STATISTICAL MODELLING FOR FORECASTING PM10 CONCENTRATIONS IN PENINSULAR MALAYSIA

NG KAR YONG

UNIVERSITI SAINS MALAYSIA

2017

STATISTICAL MODELLING FOR FORECASTING PM10 CONCENTRATIONS IN PENINSULAR MALAYSIA

by

NG KAR YONG

Thesis submitted in fulfilment of the requirements
for the degree of
Master of Science

November 2017

ACKNOWLEDGEMENT

First and foremost, I would like to express my sincere thanks to my supervisor, Dr. Norhashidah Awang for her guidance, advice and support throughout my postgraduate study in Universiti Sains Malaysia (USM). I would also like to acknowledge Malaysia Department of Environment (DOE) for providing me the data required in this research. I thank USM and the Ministry of Higher Education for their financial assistance via USM Fellowship Scheme and MyMaster. Special thanks also extended to my parents for their continuous support and care along my research study. Last but not least, I would like to extend my appreciation to my dearest friends who have always supported me during these times.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACF	Autocorrelation function
ADF	Augmented Dickey-Fuller
AIC	Akaike Information Criterion
ANN	Artificial neural network
API	Air Pollution Index
AQG	Air Quality Guidelines
AR	Autoregressive
ARCH	Autoregressive conditional heteroscedasticity
ARIMA	Autoregressive integrated moving average
ARIMAX	Autoregressive integrated moving average with exogenous variables
ARMA	Autoregressive moving average
ASMA	Alam Sekitar Malaysia Sdn Bhd
BIC	Bayesian Information Criterion / Schwartz Bayesian Criterion
CO	Carbon monoxide
CWT	Continuous wavelet transform
DOE	Department of Environment
DWT	Discrete wavelet transform
FAR	False alarm rate
FFT	Fast Fourier Transform
FIR	Finite impulse response
GARCH	Generalized autoregressive conditional heteroscedasticity
GEV	Generalised extreme value

iid	Independent and identically distributed
KPSS	Kwiatkowski-Phillips-Schmidt-Shin
LB	Ljung-Box
LM	Lagrange multiplier
MA	Moving average
MAAQG	Malaysia Ambient Air Quality Guidelines
MAPE	Mean absolute percentage error
MLR	Multiple linear regression
MRA	Multiresolution analysis
NFDA	Nonparametric functional data analysis
NO	Nitrogen monoxide
NO_2	Nitrogen dioxide
NO_x	Nitrogen oxides
O_3	Ozone
OLS	Ordinary least squares
PACF	Partial autocorrelation function
PAH	Polycyclic aromatic hydrocarbon
PC	Principal component
PCA	Principal component analysis
PM	Particulate matter
$PM_{0.1}$	Particulate matter with aerodynamic diameter less than 0.1 microns (μm)
PM_{10}	Particulate matter with aerodynamic diameter less than 10 μm
$PM_{2.5}$	Particulate matter with aerodynamic diameter less than 2.5 μm
POD	Probability of detection

PP	Phillips-Perron
QR	Quantile regression
RMSE	Root mean squared error
RTSE	Regression with time series error
SARFIMA	Seasonal autoregressive fractionally integrated moving average
SARIMA	Seasonal autoregressive integrated moving average
SO_2	Sulphur dioxide
SSA	Singular Spectrum Analysis
STS	Structural Time Series
SVR	Support vector regression
TAEF	Time-delay Added Evolutionary Forecasting
VARMA	Vector autoregressive moving average
VOC	Volatile organic compound
WHO	World Health Organization
WT	Wavelet transform

PEMODELAN STATISTIK BAGI PERAMALAN KEPEKATAN PM10 DI SEMENANJUNG MALAYSIA

ABSTRAK

Fenomena jerebu yang berlaku secara berulangan di Malaysia telah mendorong keperluan untuk mengawal pencemaran PM_{10} (habuk halus berdiameter kurang daripada 10μm) dengan berkesan. Hal ini memerlukan peramalan yang tepat dan pemahaman tentang hubungan antara pencemar PM_{10} dengan faktor-faktor lain. Oleh itu, kajian ini bermatlamat untuk meramal purata harian kepekatan PM_{10} di Semenanjung Malaysia dengan menggunakan kaedah pemodelan univariat iaitu pemodelan siri masa dan pemodelan regresi. Dalam analisis siri masa, suatu masalah yang biasa ialah penganggaran yang terlalu rendah terhadap kepuncakan. Memandangkan siri kepekatan PM_{10} yang berubah-ubah secara drastik, kajian ini mengemukakan penggunaan model siri masa berasaskan jelmaan wavelet bagi meningkatkan ketepatan ramalan, yakni penggunaan jelmaan wavelet diskret (DWT) sebelum pemodelan siri masa menggunakan model Box-Jenkins autoregresi terkamir purata bergerak (ARIMA) dan autoregresi heteroskedasticiti bersyarat teritlak (GARCH). Dengan menggunakan DWT, siri PM_{10} yang berubah-ubah diuraikan kepada beberapa sub siri dengan varians yang lebih kecil, dan dengan yang demikian, telah meningkatkan ketepatan ramalan agregat secara ketara terutamanya pada tempoh jerebu berbanding kaedah pemodelan siri masa tanpa DWT. Model regresi dengan ralat siri masa (RTSE) telah dikemukakan untuk menyelesaikan masalah reja yang berautokorelasi daripada model regresi linear berganda (MLR) yang piawai. Model ini berjaya mengambil kira autokorelasi tersebut. Hasil analisis telah menunjukkan bahawa perubahan kelembapan hari ini merupakan faktor utama yang

berkait dengan perubahan kepekatan PM_{10} pada keesokan hari. Di samping itu, perubahan-perubahan dalam nitrogen dioksida (NO_2), ozon (O_3), suhu dan arah angin hari sebelumnya juga merupakan faktor-faktor penting yang berkait dengan perubahan kepekatan PM_{10}. Selain itu, prestasi peramalan kepekatan PM_{10} keesokan hari bagi model-model RTSE adalah memuaskan kerana mereka adalah setanding dengan peramalan univariat yang menggunakan kaedah DWT. Sebagai tambahan, regresi kuantil (QR) telah digunakan untuk menyiasat secara lebih lanjut perubahan kesan pembolehubah-pembolehubah meteorologi dan gas pada kepekatan PM_{10} merentas kuantil. Daripada dapatan analisis, perubahan suhu yang lalu berkait rapat dengan perubahan yang besar dalam kepekatan PM_{10}. Pada kuantil-kuantil rendah dan tengah, QR dapat mengesan kepentingan perubahan kelajuan angin yang lalu selain perubahan kepekatan PM_{10} yang lalu dan pembolehubah-pembolehubah lain yang penting dalam RTSE. Secara ringkas, model siri masa berasaskan jelmaan wavelet berkesan untuk meningkatkan ketepatan ramalan kepekatan PM_{10} di Semenanjung Malaysia. Model RTSE dapat mengambil kira autokorelasi dan menunjukkan ketepatan ramalan yang baik. Tambahan pula, model QR dapat memberi pemahaman yang holistik tentang hubungan antara pembolehubah-pembolehubah peramal dengan PM_{10} pada pelbagai taburan kuantil.

STATISTICAL MODELLING FOR FORECASTING PM10 CONCENTRATIONS IN PENINSULAR MALAYSIA

ABSTRACT

Recurring haze phenomena in Malaysia have prompted the need for effective control of PM_{10} (particulate matter with diameter less than 10 μm) pollution. This demands accurate forecasts as well as understanding on relationship between PM_{10} and other factors. Hence, this research aims to forecast the daily average PM_{10} concentrations in Peninsular Malaysia by using univariate modelling, i.e. time series modelling and regression modelling. In time series analysis, a typical problem in forecasting is the underestimation of the peaks. Since the series of PM_{10} concentrations change rapidly, this research proposed the use of wavelet-based time series model to improve the forecast accuracy, i.e. the application of discrete wavelet transform (DWT) before the time series modelling by the Box-Jenkins autoregressive integrated moving average (ARIMA) and generalized autoregressive conditional heteroscedasticity (GARCH) models. By employing DWT, the volatile PM_{10} series were decomposed into several subsidiary series with smaller variations, and consequently, helped in improving the total forecast accuracy substantially especially during haze periods when compared to the time series modelling without DWT. Regression with time series error (RTSE) model was proposed to overcome the problem of autocorrelated residuals from the standard multiple linear regression (MLR) model. It successfully accounted for the autocorrelation. The analysis revealed that the difference of lagged humidity was the major factor related to the next-day difference of PM_{10} concentration. Furthermore, changes in lagged nitrogen dioxide (NO_2), ozone (O_3), temperature and wind direction were also the important

factors associated with the change in PM_{10} concentration. Besides, the one-day-ahead forecast performances of the RTSE models were acceptable since they were comparable to the univariate forecasting by method with DWT. In addition, quantile regression (QR) was conducted to further investigate the changes in effects of meteorological and gaseous variables on PM_{10} concentrations across quantiles. From the findings, the change in lagged temperature was associated with the large difference of PM_{10} concentrations. At low and middle quantiles, QR additionally detected the significance of change in lagged wind speed besides the change in lagged PM_{10} concentrations and those variables which are significant in the RTSE. In a nutshell, wavelet-based time series model is useful to enhance the forecast accuracy of PM_{10} concentrations in Peninsular Malaysia. RTSE model is able to account for the autocorrelation and has good forecast accuracy. Furthermore, QR model is able to provide a holistic insight about the relationship between predictor variables and PM_{10} at different quantile distributions.

CHAPTER 1

INTRODUCTION

1.1 Introduction to Statistical Modelling

A model is a mathematical equation which describes a process or system. A statistical model is a mathematical model which is built based on some sample data. Hence, statistical modelling is a process of building a statistical model, the purpose of which is to make general statements or to draw inferences about the general behaviours of a population or a random process which generates the sample data. A random process usually contains deterministic and random components. While a deterministic part is mostly the expected value which represents the general feature of a process, random component is the uncontrollable feature such as different characteristics among individuals. Therefore, a model is termed as statistical model by taking into account the random component through probability distribution. As such, a statistical model involves three main aspects, namely an equation, assumptions on the random component and the way to combine the deterministic and the random parts (model specification). The specification of models can be broadly classified as linear and nonlinear models. By formulating a statistical model, one can use it to predict future values and also to understand the process under study (Krzanowski, 1998).

The scope of this thesis focuses on statistical modelling of time series data. Time series data is a series of ordered observations which is collected at fixed time

interval. Basically, the analysis of time series can be divided into frequency-based and time-based analysis. In frequency-based analysis, there are methods such as spectral analysis and wavelet analysis. For analysis based on time domain, the classical modelling technique is the Box-Jenkins autoregressive integrated moving average (ARIMA) model. Additional advanced models embody seasonal models, long memory models, heteroscedasticity models and so on. Among models which involve other driving factors are dynamic regression models and autoregressive integrated moving average with exogenous variables (ARIMAX) model.

1.2 Background of Study

This thesis focuses on statistical modelling of the concentrations of particulate matter (PM) in Peninsular Malaysia. PM is a mixture of solid particles and liquid droplets suspended in the air (Schwartz *et al*., 1996). PM is generally categorized by its size (measured in aerodynamic diameter) into three groups, namely coarse PM (PM_{10}), fine PM ($PM_{2.5}$) and ultrafine PM ($PM_{0.1}$). PM_{10} is defined as the particles with aerodynamic diameter of less than 10 microns (μm), and it comprises $PM_{2.5}$ and $PM_{0.1}$ (Anderson *et al*., 2012). PM_{10} is of special concern by authorities and researchers because it is within the inhalation range and causes serious harm to human health (Schwartz *et al*., 1996).

Besides the size, PM is also characterized by its various chemical compositions, and thus by its various sources. The wide range of compositions includes soil particles, elemental carbon, polycyclic aromatic hydrocarbons (PAH), nitrate, sulphate and volatile organic compounds (VOCs). These varying components are due to different emission sources. Primary PM is emanated directly from primary sources, whereas secondary PM is a product of chemical reaction of its precursors

such as VOCs, nitrogen oxides (NO_x) and sulphur dioxide (SO_2) in the atmosphere. The direct sources are of both natural and anthropogenic. Natural emissions include volcanoes and forest fires, while anthropogenic activities include agriculture, industries, power generation, traffic and combustion. Basically, coarse PM comes from mechanical activities such as agriculture, mining, construction and grinding. On the other hand, fine PM is mostly originated from combustion activities. Secondary PM usually appears in the form of nitrate and sulphate components, resulting respectively from oxidation of NO_x and SO_2. These precursor gases are primarily emanated from transportation, power plants and industries (Bhattacharjee *et al*., 1999; WHO, 2013; DOE, 2015). Furthermore, the sulphur component was also found to be attributed to forest fires and biomass burning during haze periods (Afroz *et al*., 2003).

Apart from gaseous pollutants, the concentration of PM_{10} is affected by meteorological factors such as monsoons, rainfall, temperature, humidity, pressure, wind speed and direction. Normally, high concentrations are recorded during the dry season (southwest monsoon) from June to September (Liew *et al*., 2011). Conversely, the concentration of PM_{10} is lower during the rainy period from October to November as a result of dilution by rainfall (Liew *et al*., 2009). In addition, the formation of PM_{10} is more conducive under high temperature and calm weather (Shaharuddin *et al*., 2008).

Due to its detrimental effects on human health, PM_{10} has been formulated as one of the key pollutants, along with ozone (O_3), carbon monoxide (CO), nitrogen dioxide (NO_2) and SO_2, in the Air Pollution Index (API) system by Malaysia Department of Environment (DOE). The API is an indication of the status of air quality from good to emergency, described in terms of a range of values instead of the actual concentrations of the pollutants, for easy understanding to public. The API

is calculated by taking the highest sub-index among the five air pollutants. The API categories also provide corresponding health effects, health advices and actions to be taken (DOE, 2000). The categorization of API and its air quality status is illustrated in Table 1.1.

Table 1.1: Categorization of API

API	Air Quality Status
0-50	Good
51-100	Moderate
101-200	Unhealthy
201-300	Very Unhealthy
>300	Hazardous
>500	Emergency

Based on the Malaysia Ambient Air Quality Guidelines (MAAQG), the standard concentration limits for PM_{10} are $150\,\mu g/m^3$ at an averaging time of 24 hours and $50\,\mu g/m^3$ at an averaging time of 12 months. Breaching the limit of $150\,\mu g/m^3$ will lead to the unhealthy level in API system on that particular day (DOE, 2000). The World Health Organization (WHO) AQG implements stricter standard values for PM_{10} where $50\,\mu g/m^3$ for daily mean and $20\,\mu g/m^3$ for annual mean (WHO, 2013).

PM_{10} pollution causes damages to human's health, environment and economy. Coarse PM can reach the larger and upper respiratory tracts, whereas fine PM can deposit deeper in the smaller airways and alveoli to give rise to respiratory and cardiopulmonary diseases (Peng *et al.*, 2008). Studies have shown that PM_{10} pollution leads to both acute and chronic health effects. The short term exposure to PM_{10} induces the respiratory problems such as asthma and attenuated lung function as well as increased hospital admissions (WHO, 2013), while long-term exposure to PM, especially $PM_{2.5}$, is associated with mortality from cardiopulmonary diseases. The respiratory problems happen because of the oxidative stress and inflammation in

lungs, while the cardiovascular problems occur due to the plaque conglomerating in arteries (Anderson *et al*., 2012). The consequences may be more adverse for the vulnerable populations including the elders, children and those with existing lung and heart illnesses (WHO, 2013). Moreover, the exposure of pregnant women to PM_{10} emitted from motor vehicles increases the possibility of premature defects, especially attributed to musculoskeletal and chromosomal anomalies (Vinceti *et al*., 2016).

In addition to the health complications, PM_{10} also provokes environmental issues. The obvious impact is the downgrade of visibility during haze events in which PM_{10} was found to be the predominant pollutant (Shaharuddin *et al*., 2008). This is due to the light scattering or absorption effect by the particles, and thus reducing the amount of sunlight reaching to the earth. Notably, $PM_{2.5}$ worsens the visibility more than PM_{10} does. Furthermore, PM also impairs visibility during humid days by forming fog (Bhattacharjee *et al*., 1999). Another implication of PM is acid deposition. Acid deposition on materials causes damage and soiling of man-made sculptures, paints and buildings. PM acts as an agent for the accumulation of acidic gases such as SO_2 and NO_2. These acids accumulate on the material surfaces and accelerate the corrosion of the materials. On the other hand, acid deposition in water and soil destructs the aquatic system and vegetation. The excess sulphates and nitrates are poison to the aquatic lives and crops (Bhattacharjee *et al*., 1999).

The health and environmental effects above-mentioned induce a great amount of economy losses following the health and environmental treatments. Furthermore, the reduced productivity due to haze-related sickness and restricted activity days also accounts for the economy losses (Afroz *et al*., 2003).

In order to control the emission of PM_{10}, cooperative actions by authorities, industries and public are needed. The Malaysian government has developed the management policies covering different aspects such as emission legislation, prevention, enforcement and education. For example, incentives are given to industries which install the pollution control equipment (DOE, 2010). Penalties or jail sentence are to be judged for open burning activities (Prohibition, 2017). Recently in 2013, a New MAAQG was set up to include $PM_{2.5}$ as one of the principal pollutants in the awareness of its severe health implications (ITA, 2016). The New MAAQG adopts three stages of interim targets in 2015, 2018 and 2020 in order to complete the implementation of emissions reduction. For PM_{10}, the targets of limits in 2020 are 100μg/m^3 and 40μg/m^3 for the averaging time of 24 hours and one year, respectively (Air quality standards, 2017). Furthermore, regional efforts were also done to mutually maintain a cleaner environment in region. For instance, all Southeast Asia countries, including Malaysia have ratified the ASEAN Agreement on Transboundary Haze Pollution to combat the transboundary haze together (ITA, 2016).

1.3 Motivations

PM_{10} forecasting plays an important role in giving advance health warnings to public. Increased mortality and morbidity rate during hazy days (Othman *et al*., 2014; Sahani *et al*., 2014) provide evidences to the danger of PM_{10}. Hence, by early warning, appropriate planning such as limiting outdoor activities can avoid exposure to PM_{10}, and thus reducing the risk of getting ill or death. At the same time, medical expenses can be reduced.

In the long term, forecasting is significant as a part of PM_{10} pollution control program. With the reliable forecasts, authorities can take actions occasionally on those forecasted high PM_{10} days, thereby, mitigating the high cost of continuous emission control (Air Quality Research, 2001).

Furthermore, PM_{10} is known to be associated with other pollutants such as SO_2 and NO_2. However, the compositions of PM_{10} at different areas may differ. Hence, understanding of these relationships aids in reduction of related pollutants and consequently, effectively reducing PM_{10} levels as well.

For the sake of public and decision making, the forecasts of short-term trend should be provided. Therefore, the main goals of this research are to improve the forecast accuracy of PM_{10} concentrations as well as to forecast the PM_{10} concentrations based on predictor variables in Peninsular Malaysia.

1.4 Problem Statements

PM_{10} forecasting is essential, especially to forecast the high PM_{10} concentrations. Nonetheless, a common problem in forecasting is the underestimation of the abnormally high PM_{10} values (Liu, 2009). Furthermore, PM_{10} series typically have peaks and troughs with different scales. Thus, no one global model is adequately fit (Joo and Kim, 2015). Albeit numerous methods have been used, there was no single conclusion on whether linear or nonlinear model was better. Hence, in this research, wavelet-based time series model is proposed to overcome this problem.

In the context of modelling involving predictor variables, multiple linear regression (MLR) is a simple and common solution. However, PM_{10} time series data may result in serial correlation (autocorrelation) among residuals which contradicts

to the independence assumption of MLR. In Malaysia, literatures such as those by Liew *et al*. (2011) and Dominick *et al*. (2012) did not consider the autocorrelation nature of time series, while some others such as Ul-Saufie *et al*. (2013) added lagged PM_{10} as the predictor variables. Therefore, regression with time series error (RTSE) model is proposed in this research. To our knowledge, wavelet-based time series model and RTSE approaches have not been used in pollution forecasting in Malaysia.

Moreover, PM_{10} data normally exhibit heterogeneous conditional distributions over quantiles. Most researchers would be interested in the multifarious effects of predictor variables exerted on the PM_{10} concentrations, especially at the upper tail distribution. In Malaysia, Ul-Saufie *et al*. (2012) has used quantile regression (QR) to predict the future PM_{10} concentrations, but there were no detailed analysis and discussion on the relationship between PM_{10} and the predictor variables. Hence, QR is implemented to gain a better apprehension on different impacts of predictor variables on PM_{10}.

1.5 Objectives of Research

There are two main objectives in this research. The first objective is to improve the forecast accuracy of daily average PM_{10} concentrations, and the second is to examine the relationship between PM_{10} and other predictor variables (meteorological and gaseous parameters). The details of the objectives are as follows:

i. To model and forecast the daily average of PM_{10} concentrations at selected monitoring stations in Peninsular Malaysia using wavelet-based time series model. This technique transforms the PM_{10} concentrations series into several sub-series and then the autoregressive integrated moving average with

generalized autoregressive conditional heteroscedasticity (ARIMA-GARCH) method is applied.

ii. To compare the performance of the above approach with the conventional method (series without wavelet transform (WT)), and hence, verify whether the WT improves the accuracy of forecasting.

iii. To explore the association between PM_{10} and predictor variables and to forecast the one-day-ahead daily average PM_{10} concentrations at selected monitoring stations in Peninsular Malaysia using the RTSE approach.

iv. To gain more insights on the heterogeneous effects of predictor variables on PM_{10} across different quantiles of PM_{10} distribution by applying QR.

1.6 Methods and Data

As illustrated in Figure 1.1, the thesis is divided into two main parts to fulfil the two main objectives of the research. The first part is time series modelling and the second part is regression modelling constitutes of RTSE and QR modelling.

The data used in this study are the 24-hour daily averages computed from the hourly data obtained from Malaysia DOE. The dataset includes the concentrations of PM_{10}, NO_2, nitrogen monoxide (NO), SO_2, CO and O_3 as well as temperature, humidity, wind speed and wind direction in the years of 2013 and 2014. These data were collected continuously at a network of 52 air monitoring stations throughout Malaysia, managed by Alam Sekitar Malaysia Sdn Bhd (ASMA) which is endorsed by DOE.

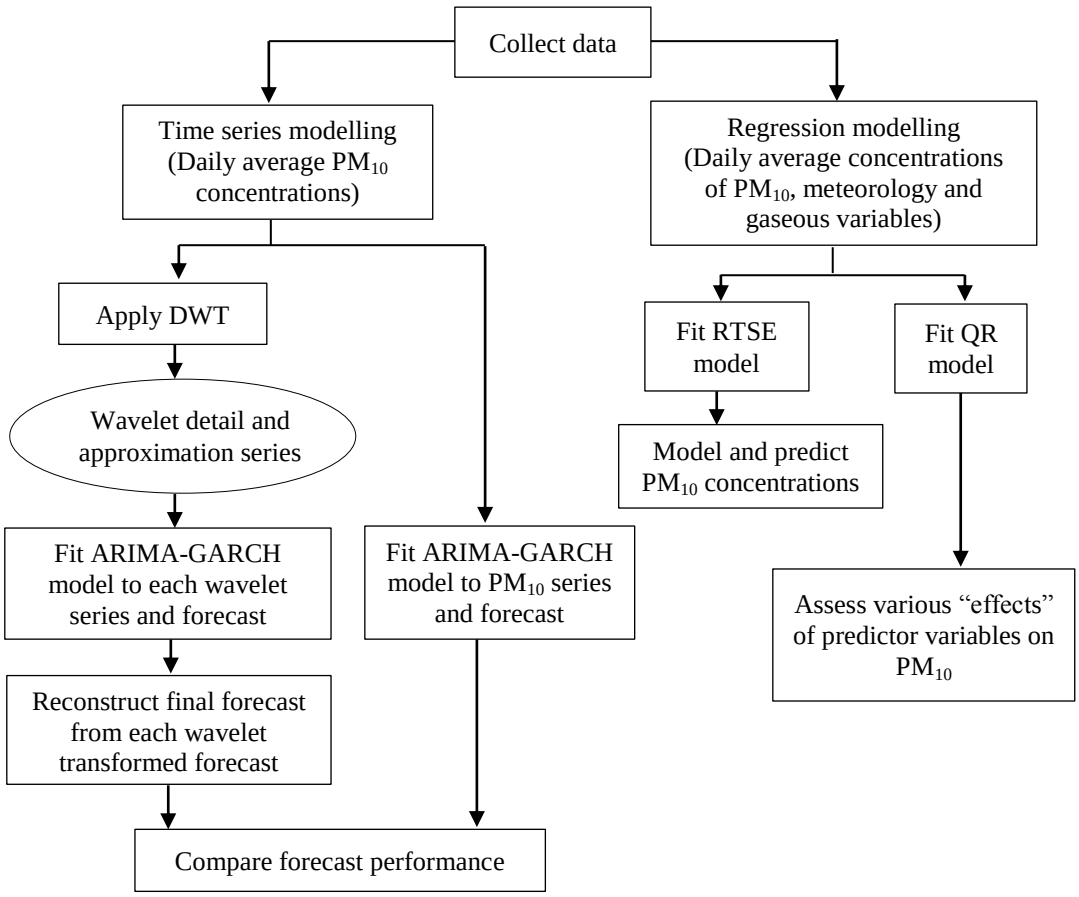

Figure 1.1: Flow chart of research procedures

In the first part of univariate time series forecasting, the PM_{10} concentrations from 2013 to 2014 were involved. Five monitoring stations situated at different locations and backgrounds, namely Seberang Jaya station (suburban), Nilai station (industrial), Klang station (urban), Petaling Jaya station (industrial) and Batu Muda station (urban) were selected.

For the second part of regression, all the above-mentioned variables of air pollutants and meteorological variables in the year 2014 were considered. The selected monitoring stations are Seberang Jaya station (suburban), Petaling Jaya station (industrial), Sungai Petani station (suburban), Seremban station (urban) and Batu Muda station (urban). The air monitoring stations involved in this research are summarized in Table 1.2 below.

Table 1.2: Summary of air monitoring stations involved

Air monitoring station	Station code	Background type
Seberang Jaya	CA09	Suburban
Nilai	CA10	Industrial
Klang	CA11	Urban
Petaling Jaya	CA16	Industrial
Batu Muda	CA58	Urban
Sungai Petani	CA17	Suburban
Seremban	CA47	Urban

In order to improve the forecast performance of time series method which is the ARIMA-GARCH model in forecasting PM_{10} concentrations, WT was applied to PM_{10} series before developing the ARIMA-GARCH model. PM_{10} series usually have large variances due to some extremely high concentrations in the series. This often makes the modelling process difficult and at the same time, reducing the forecast accuracy. Hence, the general concept here is to reduce the variability of input series to the time series model by decomposing the original time series into several wavelet detail and approximation sub-series using discrete wavelet transform (DWT). Instead of original time series with large variance, the wavelet transformed sub-series with lower variability are inputted into the time series model. Considering the reduction in variability, it is expected that the total forecast accuracy will be boosted.

With regards to the second objective of investigating the relationship between PM_{10} and its predictor variables, RTSE model was used mainly to account for the autocorrelation of time series. In addition, it was also used in forecasting PM_{10} concentrations based on the predictor variables. Moreover, QR was adopted to analyse the relationship between PM_{10} and predictor variables for various PM_{10} quantile distributions.

The DWT and regression modelling processes were implemented by using free R software, while time series modelling was conducted using Eviews 8 software.

1.7 Organisation of Thesis

In the next chapter, the related literatures are reviewed where the discussion is divided into time series modelling and modelling involving predictor variables. Chapters 3 and 4 explain the methods of time series modelling (DWT, ARIMA and GARCH) and regression modelling (MLR, RTSE and QR), respectively. Subsequently, the forecasting results of time series and regression are discussed in Chapters 5 and 6, respectively. Finally, the conclusion remarks are presented in Chapter 7.

CHAPTER 2

LITERATURE REVIEW

In the last few decades, a great variety of statistical modelling techniques have been developed and used in air pollution forecasting. As this research is primarily partitioned into time series forecasting and forecasting based on predictor variables, Section 2.1 concentrates on literatures which used solely the time series data of one variable, whereas Section 2.2 discusses literatures which involved predictor variables. Finally, Section 2.3 summarizes the findings from the review.

2.1 Time Series Modelling Techniques

The univariate time series models are advantageous and have often been used when there is lack of information of other variables because it depends only on the historical data of the time series itself. Amongst all, Box-Jenkins ARIMA methodology is the mainstream. Usually, the forecast performance of any improved methods is compared to the ARIMA model. Other improved methods evolved from it in an attempt to improve the forecast accuracy. For example, there were ARIMA models associated with GARCH model and ARIMA with wavelet decomposition method.

Kumar *et al*. (2004) used an ARIMA approach to forecast one-step-ahead daily maximum O_3 concentrations in Brunei Darussalam. The time series from July 1998 to March 1999 was found to be suitably fit by ARIMA (1,0,1) model. The

short-term forecast performance was satisfactory with fractional bias, normalized mean squared error and mean absolute percentage error of 0.025, 0.02 and 13.14%, respectively.

In India, Chelani and Devotta (2006) made comparison between the nonlinear local polynomial approximation method and the linear autoregressive (AR) model in predicting the daily average PM_{10} concentrations from 1999 to 2002 in Mumbai. The local polynomial approximation was performed based on reconstructed phase space. The errors of prediction indicated that the nonlinear method produced better prediction than the AR model.

Ismail (2011) used seasonal ARIMA (SARIMA) model to forecast the monthly O_3 concentrations in Kedah, Malaysia. The forecast results from SARIMA $(1,0,1)(2,1,2)^{12}$ model showed that there was significantly increasing trend of O_3 level in the long term. Furthermore, the model could also help in decision making such as planning strategies.

Quintela-del-Rio and Francisco-Fernandez (2011) employed nonparametric functional data analysis (NFDA) for prediction and study of extreme value of O_3 concentrations by using the data in Switzerland (mean monthly data) and United Kingdom (daily maxima), respectively. For prediction, NFDA yielded more accurate forecasts in terms of mean squared error when compared to the conventional ARIMA method. The pro of the nonparametric approach is its flexibility in the sense that it does not assume normal distribution or linear relation, while its negligible con is the slightly heavier computational load. On the other side, NFDA also performed much better than the parametric generalised extreme value (GEV) fit in estimating the return levels of extreme O_3 concentrations.

Kumar (2015) proposed a combined technique of Singular Spectrum Analysis (SSA) with ARIMA model (SSA-ARIMA) to forecast the daily maximum O_3 concentrations. The data from six European AIRBASE stations with different backgrounds were considered. The SSA was used to model the deterministic part, while ARIMA model was used to model the stochastic component. The proposed method was compared to the more popular Fast Fourier Transform (FFT) method integrated with ARIMA model. The findings proved that SSA-ARIMA provided more accurate and reliable (narrower 95% confidence interval) forecasts for one day ahead as well as multiple days ahead. Moreover, the better performance was more prominent for multiple-day-ahead forecasts.

Stoimenova (2016) also applied ARIMA method to forecast the daily average PM_{10} concentrations in Pernik, Bulgaria. ARIMA (1,0,5) as the optimum model was used to forecast the seven-day-ahead PM_{10} concentrations and it produced quite satisfactory result.

Considering the long memory and heteroscedasticity (changing of variance across time) of the time series, Reisen *et al*. (2014) applied SARFIMA (seasonal autoregressive fractionally integrated moving average)-GARCH model to the daily average PM_{10} series in Cariacica, Brazil. The comparison of forecasts between models with GARCH and without GARCH showed that the former outperformed the latter. In the meantime, the wider forecast interval of the GARCH model was able to cover more proportions of data of high volatility.

Instead of fitting the nonlinear model to account for the heteroscedasticity of the series, some authors employed WT to break down the series into more auxiliary series with smaller variation so that simpler forecasting models can be fitted to the sub-series. Joo and Kim (2015) demonstrated the superiority of wavelet filtering in

different scenarios. Eight simulated series and eight real data series were fitted by two methods, namely ARIMA with wavelet filtering (proposed method) and ARIMA without wavelet filtering (customary method). In most cases, the proposed method yielded lower mean absolute percentage errors than the customary method. This study also showed that wavelet filtering was particularly useful for the seasonal time series with great amount of noises.

Zhang *et al*. (2017) compared the performance of ARMA/ARIMA to wavelet-based ARMA/ARIMA models in forecasting the daily PM_{10} concentrations at four monitoring stations in Taiyuan, China. The results showed that the wavelet-based ARMA/ARIMA models reduced the forecast errors for all monitoring stations. However, only short-term forecasting was suitable.

Besides the conventional ARIMA models, many other different approaches were used. For instances, Fernandez de Castro *et al*. (2005) introduced a functional technique for forecasting the SO_2 concentrations near a power plant in As Pontes, Spain. The functional models embodied a kernel-based approach and linear autoregressive Hilbertian model. They emphasized on an estimation method called "historical matrix" which classified the data according to shapes but not levels. Furthermore, bootstrap technique was employed to compute the forecast confidence intervals. The predictive performances of the proposed models were generally better when compared to artificial neural network (ANN) and semiparametric methods.

In Malaysia, Md Yusof *et al*. (2010) fit the hourly PM_{10} series from 2000 to 2004 in Seberang Perai by applying Weibull and lognormal distributions. The results demonstrated that lognormal distribution fit better to the data from 2000 to 2002, while Weibull distribution fit better to the data in 2003 and 2004 which showed

higher return periods, suggesting that Weibull distribution was more appropriate to model the high concentrations of PM_{10}.

Lawson *et al*. (2011) predicted the hourly mean NO_2 and NO_x concentrations in Dublin, Ireland by using a Structural Time Series (STS) model. In this methodology, different components of the time series such as trend, seasonal components and disturbances can be modelled separately whereby the model was expressed in state-space form and was solved by Kalman filter algorithm. The prediction results demonstrated that the STS method outperformed ANN and support vector regression (SVR). The findings testified that STS model possesses several advantages. The model provides clear and neat description of different components of time series and does not require stationarity. Additionally, it can easily handle the missing values and outliers.

De Mattos Neto *et al*. (2014) presented Time-delay Added Evolutionary Forecasting (TAEF) architecture to forecast the daily mean $PM_{2.5}$ and PM_{10} concentrations in Helsinki. This method consists of two steps, namely the optimization of ANN parameters and phase adjustment based on the differences between the predicted and observed concentrations. The results showed that the proposed method was superior over the other methods shown in previous literatures. Furthermore, the authors attributed the great improvement in forecast accuracy to the phase adjustment which considers the random walk characteristic of PM time series and suggested this as a feasible way to improve the forecasting of intelligent systems.

On the other hand, Chelani (2015) advanced the use of nearest neighbour method in forecasting the one-step-ahead PM_{10} concentrations in Nagpur, India. This method does not require information on predictor variables and distributional assumptions. Five function approximation techniques, namely mean, median,

persistence model, linear combination and kernel regression were used and compared. Kernel regression was the best. In addition, the combination forecasting where more weightage was put on the model with smaller error was suggested. It outperformed the individual kernel regression model and was able to capture the nonlinear pattern of time series.

2.2 Modelling Techniques Involving Predictor Variables

In the pollution study, researchers often seeks to model and to understand the relationship between pollutant and meteorological and emissions factors, as well as to use the model to forecast its future values. While there are various techniques, MLR and ANN models have stood out from the others and have been extensively used.

A study by Van der Wal and Janssen (2000) compared Kalman filtering to MLR in predicting PM_{10} concentrations in Netherlands based on wind direction, temperature and duration of precipitation. The Kalman filtering performed better than MLR as it considers the explaining variables as time-varying. However, Kalman filtering as a linear model failed to capture the nonlinear behaviour of PM_{10} during peak seasons.

In order to accommodate the advantage of both nonlinear and linear models, Diaz-Robles *et al*. (2008) proposed a hybrid model of ANN and ARIMA to forecast the daily maximum PM_{10} moving average from 2000 to 2006 in Temuco, Chile. This hybrid method was compared to MLR as well as ANN and ARIMAX models separately. The findings showed that the hybrid method effectively reproduced 100% of alerts and 80% of pre-emergency events and it outperformed the three individual models.

Liu (2009) suggested RTSE model with principal component analysis (PCA) to forecast the daily average PM_{10} concentrations in Ta-Liao, Taiwan. RTSE was employed instead of MLR because it allows autocorrelation among observations. The models with inclusion of principal component (PC) triggers of neighbouring PM_{10} concentrations were proved to be advantageous in forecasting the PM_{10} concentrations more than 150 $\mu g/m^3$ as it improved the forecast accuracy in terms of increased probability of detection and critical success index and reduced false alarm rate, as compared to the models without PC triggers of neighbouring concentrations.

Liew *et al*. (2011) employed MLR to analyse the possible factors affecting the daily PM_{10} concentrations during summer monsoon (May to August) in Malaysia. The PM_{10} concentrations, local meteorological variables, synoptic weather patterns and hotspot counts from 2003 to 2006 at six monitoring stations in Klang Valley were included. Based on the findings, surface air temperature, relative humidity and wind speed were the consequential factors. Furthermore, the synoptic weather and foreign hotspot also explained the variability of PM_{10} concentrations besides the location of cyclone formation. The authors concluded that MLR was practical for determining the factors influencing the PM_{10} concentration.

Study by Vlachogianni *et al*. (2011) also supported the usefulness of MLR. They constructed stepwise MLR models to predict the NO_x and PM_{10} concentrations at Athens and Helsinki. The prediction results were compared to those from ANN. Although ANN produced slightly better results, MLR was preferred for practical use because it excels in the way of easy interpretation and implementation. However, MLR did not work well in situation with abrupt changes.

Dominick *et al*. (2012) used correlation analysis and MLR to study the association of meteorological parameters with daily average PM_{10} and NO_2 at three

monitoring stations (Shah Alam, Johor Bahru and Kuching) in Malaysia. The results showed that PM_{10} correlated positively to temperature but correlated negatively with humidity and wind speed. On the other hand, NO_2 correlated positively with temperature and humidity but correlated negatively with wind speed.

Ul-Saufie *et al*. (2013) predicted the future (the next-day, next-two-day and next-three-day) PM_{10} concentrations in Negeri Sembilan, Malaysia by using the MLR and ANN models incorporating with PCA. The predictor variables included previous days of PM_{10}, SO_2, NO_2, CO, wind speed, temperature and relative humidity. The results confirmed the advantage of PCA in improving the prediction accuracy for both methods. Particularly, PCA-ANN was the best for the next-day prediction, while PCA-MLR yielded the best results for the next-two-day and next-three-day predictions.

Targeting to understand the behaviour of PM_{10} anomalies, Shaadan *et al*. (2015) used an integration of robust project pursuit and robust Mahalanobis distance methods to convert the PM_{10} concentrations at three monitoring sites in Klang Valley into functional data. The functional data analyses revealed that PM_{10} anomalies were linked to monsoon, days in a week and wind speed. More frequent of high concentrations were observed during Southwest and Northeast monsoons as well as weekdays and they were related positively to the wind speed.

Catalano *et al*. (2016) presented an ensemble of ANN and ARIMAX approaches to improve the accuracy in forecasting the hourly NO_2 peaks at Marylebone road in London. The variables involved were total traffic volume, hourly mean wind speed, wind direction and temperature. The findings showed that the ensemble performed the best in predicting extreme values. Individually, ARIMAX

model was better than ANN in detecting peaks but ANN could better represent the behaviour of NO_2 based on the wind variables.

Besides the popular ANN models, another type of nonlinear model which is the GARCH model was often coupled with other linear model to overcome the problem of changing variances of air pollutants over time. For example, Wu and Kuo (2012) used the GARCH together with vector autoregressive moving average (VARMA) model to investigate the correlation among air pollutants and their variation patterns at eight monitoring stations in Central District of Taiwan. This hybrid method has successfully provided informative findings.

WT is also useful in the analysis of air pollutants. Shaharuddin *et al*. (2008) used non-decimated WT to study the relationship between PM_{10} concentration in Petaling Jaya and other meteorological parameters (temperature, rainfall and wind speed). It was found that there were relations between PM_{10} and the parameters at low frequencies but not at the high frequencies.

Zainuddin and Ong (2011) demonstrated wavelet neural networks in function approximation and pollutant prediction. Three different wavelet functions, namely Mexican Hat, Gaussian and Morlet wavelet functions were used as the activation function in the hidden layers of neural networks. They emphasized the importance of the choice of wavelet functions in order to improve the efficiency of the models. The results suggested that Gaussian wavelet neural networks surpassed the other two types of wavelet neural networks.

Siwek and Osowski (2012) forecasted the daily average PM_{10} concentrations in Warsaw by conducting wavelet decomposition and applying ensemble of neural networks. They deduced that wavelet decomposition has played a significant role in enhancing the forecast accuracy.

While MLR is still a very prevalent technique, QR has started to attract more interests recently in pollution study. QR was first proposed by Koenker and Bassett (1978) to provide a robust modelling technique for non-Gaussian-distributed data (Koenker and Bassett, 1978). In addition, its feature of producing a range of coefficient estimations at various distributions of pollutant enabling researchers to assess different covariates effects across quantiles on pollutant makes it favourable to researchers.

In 2004, Baur *et al*. used QR approach to analyse the nonlinear relationship between O_3 and meteorology in Athens. This method revealed the heterogeneous covariate effects at different O_3 levels which were masked by the MLR model using the ordinary least squares (OLS) estimation method. They also showed that QR significantly elevated the global goodness of fit when compared to the MLR.

Similar research was conducted by Sousa *et al*. (2009) on O_3 data in Portugal. Results of QR indicated the different influences of predictors at different points of O_3 distribution. The results also showed some significant variables at both end tails of distribution which MLR showed to be insignificant. Furthermore, the prediction performance of the next-day O_3 concentrations by using QR in training phase excelled the MLR model.

Ul-Saufie *et al*. (2012) predicted future PM_{10} concentrations in Seberang Perai, Malaysia by employing QR. The QR model outperformed MLR for the next-day, next-two-day and next-three-day predictions.

Munir (2016) studied the correlation between PM_{10} and meteorological parameters in Makkah. By using QR, the author confirmed the changing impacts of covariates at different PM_{10} quantile distributions and concluded that QR was able to offer new perspective to inspect air quality data.

2.3 Summary

This chapter has reviewed a number of different methods for time series forecasting and modelling of the relationship between pollutant and the predictor variables. Through the review, it can be concluded that there were two primary aspects of interests, namely the accuracy in forecasting the extreme values and the modelling of the nonlinear pattern of the pollutant.

In forecasting, there were no certain linear or nonlinear models which are good for all. While most of the studies have proved that nonlinear models such as ANN and GARCH hybrid models outperformed the linear models such as ARIMA and MLR models, there were still some papers such as those by Vlachogianni *et al*. (2011), Ul-Saufie *et al*. (2013) and Catalano *et al*. (2016) whose results supported the superiority of linear models. Since there have been a great deal of studies validating the virtue of WT (Shaharuddin *et al*., 2008; Siwek and Osowski, 2012; Joo and Kim, 2015; Zhang *et al*., 2017), this research aims to develop a wavelet-based time series model to enhance the forecast performance of the PM_{10} abnormalities in Peninsular Malaysia. By using this method, the nonlinearity such as heteroscedasticity of the time series is considered. Furthermore, since accuracy of long-term forecasting cannot be guaranteed (Zhang *et al*., 2017), only short-term forecasting are performed. Besides, even though Shaharuddin *et al*. (2008) have used WT in pollution study in Malaysia, our research focuses on its applicability in forecasting but not on the analysis of relationship as in their study.

On the other hand, this research also intends to construct a model to represent the relationship between PM_{10} and predictor variables. Although MLR has been proven to be useful in most literatures, this approach is considered inappropriate to model the time series as the independence assumption is violated. In Malaysia, most

of the studies such as Liew *et al*. (2011), Dominick *et al*. (2012) and Ul-Saufie *et al*. (2013) either neglected the autocorrelation or included the lagged response variable as the predictors. Therefore, a modified MLR model, namely the RTSE model is employed in this research to consider the serial correlation of time series. Moreover, QR is also employed for supplementary purpose to the analysis of RTSE which models the average distribution of PM_{10} concentration only. To our knowledge, there was only a study in Malaysia by Ul-Saufie *et al*. (2012) which used QR for prediction of PM_{10} concentrations but they did not conduct deep analysis on the relationship of PM_{10} with other variables.