AIR POLLUTION INDEX ESTIMATION MODEL BASED ON ARTIFICIAL NEURAL NETWORK

AL-SUBAIE MOHAMMED NASSER

UNIVERSITI SAINS MALAYSIA

2021

AIR POLLUTION INDEX ESTIMATION MODEL BASED ON ARTIFICIAL NEURAL NETWORK

by

AL-SUBAIE MOHAMMED NASSER

Project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Chemical Engineering

2021

ACKNOWLEDGEMENT

Foremost, I would like to express my sincere gratitude to my supervisor Dr. Norazwan Bin Md Nor for the continuous support of my study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this report. I could not have imagined having a better advisor and mentor for my final year project. Moreover, I would like to extend my gratitude to final year project coordinator,Prof. Dr. Mohd. Roslee Othman and all inviting lecturers from the School of Chemical Engineering ,Universiti Sains Malaysia for their knowledge sharing, support and help during the completion of this report.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF SYMBOLS

C	constant
RMSE	Root mean square errors
R^2	Coefficient of determination
S	the resulting weighted sum.
Y	the output
X	the input

LIST OF ABBREVIATION

ANN	Artificial neural network
CO	Carbon monoxide
KPCA	Kernel principal component analysis
MPCA	Multiway principal component analysis
NOx	Nitrogen oxides
O_3	Ozone
PCA	Principal component analysis
PM10	Particulate matter 10
SO_2	Sulphur dioxides

MODEL PENGANGGARAN INDEKS PENCEMARAN UDARA BERDASARKAN RANGKAIAN NEURAL BUATAN

ABSTRAK

Usaha pemeliharaan alam sekitar sering menghadapi masalah yang kompleks disebabkan ia berhubungkait dengan pembolehubah yang pelbagai. Walaubagaimanapun, memilih struktur model yang sesuai dan algoritma latihan yang optima dengan tahap kerumitan minima adalah sangat penting.

Jadi, perlaksanaan kaedah pengurangan dimensi berdasarkan kepada prinsip analisis komponen berbilang hala (MPCA) telah digunakan.Tiga model telah dibina pada permulaan iaitu model ramalan ozon, model ramalan zarah halus yang bersaiz lebih kecil dari 10 mikron (PM10) dan model indeks pencemaran udara. Enam input telah digunakan untuk model ozon dan PM10 iaitu nitrogen oksida (NO_x), karbon monoksida (CO), sulfur dioksida (SO_2), halaju angin, suhu udara dan kelembapan relatif. Kemudian, ozon dan PM10 digunakan sebagai input didalam model indeks pencemaran udara. Keputusan menunjukkan bahawa aplikasi MPCA tidak memberikan pembaikan yang signifikan dalam keseluruhan faktor korelasi disebabkan oleh ketidaklurusan data yang tinggi.

AIR POLLUTION INDEX ESTIMATION MODEL BASED ON ARTIFICIAL NEURAL NETWORK

ABSTRACT

Environmental conservation efforts are always dealing with a complex problem because it involves a large number of variables. However, choosing a correct model structure, and optimum training algorithm with minimum complexity is crucial. Therefore, a dimensional reduction method was implemented based on the multiway principal component analysis (MPCA) method. Three models were built in first part; ozone estimation model, particulate matter 10 (PM10) estimation model, and air pollution index (API) estimation model. Six inputs were used in ozone and PM10 models, which are nitrogen oxides(NO_x), carbon monoxide (CO), sulphur dioxides (SO_2), wind speed, air temperature, and relative humidity. After that, ozone and PM 10 were used as input to the API estimation model. The result shows that the implementation of the MPCA has insignificant improvement on the overall correlation factor due to the high nonlinearity of data.

CHAPTER 1

INTRODUCTION

1.1 Research Background

Air pollution is a major problem that needs to be tackled urgently and seriously because it is one of the most significant factors contributing to the quality of health and life. Air pollution typically defined as a situation in which air pollutants are present at sufficiently high concentrations in the atmosphere above their normal ambient concentrations (Seinfeld and Pandis, 2016). Air pollution often pose serious issues especially in urban areas where there are very high population density and manufacturing industries. The primary sources of air pollutants in Malaysia are mobile, stationary, and cross-border sources (Afzali et al., 2012, DOE, 2014). As illustrate in Figure 1.1, the mobile sources are mainly attributed to motor vehicle emissions ,which contributed 82 % of the air emission load in Peninsular Malaysia (Al Madhoun et al., 2012). The emissions of power stations, industrial fuel burning, industrial production processes, open burning at solid waste disposal sites ,and domestic and commercial furnaces, which contributed to 9, 5, 3, 0.8, and 0.2 % of the overall air pollution load , respectively, are types of stationary sources in Peninsular Malaysia (Al Madhoun et al., 2012). Uncontrolled burning of plantation owners and farmers in Indonesia to clear land in 2019 is a case of a cross-border source of emissions affecting neighbouring countries.

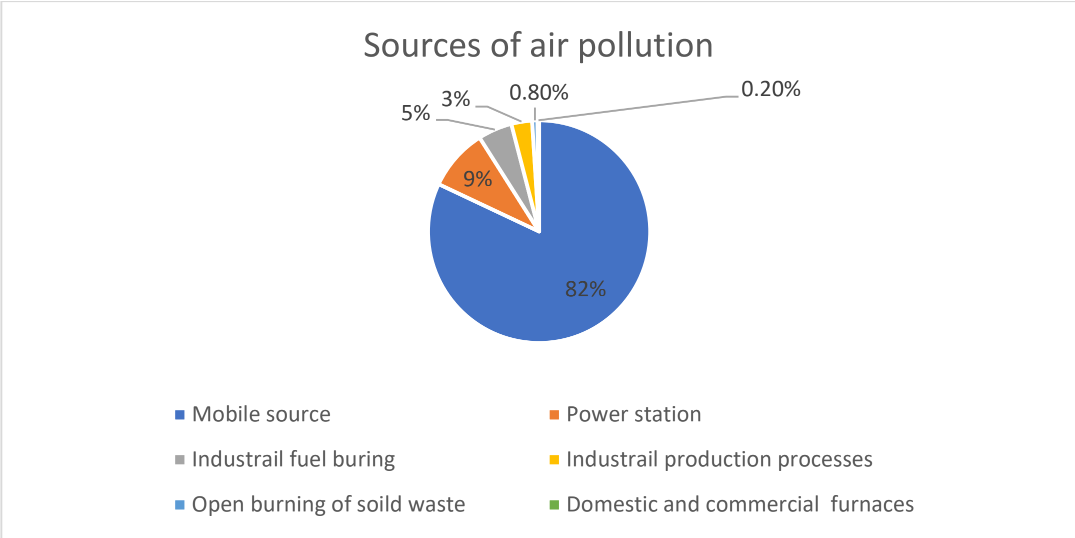

Figure 1.1: *Sources of air pollution in Malaysia*.

The Air Pollutant Index (API) can be defined as a number used to report air quality in relation to its impact on human health (Bishoi et al., 2009). The API is a useful tool to increase awareness about how clean or contaminated the environment is and what health consequences might be of interest to us. The API is a daily monitoring index measured using the five primary air pollutants (ozone (O_3), particulate matter below 10 µm (PM10), carbon monoxide (CO), sulphur dioxide (SO_2) and nitrogen dioxide (NO_2)) under the Clean Air Act (DOE, 2014). For each of these pollutants, the Malaysian Department of Environment (DOE) has set up the API and established national air quality standards through the Recommended Malaysian Air Quality Guidelines (RMAQG) to recover and maintain air quality and protect public health. The higher the API level, the higher the level of air pollution and the greater the health issue.

1.2 Problem Statement

The estimation of air pollutants index using a simple mathematical formula faces various challenges. It is incapable to capture the non-linear relationship among different variables due to the complexity of the data (Afzali et al., 2012). Air quality data are discovered as stochastic time series, so forecast can be made based on historical data (Sharma, 1998). Air pollution index estimation is important for proper planning and strategic control to minimise human health problems.

For air quality modelling, many approaches based on black-box have been used. The artificial neural network (ANN) is one of the black-box methods used to model environmental systems and has become a common and useful tool. ANN is a more flexible, effective, and reliable technique, as it has an appealing function that is similar to the brain. ANN can be trained to recognise non-linear patterns between the values of input and output and can solve complex problems far more quickly and able to train effectively when presented with new data (Abd Rahman et al., 2013, Zhang et al., 1998). The ANN method may also perform tasks based on training or initial experience, and an algorithm is not necessary to solve a problem. This is because by learning it can produce its distribution of the weights of the links and is easily inserted into the current technology (Azid et al., 2014a).

In this project, the air pollution index estimation model will be developed based on the combination of artificial neural networks and dimensional reduction methods.

1.3 Objectives

The aims of this research are:

i. To study the effectiveness of multiway principal component analysis as a dimensional reduction method for the ANN estimation model.

ii. To develop and demonstrate the potential of artificial neural networks for the API estimation model, called MPCA-ANN estimation model

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, a survey of air quality estimation model based on artificial neural network (ANN) is presented. The literature surveys on dimensional reduction methods such as principal component analysis (PCA), kernel principal component analysis (KPCA), and multiway principal component analysis (MPCA).

2.2 Air Quality Monitoring and Estimation System

Air pollution is a serious health risk caused by rapid urbanisation, population growth, and industrialization. The human health concern is one of the most serious short-term consequences of air pollution, particularly in urban areas. The long-term effects on the environment are global warming and greenhouse gases. Air pollution-related issues have raised public awareness of air quality around the world. Air quality monitoring and estimation tools are required to take preventive measures, such as reducing the impact of predicted pollution peaks on the surrounding population and ecosystem.

In Malaysia, ambient air quality is measured using the Air Pollution Index (API). The API is designed to describe and report air quality in a simple manner instead of using pollutants concentration directly. The API has a category for each hazardous level based on human health ranging from good to hazardous.

Different types of air pollution estimation models have been developed so far, for instance, ANN-based models, linear models, and hybrid models. These models can detect primary pollutants such as PM10, SO_2 or secondary pollutants such as O_3. Other categorisations that can distinguish air prediction models include the emission and climatic variables used by several models, the region of prediction (urban, industrial, coastal, or desert), and the predicted duration (short or long term).

It is critical to accurately predict air quality in order to provide actions and controlling strategies to mitigate the negative effects of pollutants on human health. Therefore, several studies on air quality estimation models using artificial neural networks have been conducted, such as by (Antanasijević et al., 2013, Putra et al., 2018). Unlike other modelling techniques, ANN makes no assumptions about the distribution of data. In addition, ANN model can be trained to accurately generalise and model highly nonlinear relationships when given a new data set (Putra et al., 2018).

There are so many variables have been used on API estimation models. These variables can be divided into meteorological such as wind speed, wind direction, temperature, relative humidity, ambient temperature, rainfall, solar radiation, general condition, pressure, rain, and pollution variables for instance NO, NO_2, CO, PM10, O_3, SO_2, NO_x, CO_x, etc (Ao et al., 2019).

Several techniques that can be used to assess the significance of the relationship between potential inputs and outputs have been review by (Cabaneros et al., 2019). These techniques are classified as model-free techniques and model-based techniques. Model-free techniques such as analytical methods based on correlation and mutual information (MI) and ad-hoc methods based on available data and domain knowledge. For model-based methods, the techniques included stepwise methods such as constructive or pruning methods, sensitivity analysis, global optimization methods such as genetic algorithms, and ad-hoc methods such as trial-and-error methods (Cabaneros et al., 2019).

The methods that are used to account for input independence can also be divided into two categories: dimensionality reduction methods and filtering methods. The dimensionality reduction technique can be achieved by using principal component analysis with input rotation (Azid et al., 2014b), or by clustering the input or output space, as with clustering analysis and self-organizing maps (Bhavaraju et al., 2010).

Many studies have been conducted based on the possibility of significantly reducing the number of input variables required while maintaining the predictive power of the model. Some methods such as principal component analysis (PCA), kernel principal component analysis (KPCA), and multiway principal component analysis (MPCA) have been proven to be effective tools to reduce the model dimension without affecting the model estimation efficiency.

2.3 Air Quality Monitoring and Estimation System based on Artificial Neural Network

Table 2.1: Advantages and disadvantages of ANN

Advantages	Disadvantage
Complex process	Time-consuming
Accurate classification	Data volume
Direct model	
Black-box model	

The benefit of using the ANN model is that it can handle difficulty with very many parameters. Moreover, ANN can successfully give accurate values and classify objects, despite the chaotic distribution of the objects. Also, It goes straight from the data to the model, with no recording, bingeing, simplification or questionable interpretation in between. Therefore, it necessitates less human involvement than traditional analysis. However, the training process of ANN consumes a significant amount of time, especially in supervised training. ANN has limitations with training data which will affect the effectiveness of the results since it depends on the training data and the training methods of the network.

2.4 Application of Principal Component Analysis Methods in Air Quality Estimation Model

The high dimensional dataset can be downsized by using principal component analysis (PCA), which is considered to be one of the most popular and useful statistical approaches to discover the possible structure of a set of variables. This approach is used to characterize the variance of a large set of interrelated variables by converting them into a different smaller set of uncorrelated variables namely principal components (PCs). PCs are non-correlated and orthogonal and have linear combinations of the original variables (Azid et al., 2014b).

PCA can explain the most significant variables which can indicate the source of the pollutants. This is because the less important variables are removed from the data set with a minimum loss of original data in the analysis process (Abdi and Williams, 2010, Azid et al., 2014b).

In summary, the aims of principal component analysis are to extract the most important information from the data table, to compress the size of the data set by keeping only the important information, to simplify the description of the data set, and to analyse the structure of the observations and the variables (Abdi and Williams, 2010).

Kernel principal component analysis (KPCA) is non-linear dimension reduction method based on the classical principal component analysis. It transform non-linear data in the original space into linear data in the mapping space by introducing kernel function (Yuan et al., 2018). The kernel function is non-linear function where it maps the data of lower dimensions into higher dimension space. Moreover, the kernel function's basis is to implement the inner product transformation of a vector. Therefore, it will not increase the complexity of input data. After kernel transformation, the kernel principal components will be extracted into new space (Xiao et al., 2017).

Multiway PCA is the normal PCA that has been extended to deal with an additional dimension. First, it is performed by unfolded the N-dimensional data array into a two-dimensional matrix then apply PCA on the resultant two-dimensional matrix (Swiercz and Mroczkowska, 2020). The unfolded process involves reshaping a folded N-dimensional array into a two-dimensional array. The number of unfolding possibilities doubles for an N-dimensional data matrix. However, not all of these unfolding make sense in monitoring applications, so only the most relevant ones should be used. Since the PCA studies the relationship between these variables, the uncorrelated variables must be avoided so the complexity and cost of computation can be reduced (Burgas et al., 2018).

CHAPTER 3

METHODOLOGY

3.1 Introduction

The objectives of this study are to determine a suitable dimensional reduction method based on the application of multiway principal component analysis (MPCA), and to develop an estimation model of air pollution index based on an artificial neural network (ANN). This research will be carried out in as shown in Figure 3.1

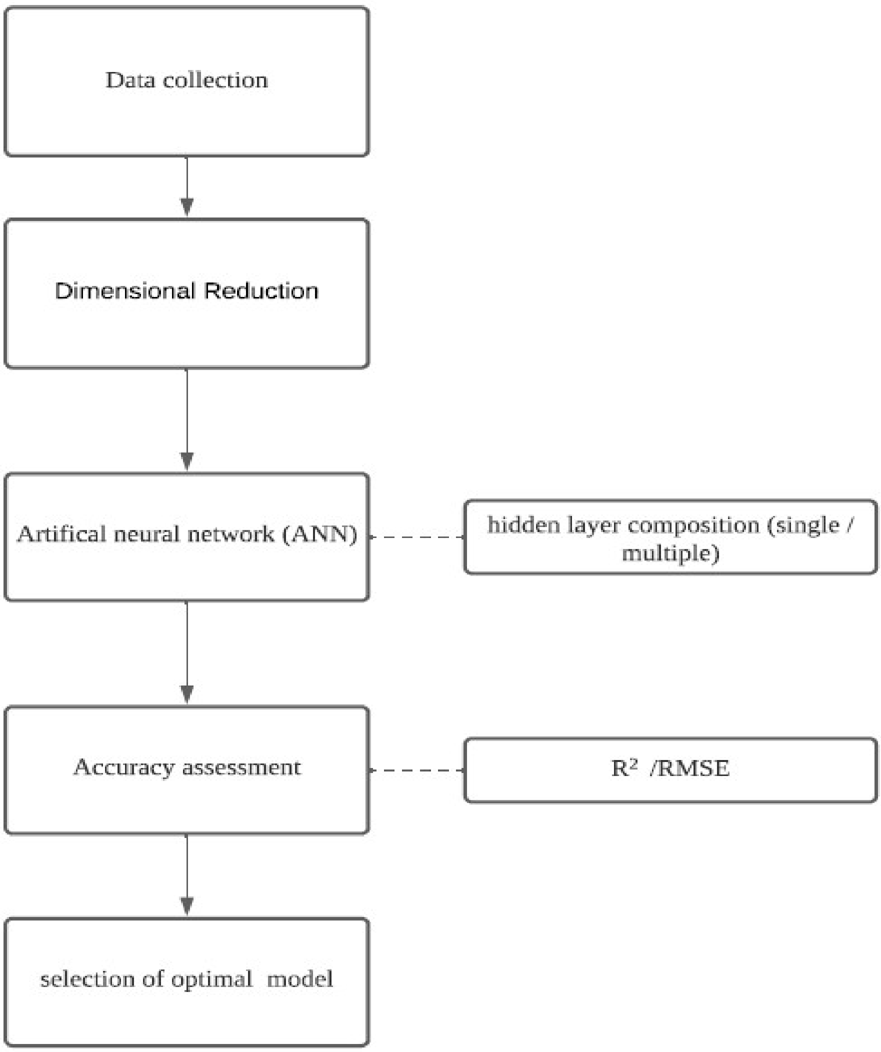

Figure 3.1: Flow diagram of the research project

3.2 Data Collection

The air quality data were gathered from the Air Quality Division of the DOE, Malaysia. The data was collected from the year 2015 to 2020 in several monitoring stations located in Penang. Six different parameters have been monitored since they are most effected parameters for air pollution index (API) in Malaysia, namely nitrogen oxides, carbon monoxide, sulfur dioxide, wind speed, air temperature, and relative humidity.

Those six parameters with 1370 sampling data will be used to predict ozone (O3) and particulate matter 10 (PM10) models. After that, the O3 and PM10 model outputs will be used as input for the air pollution index (API) estimation model. All air quality parameters that will be involved in this study are listed in Table 3.1.

Table 3.1: Input Parameters

O_3 estimation model PM10 estimation model		**API estimation model**	
Input parameters	Output parameters	Input parameters	Output parameters
Nitrogen Oxides, NOx	Ozone, O_3	Ozone, O_3	Air Pollution Index (API)
Carbon Monoxide, CO	PM10	PM10	
Sulphur dioxide, SO_2			
Wind speed			
Air Temperature			
Relative humidity			

3.3 Dimensional Reduction

The multiway principal component analysis (MPCA) for dimensional reduction have been used in this study. The implementation has been treated as pre-training procedures before the input data will be used for ANN training.

Even though principal component analysis (PCA) is a well-known dimension reduction approach for removing unnecessary or redundant input variables while maintaining the data structure, the MPCA also statistically and algorithmically consistent with PCA and has the same goals and benefits. The relation between MPCA and PCA is that MPCA is equivalent to the ordinary PCA performance on a large two-dimensional matrix formed by unfolding the three-way array in one of three possible ways.

Therefore, in this study, the unfolding of the input data and the PCA method was performed in MATLAB 2019a. This platform was utilized to find the most significant principal components (PCs, through the determination of eigenvalues and eigenvectors. Eigenvectors and eigenvalues are the linear algebra concepts that are used to determine principal components. The eigenvectors of the covariance matrix are the directions of the axes with the most variance (most information), which are called PCs. Eigenvalues are simply the coefficients attached to eigenvectors that represent the amount of variance carried by each PC. The result will show eigenvectors in order of their eigenvalues, highest to lowest. where a principle component that represents more or equal to 99.9% of data was taken.

3.4 Artificial Neural Networks

In general, ANNs are massive parallel computational models that mimic human brain function. ANN consists of a large number of simple processors connected by weighted connections. The processing nodes can be referred to as 'neurons.' The output of each node depends only on the information available locally at the node, either stored internally or obtained through the weighted connections. Each unit receives inputs from several other nodes and transmits its output to another node (Dongare et al., 2012).

3.4.1 Architecture of ANN

The process of information with neural networks represent by trillions of neurons (nerve cells) formed the networks, which will create electrical pulses to exchange between cells called action potentials. Computer algorithms that mimic these structures of biological are properly called artificial neural networks. Generally, the three basic types of networks based on network architecture are single-layer feedforward, multi-layer feedforward, and recurrent network (Birdi et al., 2013).

3.4.1(a) Single Layer Feedforward

These are the earliest and most basic artificial neural networks that have been discovered. It consists of a single layer feedforward network that has a single layer of artificial neurons, and it processes input signals in a forward directional manner. The net weighted sum of the node's inputs is subjected to a step function. The input pattern is assigned to one of two classes which depend on whether the node's output is 0 or 1 (Haykin and Network, 2004).

3.4.1(b) Multi-layer Feedforward

The multi-layer feedforward network is a development of the single-layer network, which is used to solve far more complex and complicated problems that cannot be solved using a single layer approach or that take longer to solve. It is made up of the three most important parts of any network: an input layer of neurons, one or more hidden neurons layers, and an output layer of neurons, as illustrated in Figure 3.2. The network's power comes from the hidden layer, which allows it to extract additional features from the input (Haykin and Network, 2004).

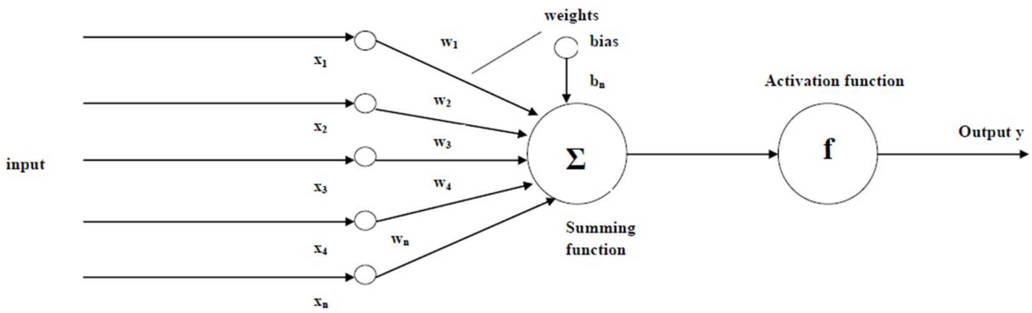

Figure 3.2: Typical multi-layer feedforward architecture

3.4.1(c) Recurrent Network

Recurrent networks are similar to feedforward neural networks, but they vary in that they have at least one feedback loop. The outputs of some nodes or the network are propagated back to the input layers or nodes through the feedback connections, allowing for repeated computations. A recurrent network's output will be affected by an input presented at time t for future time steps greater than time *t*. As a result, recurrent networks must be operated over time (Haykin and Network, 2004).

3.4.2 Transfer Functions

The most commonly used transfer functions are the linear function, the threshold function, the log-sigmoid function, and the tanh-sigmoid function. Different transfer functions can be used based on the applications (Zekić-Sušac et al., 2013).

The linear transfer function:

$$y = c.s \tag{3.1}$$

Where ,

c = constant.

s = the resulting weighted sum.

y = the output.

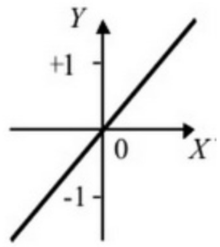

Figure 3.3: Linear function (Dorofki et al., 2012)

The threshold function:

It only computes 1 and 0. This type of activation-function is known as the McCulloch-Pitts model that represents a neural network's ability to be "all-or-nothing".

$$y = \begin{cases} +1, if\ s \geq 0 \\ -1, otherwise \end{cases} \tag{3.2}$$

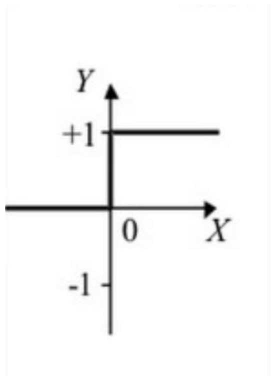

Figure 3.4: Threshold function (Dorofki et al., 2012)

The logistic sigmoid transfer function:

When using the sigmoid activation function, the artificial neuron can resemble a natural neuron more closely. The advantage of this function is that it creates a very soft transition. It gives outputs in the range [0, 1]:

$$y = \frac{1}{1 + e^{-s}} \tag{3.3}$$

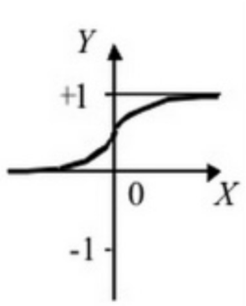

Figure 3.5: Sigmoid function (Dorofki et al., 2012)

The tanh-sigmoid transfer function:

The tangents hyperbolic activation-function is similar to the Sigmoid, but it can produce negative results. It gives outputs that lie in the range [−1, 1]:

$$y = \frac{e^{s} - e^{-s}}{e^{s} + e^{-s}} \tag{3.4}$$

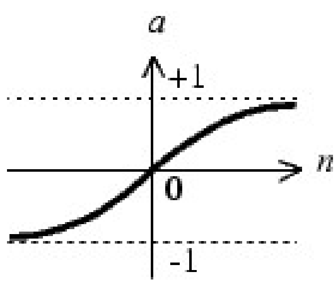

Figure 3.6: Tanh-sigmoid function (Dorofki et al., 2012)

3.4.3 Training of ANN

An ANN must be designed in such a way that it generates desired outputs in response to a set of training inputs. There are two types of training algorithms: supervised training and unsupervised training (Haykin and Network, 2004).

- Supervised training

Under the supervised training style, a distinction is made between the actual outputs of an ANN and the expected output of an ANN. Therefore, it attempts that desired solutions are known for the training data sets. This consists of reducing error over time by changing the weights input until network accuracy is acceptable.

- Unsupervised training

Unsupervised training does not necessitate the correct desired data set. In reality, the data's fundamentals, or the connections between the data's patterns, are identified and organised into categories. This is particularly useful when there are no known solutions.

3.4.3(a) Learning Rate

A learning rate, also known as learning constant, appears in most learning functions. This term is usually positive and between 0 and 1. If a low rate is chosen, more time is spent training an ANN, but the outcomes are more stable. As the learning rate increase, the training takes less time with faster learning rates, but the results are less accurate.

Other factors that influence how long it takes to train a network include network complexity (number of hidden neurons and layers), data size, architecture, and learning rule type.

3.4.4 Training Methods

3.4.4(a) Feedforward, Back-propagation Network

The feedforward backpropagation architecture was introduced in the early of 1970's by several independent sources. Currently, backpropagation architecture is becoming increasingly popular, useful, and easy to learn, even for complex models like multi-layered networks. The ability of ANN to deal with nonlinear solutions to infinite problems is its greatest strength. A typical back-propagation network has an input layer, an output layer, and at least one hidden layer.

One of the most widely used ANN algorithms is the BP algorithm. (Rojas, 2013) stated that the BP algorithm could be broken down into four steps. After the weights are chosen at random, the backpropagation algorithm is used to calculate the required corrections. The algorithm can be divided into the four steps below:

- Computation of feedforward
- Output layer backpropagation .
- The hidden layer propagation.
- Weight updates.

These four steps will continue until the function error value may become small enough.

3.4.4(b) Levenberg-Marquardt (LM) Algorithm

It is one type of back-propagation (BP) algorithm. The Levenberg-Marquardt (LM) algorithm is a gradient-based algorithm that takes into account the error term's first and second derivatives It adjusts the weight in the direction of the steepest descent, where the performance function is rapidly decreasing. The primary advantage of the LM algorithm over the conventional BP algorithm is that it provides a faster (second order) convergence rate while maintaining relative stability. On other hand, the main disadvantage is that it may have a potential convergence to a local minimum and overfitting which means finding the optimal weight matrix is not guaranteed by this algorithm (Putra et al., 2018).

3.4.4(c) Bayesian Regularization

The Bayesian approach to control model complexity which will solve the issue of overfitting of the data. Overfitting means that the network error is driven to a minimal value for the training samples but becomes big when new input is presented. This shows that the network has memorised the training samples but is unable to generalise to provide suitable answers for input parameter combinations that have not been seen before. In Bayesian regularization, the network's weights are treated as random variables. After that, statistical techniques can be used to estimate distribution parameters such as variances. In this approach, the optimal regularisation parameters may be determined automatically. In addition, no test set or validation set is required with the Bayesian technique, and all available training data can be used for both model fitting and model comparison. The main benefit of this technique is that it provides an implicit assessment of how many network parameters (weights) the network is efficiently using. This indicates whether the network has received a sufficient number of training samples and automatically selects the best network size (Hirschen and Schäfer, 2006).

3.5 ANN Estimation Model Development

Artificial Neural Networks (ANNs) have been employed in a range of areas and have proven to be effective. ANNs have distinct advantages, such as non-parametric nature, ease of adaption to various data sources, and flexible decision boundary capabilities but they have inherent limitations. These limitations result from some factors, which may affect the accuracy of the ANN’s outputs. These factors can be divided into two main categories which are internal factors and external factors. Internal factors include the choice of an appropriate network structure, initial weights, number of iterations, transfer function, and learning rate, while external factors include the properties of the input data set (multisensor, multispectral, etc.) and the study's scale. For a successful ANN application, it's critical to understand these factors and choose proper parameter values (Kavzoglu, 1999).

The ideal neural network architecture for a given task can only be established through experimentation. The quality of a neural network's output is highly influenced by the network size used. In general, network size has an impact on network complexity, learning time, and most crucially the network's generalisation capabilities. Generalisation refers to the neural network's ability to interpolate and extrapolate data that it has never seen before. An ANN's power is determined by how well it generalises from training data. Moreover, generalisation capabilities of a neural network can be affected by the training data set size and characteristics together with the number of iterations.

Small networks may become trapped in a local error minimum and fail to learn from the training data, whereas large networks take a long time to learn the characteristics of the data. As the number of nodes in the hidden layer(s) increases, the network may learn more complex data by detecting decision boundaries more accurately. However, when reducing the network's generalisation capabilities, the time necessary to train the network will increase. On other hand, large networks require more training samples than small networks to attain good generalisation performance, due to the approximately linear relationship between the number of samples required for the training process and the number of hidden units.

The number of input nodes is dependent on the size of the hidden layer since each node represents different characteristics of the pattern used. The input layer can be increased in size by simply adding new data sources as extra neurons, although this significantly increases the computation time (N) by the order of N^2. To put it another way, if the size of the input data is doubled, the time it takes to train the network will be four times longer than it was before. Therefore, new data sets should be added only if they significantly improved outputs. Another parameter to consider is the number of outputs, which is determined at the start. A larger number of outputs usually makes the problem more complex which means more computation time (N). Therefore, it is very important to choose the optimum number of outputs to avoid unnecessary training (Kavzoglu, 1999).

Some problems necessitate more than one hidden layer to adequately train a network, whereas others only necessitate a single hidden layer. The use of extra hidden layers that are not needed can make the network excessively specific and need more training time. If there are too few hidden units, the network will not perform well, whereas if there are too many hidden units, the network will memorize the patterns in the training set (and so become overspecific to the training data) and so perform poorly for patterns that are not included in the training data. Optimization of generalisation performance is obtained by trading the training error against network complexity. It is important to note that a smaller network is more likely to generalise effectively since it extracts the most important and relevant features from the training data (Thomas et al., 2015).

The neural network toolbox in MATLAB by The MathWorks divides the data set into three categories namely the training set, validation set, and the testing set. The training set is used to train the network by which the ANN will learn from the input data and update its weight accordingly. The validation set is presented to the network without outputs during the training process then the error in validation data will be monitor throughout the training process. When the network starts to overfitting the data, the validation data error will increase until it reaches maximum error then the training process stops to avoid further overfitting the data, and the network is returned at the minimum number of validation errors. On other hand, the test data set will not be used during the training process but is used to test the performance of the trained network. The neural network will not be able to offer sufficient performance if the test set reaches the minimal value of MSE at a considerably different iteration than the validation set (Beale et al., 2010).

3.6 Evaluation of ANN Estimation Model Performance

The performance of the ANN model is typically measured using a quantitative error metric. However, ANN models should not be determined purely on their predictive error, but also on their ability to obtain underlying dynamics between predictors and target variables. As such, two aspects of model validity will be used when assessing the performance ANN model which are:

a) Root mean square errors (RMSE).

b) Coefficient of determination (R^2).

The model with the highest value of R^2 and the lowest value of RMSE can be declared the best linear model.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 Data Pre-processing – Dimensional Reduction

The performance of a neural network can be optimized by dimensional reduction of the database, which can be achieved by feature extraction methods such as MPCA. This will ensure that all the essential and relevant information regarding the air quality monitoring database can be used effectively.

This dimension reduction can be done using the MPCA technique which will eliminate unwanted or redundant input variables while keeping the structure of the data intact. The outputs from the MPCA input processing will be represented by principal components, eigenvalues, and eigenvectors.

Table 4.1: Explained for Principle components (PCs)

Principle component	**Eigenvalues**
PC1	94.2629
PC2	5.2678
PC3	0.4676
PC4	0.0017
PC5	5.6526×10^{-7}
PC6	3.4586×10^{-8}

From Table 4.1 PC1 has highest eigenvalue which is equal to 94.2629. Therefore, the most significant principal component is PC1. In addition, we observe that the inclusion of up to 4 PCs will contribute to the explained percentage of approximately