r = 15, C_3 = S_4$:

Where $u_1 = S_3$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	1	0	0	1	0	1	1	1
Humidity of S_1	1	0	0	0	1	0	1	0	0
PM$_1$ of S_1	0	0	0	1	1	1	0	0	0
PM$_{2.5}$ of S_1	0	0	1	0	1	0	0	0	0
PM$_{10}$ of S_1	1	1	1	1	1	1	1	1	1
Temperature of S_3	0	0	1	0	0	0	1	1	0
Humidity of S_3	0	0	1	0	1	0	0	1	1
PM$_1$ of S_3	1	0	0	1	1	1	1	1	1
PM$_{2.5}$ of S_3	0	1	0	0	0	1	1	1	1
PM$_{10}$ of S_3	0	0	1	0	1	1	1	1	1
Temperature of S_4	1	1	0	0	0	1	1	1	1
Humidity of S_4	1	1	0	0	0	0	1	1	1
PM$_1$ of S_4	0	0	0	1	1	1	0	0	0
PM$_{2.5}$ of S_4	1	1	0	0	1	1	0	1	0

PM$_{10}$ of S_4	1	1	1	1	1	1	1	1	1
RMSE	5.301	5.285	5.328	5.282	5.254	5.162	5.254	5.246	5.248
R	0.913	0.913	0.912	0.914	0.914	0.918	0.914	0.915	0.915

Where $u_2 = S_2$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	1	0	0	1	1	0	1	1
Humidity of S_1	1	0	0	0	1	0	1	0	0
PM$_1$ of S_1	0	0	0	1	1	1	0	0	0
PM$_{2.5}$ of S_1	0	0	1	0	1	0	0	0	0
PM$_{10}$ of S_1	1	1	1	0	1	1	1	1	1
Temperature of S_2	0	1	0	0	1	1	0	0	0
Humidity of S_2	0	1	0	0	1	0	1	1	1
PM$_1$ of S_2	0	1	0	0	1	1	1	0	1
PM$_{2.5}$ of S_2	0	1	0	0	1	1	1	1	1
PM$_{10}$ of S_2	1	1	1	0	1	1	1	1	1

Temperature of S_4	0	0	1	0	0	0	1	1	1
Humidity of S_4	0	1	1	0	1	0	1	1	1
PM_1 of S_4	1	0	0	1	1	0	1	1	0
$PM_{2.5}$ of S_4	1	0	0	0	1	1	0	1	1
PM_{10} of S_4	1	1	1	1	1	1	1	1	1
RMSE	5.284	5.288	5.297	5.375	5.336	5.200	5.248	5.246	5.249
R	0.915	0.913	0.913	0.91	0.912	0.916	0.914	0.914	0.914

Where $u_3 = S_5$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	0	0	0	1	0	0	1	0
Humidity of S_1	1	0	0	0	1	0	1	0	1
PM_1 of S_1	0	0	0	0	0	1	0	0	0
$PM_{2.5}$ of S_1	0	0	0	0	1	0	0	0	0
PM_{10} of S_1	1	1	1	1	1	1	1	1	1
Temperature of S_4	0	1	1	0	1	1	1	1	1

Humidity of S_4	0	0	1	0	1	0	1	1	1
PM_1 of S_4	0	0	0	1	1	0	1	1	1
$PM_{2.5}$ of S_4	1	1	1	0	1	0	1	1	1
PM_{10} of S_4	1	1	1	0	1	1	1	1	1
Temperature of S_5	0	1	0	0	1	0	0	0	0
Humidity of S_5	0	1	0	0	1	1	0	0	0
PM_1 of S_5	0	0	0	0	1	0	1	1	1
$PM_{2.5}$ of S_5	0	1	0	0	1	1	1	1	1
PM_{10} of S_5	1	0	1	1	1	1	1	1	1
RMSE	5.262	5.276	5.249	5.295	5.387	5.162	5.214	5.22	5.214
R	0.916	0.914	0.915	0.913	0.911	0.918	0.915	0.915	0.915

For $r = 20, C_4 = S_3$:

Where $u_1 = S_2$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	1	1	0	1	0	0	1	1
Humidity of S_1	1	0	0	0	1	0	1	0	0
PM_1 of S_1	0	0	0	1	1	1	0	0	0
$PM_{2.5}$ of S_1	0	0	0	0	1	0	0	0	0
PM_{10} of S_1	1	1	1	0	1	0	1	1	1
Temperature of S_2	0	1	1	0	1	0	0	0	0
Humidity of S_2	0	1	1	0	0	1	0	1	1
PM_1 of S_2	0	1	0	0	1	0	0	0	0
$PM_{2.5}$ of S_2	0	1	0	0	1	1	0	0	0
PM_{10} of S_2	1	1	1	0	1	1	1	1	1
Temperature of S_3	0	0	0	0	1	1	0	0	0
Humidity of S_3	0	0	0	0	1	1	1	1	1
PM_1 of S_3	0	0	0	0	1	0	1	1	1
$PM_{2.5}$ of S_3	0	0	0	0	1	0	1	1	1

PM_{10} of S_3	0	0	0	1	1	1	1	1	1
Temperature of S_4	0	0	0	0	1	0	1	1	1
Humidity of S_4	0	1	1	0	1	0	1	1	1
PM_1 of S_4	1	0	0	0	1	0	0	0	0
$PM_{2.5}$ of S_4	1	1	1	1	1	0	0	0	0
PM_{10} of S_4	1	1	1	1	1	1	1	1	1
RMSE	5.284	5.291	5.272	5.305	5.300	5.322	5.230	5.209	5.209
R	0.915	0.913	0.914	0.913	0.913	0.913	0.915	0.916	0.916

Where $u_2 = S_5$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	0	0	0	1	0	0	0	1
Humidity of S_1	1	0	0	0	1	0	1	1	0
PM_1 of S_1	0	0	0	1	1	1	0	0	0
$PM_{2.5}$ of S_1	0	0	0	1	1	0	0	0	0
PM_{10} of S_1	1	1	1	1	1	1	1	1	1

Temperature of S_3	0	0	0	0	0	0	0	0	1
Humidity of S_3	0	1	0	1	1	0	0	1	1
PM_1 of S_3	0	0	0	0	1	1	1	1	1
$PM_{2.5}$ of S_3	0	1	0	0	1	0	1	1	1
PM_{10} of S_3	0	0	0	1	1	1	1	1	1
Temperature of S_4	0	1	1	0	1	0	1	1	1
Humidity of S_4	0	1	1	0	1	1	1	1	1
PM_1 of S_4	0	0	0	1	1	1	0	0	0
$PM_{2.5}$ of S_4	1	1	1	1	1	1	1	0	0
PM_{10} of S_4	1	1	1	1	1	1	1	1	1
Temperature of S_5	0	1	0	0	1	0	0	0	0
Humidity of S_5	0	1	0	0	0	0	0	0	0
PM_1 of S_5	0	1	0	1	1	0	0	0	0
$PM_{2.5}$ of S_5	0	0	0	0	0	0	0	1	1
PM_{10} of S_5	1	0	1	1	1	1	1	1	1
RMSE	5.262	5.296	5.249	5.143	5.368	5.119	5.190	5.182	5.178
R	0.916	0.913	0.915	0.918	0.911	0.919	0.916	0.916	0.917

Where $u_3 = S_6$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	0	1	1	1	0	0	0	1
Humidity of S_1	1	0	0	0	1	1	1	1	0
PM_1 of S_1	0	0	0	1	1	1	0	0	0
$PM_{2.5}$ of S_1	0	0	0	0	1	0	0	0	0
PM_{10} of S_1	1	1	1	1	1	1	1	1	1
Temperature of S_3	0	0	0	0	1	0	0	0	0
Humidity of S_3	0	0	0	0	1	0	1	1	1
PM_1 of S_3	1	0	0	1	1	1	1	1	1
$PM_{2.5}$ of S_3	0	0	0	0	1	1	0	1	0
PM_{10} of S_3	0	0	0	1	1	0	1	1	1
Temperature of S_4	0	1	1	0	1	0	0	1	1
Humidity of S_4	0	1	1	0	0	1	1	1	1
PM_1 of S_4	0	1	0	0	1	1	0	0	0
$PM_{2.5}$ of S_4	1	1	1	1	1	1	0	0	0
PM_{10} of S_4	1	1	1	1	1	1	1	1	1

Temperature of S_6	1	0	0	0	1	0	1	0	1
Humidity of S_6	0	1	0	0	1	0	0	0	0
PM_1 of S_6	0	1	0	0	1	1	1	1	1
$PM_{2.5}$ of S_6	0	1	0	1	1	0	0	1	1
PM_{10} of S_6	1	1	1	0	1	1	1	1	1
RMSE	5.253	5.280	5.254	5.170	5.300	5.141	5.221	5.212	5.215
R	0.914	0.914	0.914	0.917	0.914	0.918	0.915	0.915	0.915

For $r = 20, C_4 = S_5$:

Where $u_1 = S_2$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	1	1	0	1	1	0	0	0
Humidity of S_1	1	0	0	0	0	0	1	0	1
PM_1 of S_1	0	0	0	1	1	1	0	0	0
$PM_{2.5}$ of S_1	0	0	0	1	1	1	0	0	0
PM_{10} of S_1	1	1	1	0	1	1	1	1	1
Temperature of S_2	0	1	0	0	1	0	0	0	0
Humidity of S_2	0	1	0	0	0	0	0	1	1
PM_1 of S_2	0	1	0	0	1	0	0	0	0
$PM_{2.5}$ of S_2	0	1	0	0	1	1	1	1	0
PM_{10} of S_2	1	1	1	0	1	0	1	1	1
Temperature of S_4	0	0	0	0	1	0	1	1	1
Humidity of S_4	0	1	1	0	1	0	1	1	1
PM_1 of S_4	0	0	0	0	1	1	1	1	1
$PM_{2.5}$ of S_4	1	0	1	1	1	1	0	1	1

PM_{10} of S_4	1	1	1	1	1	1	1	1	1
Temperature of S_5	0	0	0	0	1	0	0	0	0
Humidity of S_5	0	0	0	0	1	0	0	0	0
PM_1 of S_5	0	1	0	0	0	0	0	1	0
$\mathrm{PM}_{2.5}$ of S_5	0	0	0	0	0	1	1	1	1
PM_{10} of S_5	1	0	1	1	1	1	1	1	1
RMSE	5.246	5.27	5.238	5.276	5.357	5.186	5.186	5.183	5.183
R	0.916	0.914	0.915	0.914	0.911	0.917	0.916	0.916	0.916

Where $u_2 = S_3$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	0	0	0	1	0	0	0	1
Humidity of S_1	1	0	0	0	1	0	1	1	0
PM_1 of S_1	0	0	0	1	1	1	0	0	0
$\mathrm{PM}_{2.5}$ of S_1	0	0	0	1	1	0	0	0	0
PM_{10} of S_1	1	1	1	1	1	1	1	1	1

Temperature of S_3	0	0	0	0	0	0	0	0	1
Humidity of S_3	0	1	0	1	1	0	0	1	1
PM_1 of S_3	0	0	0	0	1	1	1	1	1
$\text{PM}_{2.5}$ of S_3	0	1	0	0	1	0	1	1	1
PM_{10} of S_3	0	0	0	1	1	1	1	1	1
Temperature of S_4	0	1	1	0	1	0	1	1	1
Humidity of S_4	0	1	1	0	1	1	1	1	1
PM_1 of S_4	0	0	0	1	1	1	0	0	0
$\text{PM}_{2.5}$ of S_4	1	1	1	1	1	1	1	0	0
PM_{10} of S_4	1	1	1	1	1	1	1	1	1
Temperature of S_5	0	1	0	0	1	0	0	0	0
Humidity of S_5	0	1	0	0	0	0	0	0	0
PM_1 of S_5	0	1	0	1	1	0	0	0	0
$\text{PM}_{2.5}$ of S_5	0	0	0	0	0	0	0	1	1
PM_{10} of S_5	1	0	1	1	1	1	1	1	1
RMSE	5.262	5.296	5.249	5.143	5.368	5.119	5.190	5.182	5.178
R	0.916	0.913	0.915	0.918	0.911	0.919	0.916	0.916	0.917

Where $u_3 = S_6$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	1	0	0	1	0	0	0	0
Humidity of S_1	1	0	0	0	0	0	1	0	1
PM_1 of S_1	0	0	0	0	1	1	0	0	0
$PM_{2.5}$ of S_1	0	0	0	0	1	1	0	0	0
PM_{10} of S_1	1	1	1	1	1	1	1	1	1
Temperature of S_4	0	1	1	0	1	0	1	1	0
Humidity of S_4	0	1	1	0	0	0	1	1	1
PM_1 of S_4	0	1	0	1	1	1	1	1	0
$PM_{2.5}$ of S_4	1	1	1	0	1	0	0	0	1
PM_{10} of S_4	1	1	1	0	1	1	1	1	1
Temperature of S_5	0	1	1	0	1	0	0	0	0
Humidity of S_5	0	1	1	0	1	1	0	0	0
PM_1 of S_5	0	0	0	0	1	1	0	0	1
$PM_{2.5}$ of S_5	0	1	0	0	1	0	0	1	1
PM_{10} of S_5	1	0	1	0	1	0	1	1	1

Temperature of S_6	1	0	0	0	1	1	1	1	1
Humidity of S_6	0	1	1	1	1	0	0	0	0
PM$_1$ of S_6	0	1	1	1	1	0	0	1	1
PM$_{2.5}$ of S_6	0	1	1	1	0	1	1	1	1
PM$_{10}$ of S_6	1	1	1	0	1	0	1	1	1
RMSE	5.231	5.270	5.262	5.246	5.217	5.104	5.189	5.184	5.182
R	0.915	0.914	0.914	0.915	0.916	0.919	0.916	0.916	0.916

For $r = 25, C_5 = S_6$:

Where $u_1 = S_2$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA
Temperature of S_1	0	1	0	0	1	0	0	0	0
Humidity of S_1	1	0	0	0	1	0	1	0	0
PM_1 of S_1	0	0	0	1	1	0	0	0	0
$PM_{2.5}$ of S_1	0	0	0	0	1	0	0	0	0
PM_{10} of S_1	1	1	1	1	1	1	1	1	1
Temperature of S_2	0	1	0	0	1	0	0	0	0
Humidity of S_2	0	1	0	0	1	0	0	1	0
PM_1 of S_2	0	1	1	0	1	1	0	0	0
$PM_{2.5}$ of S_2	0	1	1	0	1	1	0	0	1
PM_{10} of S_2	1	1	1	1	1	1	1	1	1
Temperature of S_4	0	0	0	0	1	1	0	1	1
Humidity of S_4	0	1	1	0	1	1	1	1	1
PM_1 of S_4	1	0	0	1	1	0	1	1	1
$PM_{2.5}$ of S_4	1	0	0	0	0	0	1	1	0

PM$_{10}$ of S_4	1	1	1	1	1	1	1	1	1
Temperature of S_5	0	1	0	0	1	0	0	0	0
Humidity of S_5	0	0	0	0	1	0	0	0	0
PM$_1$ of S_5	0	0	1	0	0	0	0	1	0
PM$_{2.5}$ of S_5	0	1	1	0	1	1	0	1	1
PM$_{10}$ of S_5	1	0	1	0	1	1	1	1	1
Temperature of S_6	0	0	1	0	1	1	1	1	1
Humidity of S_6	0	1	1	1	1	1	0	0	0
PM$_1$ of S_6	0	1	0	0	1	0	0	1	0
PM$_{2.5}$ of S_6	0	1	0	0	1	0	0	1	1
PM$_{10}$ of S_6	1	1	1	0	1	1	1	1	1
RMSE	5.217	5.261	5.231	5.279	5.300	5.164	5.179	5.169	5.172
R	0.917	0.914	0.916	0.913	0.913	0.917	0.917	0.917	0.917

Where $u_2 = S_3$,

Supervised learning model	MLR			MLP			SVR		
Feature selection	FW	BW	GA	FW	BW	GA	FW	BW	GA

Temperature of S_1	0	1	0	0	1	0	0	0	0
Humidity of S_1	1	0	0	0	1	0	1	1	0
PM_1 of S_1	0	0	0	1	1	1	0	0	0
$PM_{2.5}$ of S_1	0	0	0	0	1	0	0	0	0
PM_{10} of S_1	1	1	1	1	1	1	1	1	1
Temperature of S_3	0	0	0	0	1	0	0	0	0
Humidity of S_3	0	1	0	0	1	0	0	1	0
PM_1 of S_3	0	1	0	0	1	1	0	1	1
$PM_{2.5}$ of S_3	0	0	0	0	1	1	1	1	1
PM_{10} of S_3	0	0	0	0	1	0	1	1	1
Temperature of S_4	0	1	0	0	1	0	0	1	1
Humidity of S_4	0	1	1	0	0	0	1	1	1
PM_1 of S_4	0	0	0	0	1	1	0	0	1
$PM_{2.5}$ of S_4	1	1	1	0	1	1	0	1	0
PM_{10} of S_4	1	1	1	1	1	1	1	1	1
Temperature of S_5	0	1	0	0	1	0	0	0	0
Humidity of S_5	0	1	0	0	1	0	0	0	0
PM_1 of S_5	0	0	0	0	0	1	0	0	0

PM$_{2.5}$ of S_5	0	1	0	0	1	0	0	0	0
PM$_{10}$ of S_5	1	0	1	0	1	0	1	1	1
Temperature of S_6	1	0	1	0	1	1	1	1	1
Humidity of S_6	0	1	1	1	1	0	1	0	0
PM$_1$ of S_6	0	1	0	0	1	0	0	1	1
PM$_{2.5}$ of S_6	0	1	0	0	1	0	0	1	1
PM$_{10}$ of S_6	1	1	1	0	1	1	1	1	1
RMSE	5.231	5.278	5.229	5.311	5.314	5.089	5.183	5.161	5.154
R	0.915	0.914	0.915	0.913	0.912	0.919	0.916	0.917	0.917

VITA

NAME	Miss Chadaphim Photphanloet
DATE OF BIRTH	16 February 1991
PLACE OF BIRTH	Nakhon Si Thammarat, Thailand
INSTITUTIONS ATTENDED	B.Sc. Mathematics, Prince of Songkla University, 2014 M.Sc. Applied Mathematics and Computational Science, Chulalongkorn University, 2016
HOME ADDRESS	18111/76 Village No.11, Kapang, Thung Song, Nakhon Si Thammarat 80310
PUBLICATION	1. Photphanloet, C., and Kittipoom, P. (2014). Necessary Conditions of Existence of Solutions (Poster). The 9th Conference on Science and Technology for Youths, Bangkok, Thailand. 2. Photphanloet, C., Treeratanajaru, W., Cooharojananone, N., and Lipikorn, R. (2016). Biochemical Oxygen Demand Prediction for Chaophraya River Using α –Trimmed ARIMA Model, in The 13th International Joint Conference on Computer Science and Software Engineering, 2016. 3. Photphanloet, C., and Lipikorn, R. (2020). PM10 concentration forecast using modified depth-first search and supervised learning neural network, in Science of the Total Environment, ISSN: 0048-9697, Volume 727, 2020.
AWARD RECEIVED	Science Achievement Scholarship of Thailand (SAST)