UNIVERSITI TEKNOLOGI MARA

ASSESSMENT OF OZONE (O_3) FUNCTIONAL DATA IN MIRI, SARAWAK

NUR MA FADHILAH MAT SELEEI

BACHELOR OF SCIENCE (Hons.)
MANAGEMENT MATHEMATICS

JUNE 2020

Universiti Teknologi MARA

Assessment of Ozone (O_3) Functional Data Analysis in Miri, Sarawak

Nur Ma Fadhilah Binti Mat Seleei

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

June 2020

SUPERVISOR'S APPROVAL

ASSESSMENT OF OZONE (O_3) FUNCTIONAL DATA IN MIRI, SARAWAK

By

NUR MA FADHILAH BINTI MAT SELEEI
2017957591

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

Approved by:

.....
BALKIAH BINTI MOKTAR
Supervisor

JUNE 25, 2020

STUDENT'S DECLARATION

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

.....

NUR MA FADHILAH BT MAT SELEEI

2017957591

JUNE 25, 2020

ACKNOWLEDGEMENT

Alhamdulillah, praise and thank to Allah because of His Almighty and His utmost blessings, I was able to finish this research within the time duration given. Firstly, my special thanks go to my supervisor, Madam Balkiah Bt Moktar who has been giving encouragement, support, and supervising me to finish this project. It is an honor for me especially when I am having a problem regarding this research.

Special appreciation also goes to my beloved parents who have been giving their full physical and mental support to me. They also keep motivate me to complete this report successfully.

Last but not least, I would like to give my gratitude to my dearest friend for helping, giving their opinions and suggestions, about this project. And, I would like to thanks all persons who involve in this project either directly or indirectly in which I can finish this study.

ABSTRACT

Developing countries cannot be spared the presence of poor air quality due to rapid technological change. On the other hand, it also causes new and growing health problems. However, over the decades, this has been a persistent issue as many people are still ignorant of it. According to Shaadan, Deni, & Jemain (2012), a geographical area, a high level of industrial and commercial activity, a high-density population, heavy-duty vehicles, and others are responsible for poor air quality. Therefore, this study is conducted to assess the functional curve of ozone, O_3 behavior at a monitoring station in Miri, Sarawak, Malaysia. Functional Data Analysis (FDA) is used in this study because it can produce a model that can be continuously represented as a smooth dynamic. This also enables precise estimation of the parameters to be used in the analysis process, an efficient way of reducing data noise by curve smoothing and useful for data with various sampling schedules. In this study, the results of the analysis revealed implicit information on the existence of two significantly different O_3 behaviors between 2014 and 2015. The results showed that anomalies were detected in the first half of 2014, while anomalies were not detected in 2015. This showed that the diurnal behavior was influenced by the various dominant emission sources and other methodological conditions that existed in those years.

Keywords: ozone, functional curve, Functional Data Analysis (FDA), curve smoothing, anomalies

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

O_3	Ozone
PM10	Particulate Matter with the Size of Fewer than 10 Microns
SO_2	Sulfur Dioxide
CO	Carbon Monoxide
NO_2	Nitrogen Dioxide
API	Air Pollutant Index
FDA	Functional Data Analysis
FPCA	Functional Principal Component Analysis
FLM	Functional Linear Modeling
CVD	Cardiovascular Disease
SIE	Sea Ice Extent
PWT	Penn World Tables
OLS	Ordinary Least Squares
BIC	Bayesian Information Criteria
MCD	Minimum Covariance Determinant
PC	Principal Components

CHAPTER 1
INTRODUCTION

1.1 Background of the Study

Economic, social, technological, policies and environmental evolution are transforming a country into an urban and industrial area. Ruth and Franklin (2014) explained that the key elements of urban live ability are the ability of the city to meet its residents' needs and the condition of the city's environment in maintaining its residents' livelihoods. As a result, many people moved from rural to urban areas to get well-paid jobs, better health care, and good education due to the rapidly expanding industries. Besides, well-paid jobs are usually correlated with rises in society's total income and living standards, while technology changes allow the growth of more factories thus increasing more job opportunities. Due to the rapid growth of urban areas, human health, and the environment are exposed to the risk of atmospheric pollution (Azmi, Latif, Ismail, Liew, & Jemain, 2010). This has become ongoing problems over the decades because many people still unaware of it.

Pollutions are divided into air pollution, water pollution, noise pollution, and so on. Haze is one of the air pollution caused by fine particles that appear in the atmosphere that cannot be seen directly by the naked eye. During the dry period, air quality worsened due to motor vehicle smokes, industrial emissions, and soot from open burning activities that significantly contributed to the hazy situation. Therefore, by monitoring air quality, it helps detect changes in the status of ambient air quality that can affect the public and the environment. As stated by The Department of Environment (DOE), about 51 monitoring stations specifically located in residential, traffic, and industrial areas are set up to detect the change in the air quality. The data is collected continuously (24 hours a day) during the monitoring period.

In this study, ground-level ozone (O_3) is focused on rather than particulate matter with the size of fewer than 10 microns (PM10), sulfur dioxide (SO_2), carbon monoxide (CO), and nitrogen dioxide (NO_2), which are other main pollutants that included in the Air Quality Standard. In the presence of sunlight and pollutants, ozone was found in the atmosphere by photochemical reactions, according to the WHO Air Quality Guidelines 2005. Nonetheless, if ozone concentrations increase above the guideline level, various effects on human health such as lung malfunction, lung inflammation, respiratory disease in children, and deaths.

1.2 Problem Statement

The rapid changes in technology caused massive urbanization and advancement in the industrial process. Therefore, enduring poor air quality cannot be denied to a developing country, which is also causing new and increasing health problems. A geographical position, high industrial and commercial operation, high-density populations, heavy vehicle traffic, and others are responsible for poor air quality (Shaadan, Deni, & Jemain, 2012). The Air Pollutant Index (API) is an air quality measure used to monitor the relationship between air pollution and human health (Rahman, Lee, Suhartono, & Latif, 2016). The calculation of API that consists of PM10, O_3, CO, SO_2, and NO_2. According to DOE, the API reading status is categorized into 0-50 shows good air quality, 51-100 moderate, 101-200 unhealthy, 201-300 very unhealthy, and over 300 hazardous.

This study will focus on the monitoring station at Miri, Sarawak. Miri is one of the urban and industrial areas in Sarawak. According to the official website of Sawarak Natural Resources and Environment Board (2019), Sarawak's air quality status has generally been good to medium throughout the past year, except for a few unhealthy days reported during the dry season, particularly in July, August, and September. However, the hotspots detected in the nearby region Sumatra and Kalimantan, Indonesia was significantly high particularly, in August and September each year. Recently, Malaysia experienced severe haze due to forest

burning in Kalimantan as well as Sarawak. Business Insider (2019) reported that API reading for Sri Aman, Sarawak hit 367, which considered hazardous. Moreover, a four-month-old and a 59-year-old man reportedly died due to the worsening haze crisis in Indonesia. As a result, hundreds of schools in Malaysia have been ordered to closed, 298 schools in Sarawak, 138 schools in Selangor, and 65 schools in Port Dickson (The Star, 2019). Besides that, 25 schools in Putrajaya also ordered to closed (Bernama, 2019). Thus, this study is performed to analyze the patterns of the O_3 by using Functional Data Analysis (FDA).

1.3 Objective of the Study

The objective of this study is to assess the functional curve of O_3 behavior at a monitoring station in Miri, Sarawak, Malaysia.

1.4 Scope of the Study

This study focuses on studying the patterns of O_3 behavior in Sarawak. The secondary data was collected from the DOE specifically at a monitoring station in Miri, Sarawak where the chosen station is Sekolah Menengah Dato Permaisuri, Miri, Sarawak. The daily hourly data for this study will be obtained from 2014 to 2015.

1.5 Significance of the Study

This research is important as it can show the status of air quality and its effects on human health, ecology, and others so that precautions can be taken when the worst air quality is discovered. In the previous haze episode, the Malaysian government took the necessary steps to amend the laws on open burning, industrial activities, and cloud seeding (Rani, Azid, Khalit, Juahir, & Samsudin, 2018). Individuals also can contribute to reducing poor air quality by stopping open burning, forest burning, starting the carpooling system, and so on. As a result, people can continue their daily routine without having to worry about their health. Finally, for

university, this research may become guidance for future researchers to produce better researches in improving the quality of air.

CHAPTER 2
LITERATURE REVIEW

2.1 Functional Data Analysis (FDA)

The change of data collection technology in recent decades has enabled more complex observations to be sampled over time, space, and other continuum measures (Ullah & Finch, 2013). A large proportion of these data are classified as functional data and converted into the FDA due to sudden changes in statistical (Morris, 2015). In general, the FDA's main idea is to analyze discrete observations that are rising from time series as functional data, so that the conclusion can be drawn based on the collection of functional data using statistical concepts. Morris (2015) clarified that the FDA is unique due to the need to integrate data within and across roles or to call replication and regularization. Replication means merging information between functions to produce correlations and draw a conclusion on populations that have been used as the sample, whereas regularization includes power within a function, capitalizing on the predicted underlying structural relationships to achieve efficiency and understanding between a function. According to Ullah and Finch (2013), FDA methods produce a model that can be described by smooth dynamics continuously. It also enables reliable parameter estimates to be used in the analysis process, an efficient way of reducing data noise by curve smoothing, and useful to data with inconsistent sampling schedules. By using functional data analysis, there is no doubt about the relation between repeated measurements because the FDA views the entire curve as a single entity that allows model creation which can equally treat complex structures both within functions and between functions (Morris, 2015; Jamaludin & Zulkifli, 2017).

2.2 Features of Functional Data Analysis (FDA)

The important features of the FDA identified by Ullah and Finch (2013) are smoothing methods, use of Functional Principal Component Analysis (FPCA), method of clustering modification, functional linear modeling (FLM) approach, and type of forecasting. Morris (2014) explained that basis functions are the parts of the FDA and used to determine the mechanism of regularization. The basis functions of the FDA are splines, Fourier series, wavelets, and principal components in which each has certain characteristics that suit functions. However, this study will be focused on smoothing techniques and data reduction.

2.2.1 Smoothing Techniques

Smoothing is the FDA's fundamental phase. The aim is to transform discrete raw data into a smoothly varying function. Ullah and Finch (2013) argued that the smoothing method used to detect longitudinal data errors is significant. However, the option in choosing smoothing techniques depends on the underlying behavior of the data being analyzed. Fourier series is used because the data are cyclical or periodic (Levitin, Nuzzo, Vines, & Ramsay, 2007).

2.2.2 Data Reduction

The most popular multivariate analysis technique is FPCA. This method reduces the number of interrelated variables although it still retains as much as possible of the overall variance. The reduction is accomplished by moving the data into a new set of uncorrelated and ordered variables or principal components to ensure that the first few retain most of the differences present in all the original variables (Ullah & Finch, 2013).

2.3 Application of Functional Data Analysis (FDA)

Ullah and Finch (2013) have listed different types of applications of the FDA in various science fields. For example, child-size evolution analysis, climate variability, Chinese handwriting, acidification processes, satellite-based land-use forecasting, medical research, behavioral science, term-structured yield curves, spectrometry data, and age-fallen injury incidence relationship over time.

In the medicine area, by using a functional data analysis method, Rowland et al. (2019) investigate the relationship between size-fractioned lipoprotein particles and cardiovascular disease (CVD). In the study, the hazard of CVD is concluded that the abundance measurements differ because abundance values are highly correlated with particles of similar diameters. FDA is used to mimic the ion mobility data and traditional risk factors with CVD at the same time to help identify size areas by mapping smooth regression coefficients by particle diameter. The result shows a positive relationship between participants in the highest quartile with a functional risk score and cardiovascular injury risk.

Das, Lahiri, & Das (2018) stated the problem of sea ice extent (SIE) for the Arctic and Southern Oceans. This is because the analysis of SIE has been rarely considered. In the study, researchers treat the 'satellite passive microwave remote sensing' daily data for a year as functional data for both the Arctic and the Southern Oceans. Smoothing techniques in FDA, Fourier basis had been used to estimate the smooth functions for 37 years (1979–2015) because FDA is useful for missing value and unequal length prediction. From that, the study shows strong statistical evidence of a decrease in SIE the of Arctic Ocean over year-blocks while no statistical evidence has been found in the Southern Ocean.

Real-time and high-frequency data generated by smart energy meters provide demand management, and the response of consumers that can give benefits to individual and social. Thus, the FDA used to ensure these data into usable

recommendations apart from that to identify drivers of residential electricity load curves. The result of the study revealed that residential electricity consumption throughout the day that depends on the ownership of electrical appliances and an overall reduction of consumption after the introduction of real-time feedback was unrelated to appliance ownership characteristics (Fontana, Tavoni, & Vantini, 2019).

In the economic field, data in the Penn World Tables (PWT) and the amendments are important for many countries that will lead to different versions of the PWT. Thus, in the study, the researchers used functional data analysis to study the distribution functions of GDP across different versions of the PWT. Furthermore, GDP in different versions does not share a common underlying distribution. Based on the result, some evidence is shown to support the hypothesis that GDP's distribution functions differ across versions while complementing the results of previous studies (Chen, DeJuan, & Tian, 2018). In conclusion, J. Martínez et al. (2014) stated the advantages of the FDA which is it takes into account the time correlation structure of the data and the comparisons are more general in which leading to a global vision of the problem.

2.4 Summary

The above observation showed that FDA widely used in various fields of scholarly studies. Different features of the FDA can be used in solving future studies. Apart from that, a better understanding of O_3 helps people become more concerned about the effects of exposure that may affect human health and nature. The next chapter will explain the research methodology that will be used in conducting this study.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Data Collection

This research was used hourly data of O_3 concentration measured in μgm^{-3} that is obtained from the DOE Malaysia - Alam Sekitar Malaysia Sdn. Bhd (ASMA). The records were collected from the monitoring station at Sekolah Menengah Dato Permaisuri, Miri, Sarawak, Malaysia started from 1st January 2014 to 31st December 2015.

3.2 Data Analysis

In this research, three stages are involved in data analysis. First, converting data to functional or curve forms from point values. Secondly, the stage involved detecting anomalies in O_3 functional data in the selected station. This stage then continued with an analysis of the effectiveness of the method of detection and profiling in which the anomalies were identified (Shaadan, Deni, Jemain, & Latif, 2015). At last, a quantitative analysis was carried out to assess the functional curve of O_3. The analysis is used the "FDA" package in R software that is accessible. Figure 3.1 below shows the research phase that required in this study.

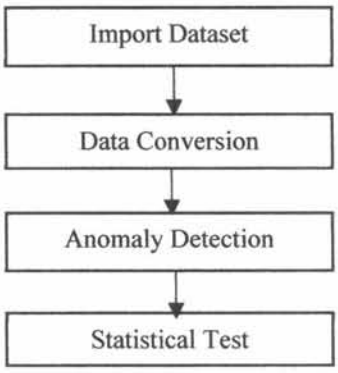

Figure 3.1 Steps of Data Analysis

3.3 Data Conversion

According to Shaadan, Rosslan, Deni, & Ismail (2018), discrete observed hourly data within one day have been converted into continuous data in the form of a curve represented by a predetermined function.

In this analysis, the curve data were modeled using a collection of spline functions given as *x(t)* where t ϵ [0, 24] is a continuous-time function of the O_3 level. With basis function, the daily O_3 cycle at any monitoring station on any day *i* can be expressed as:

$$x_i(t) = \sum_{k=1}^{K} C_k \emptyset_k (t) \tag{3.1}$$

where:

- $\emptyset_{k,}(t)$ is the spline basis system consisting of *K* number of basis functions,
- C_k is the basis coefficient.

The coefficient C_k is calculated using the ordinary least squares (OLS) method while a *K* number of appropriate spline basis is calculated using the Bayesian Information Criteria (BIC). The *K* value must be first calculated before applying the basis function expansion method where *K* with the smallest BIC value is selected. Let *m* denotes the number of data points recorded (*m*=24), *K* is the number of basis and *RSS* is the residual sum of the squared or error variance of the estimated mean curve. The following is the BIC formula:

$$BIC = \log\left(\frac{RSS}{m}\right) + \frac{k}{m}\log(m) \tag{3.2}$$

Given the observed data *y* at the point time *j*, the formula of *RSS* is as follows,

$$RSS = \sum_{j=1}^{24}(y_i - x(t_j))^2 - \sum_{j=1}^{24}(y_j - \sum_{k=1}^{K} C_k \varphi_k(t_j)^2) \tag{3.3}$$

The basis function can be used to determine which basis is best depending on the nature of the data. Fourier basis is appropriate for periodic data, while spline is more appropriate for non-periodic data (Shaadan, Jemain, Latif, & Deni, 2015).

3.4 Anomaly Detection

Multivariate robust Mahalanobis distance method is used to detect anomalies (Hyndman & Shang, 2010). This approach is made up of two sub-procedures, according to Shaadan, Jemain, Latif, and Deni (2015). The first sub-procedure is the method for the study of projection and the second is the calculation of the measures for the outward curve procedures. The first procedure is aimed at identifying the discretized curves for specific linear projections. Using the principal component analysis (PCA), the discrete curves are projected into p dimensional space. The search for the predicted curves follows the formula of a matrix X of length N rows by n columns:

$$Vu = \lambda u \tag{3.4}$$

where:

- V is the sample variance-covariance matrix, that is $V=N^{-1}X^{T}X$,
- u is an eigenvector of V,
- λ is an eigenvalue of V.

The solution can be obtained by finding the first p expected weight vector or proprietary vector that maximizes data variance. The scores projected are given by $s_1=u_1^{T}X, s_2=u_2^{T}X, \ldots s_p=u_p^{T}X$. Eigenvectors are orthogonal to one another in which the first is determined by the largest variation in the data collection, the latter by the second-largest variation, etc.

For the second procedure, the square robust Mahalanobis distance for each curve $D^2(X_i)$ was determined by considering the first two projections or key components. This was achieved by using the projected scores obtained from the first procedure where the new matrix data set was defined to be bivariate covariate $X = [s_1, s_2]$. $D^2(X_i)$ is an indicator of the curve's outwardness. The increasing the value, the more outward a curve is from the group's center. The $D^2(X_i)$ interpretation is as follows:

$$D^2(x_i) = (x_i - \bar{x})^T \sum^{-1} (x_i - \bar{x}) \quad (3.5)$$

where:

- x_i is a vector of measured points for the curve i,
- $\bar{x}$ is the location estimator (mean vector),
- $\sum$ is the robust estimate of the covariance matrix of X.

The cut-off point to differentiate between anomalous and non-anomalous curve was based on the critical value of the $X_{1-a,p}^2$, which predefined α quantile of the distribution with p degree of freedom, by assuming that the data are generated from a chi-square distribution. Higher α values are higher cut-off levels, resulting in lower percentages of anomalies observed. The greater the square value of the stable distance from Mahalanobis, the more outward the curve.

The stable covariance matrix calculation is assumed as the ideal subsample h covariance. If the minimum determinant of the covariance matrix or minimum covariance determinant (MCD) method is used, the subsample will be considered optimal. It is assumed that the value of h is the total number of curves that should not be outward. By using MCD in the calculation process, the outlying points are ignored and MCD was established as follows:

$$MCD = (\bar{x}_L^*, s_L^*) \tag{3.6}$$

where:

- L is the subsample matrix h with the minimum covariance matrix determinant with $h = [(N + p + 1)/2]$, $\bar{x}_L^* = \frac{1}{h}\sum_{i \in L}(x_i)$ and $s_L^* = \sum_{i \in L}\left(x_i - \bar{x}_L^*\right)\left(x_i - \bar{x}_L^*\right)^T$,
- $\sum = s_L^*$ is the sample covariance estimate.

3.5 Statistical Technique

Functional principal components analysis or FPCA is the first approach to be used after conducting descriptive statistics and plots. According to Gentleman, Hornik, & Parmigiani, (2009), FPCA is used to identify the primary mode of variation and to assess whether or not they appear to be important. Shaadan, Rosslan, Deni, & Ismail (2018) clarified that FPCA is expanding the notion of extracting key features of data functions based on the conventional PCA technique from multivariate data. Each principal component is defined by an eigen weight function $\xi(t)$ that reflects the dominant features of variance in the functional data or curves. First, find the principal component eigen weight function $\xi_1(t)$ by selecting those that maximize the mean square of the z_{i1} value, where

$$z_{i1} = \int \xi_1(t)\, x_i(t) dt \tag{3.7}$$

Subject to constraint:

$$\int \xi^2(t) dt = 1 \tag{3.8}$$

Next, compute eigen weight function $\xi_j(t)$ for $j = 2,\ldots,p$ where p is the number of principal components required. The computation is subjected to another constraint below:

$$\int \xi_j^2(t)dt = 1 \tag{3.9}$$

and the additional constraint:

$$\int \xi_j(t)\xi_1(t)dt = \int \xi_j(t)\xi_2(t)dt = \cdots = \int \xi_j(t)\xi_{j-1}(t)dt = 0 \tag{3.10}$$

The eigen weight function can be obtained by solving a Fredholm function given by:

$$\int v(s,t)(t)\xi(t)dt = \gamma\xi(s) \tag{3.11}$$

where,

- γ is an eigenvalue
- $v(s,t)$ is the variance-covariance function of the data set.

The number of principal components (PC) was determined based on the scree plot analysis that required to be retained to clarify the data variation. The plot shows the relation between the eigenvalues (y-axis) and the number of eigenvalues (x-axis). The first PC represents the most significant variation contained in the data collection, the second PC represents the second most important variation, followed by the third PC, and so on. The amount of variation described by each PC is calculated by the related eigenvalue (Shaadan, Rosslan, Deni, & Ismail, 2018).

CHAPTER 4
RESULT AND DISCUSSION

4.1 Data Conversion

Figure 4.1 shows that O_3 behavior varied between the years 2014 and 2015. In 2014, the curves at the higher level showed extreme behavior, while in 2015 the curves were almost smooth. Thus, the obvious difference in the daily hourly curve could be attributed to the different dominant source of emissions. Changes in the weather condition contributed to the incidence of the maximum peaks at about 2:00 p.m. in 2014.

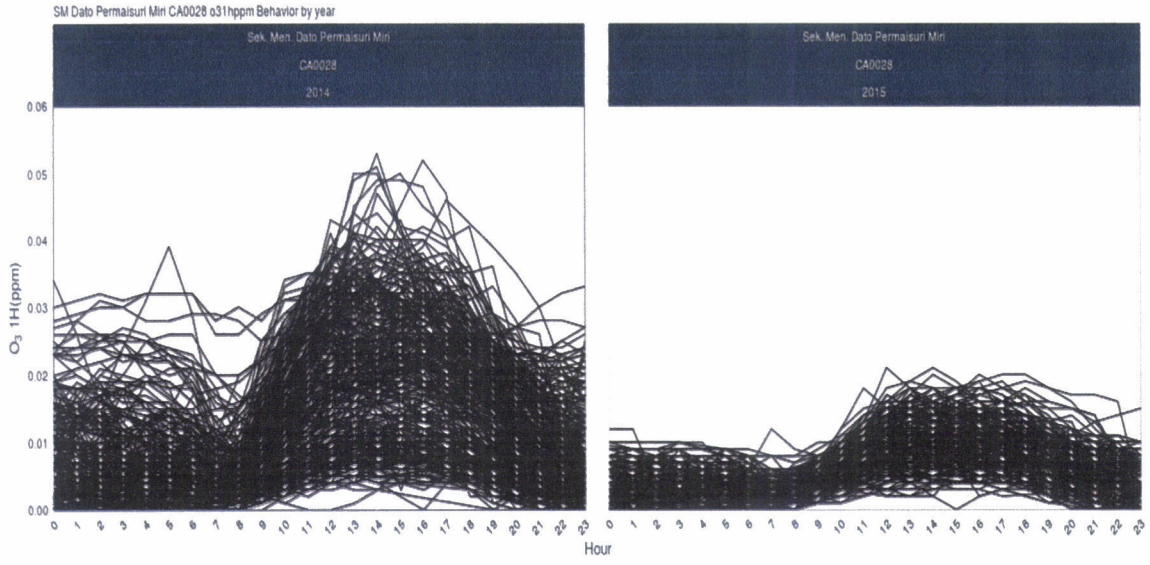

Figure 4.1: O_3 curves behavior by year

Figure 4.2 shows examples of the plot between BIC and the number of basis (K) for Sekolah Menengah Dato Permaisuri, Miri, Sarawak station. In the range of [1,51], K is chosen based on the lowest BIC.

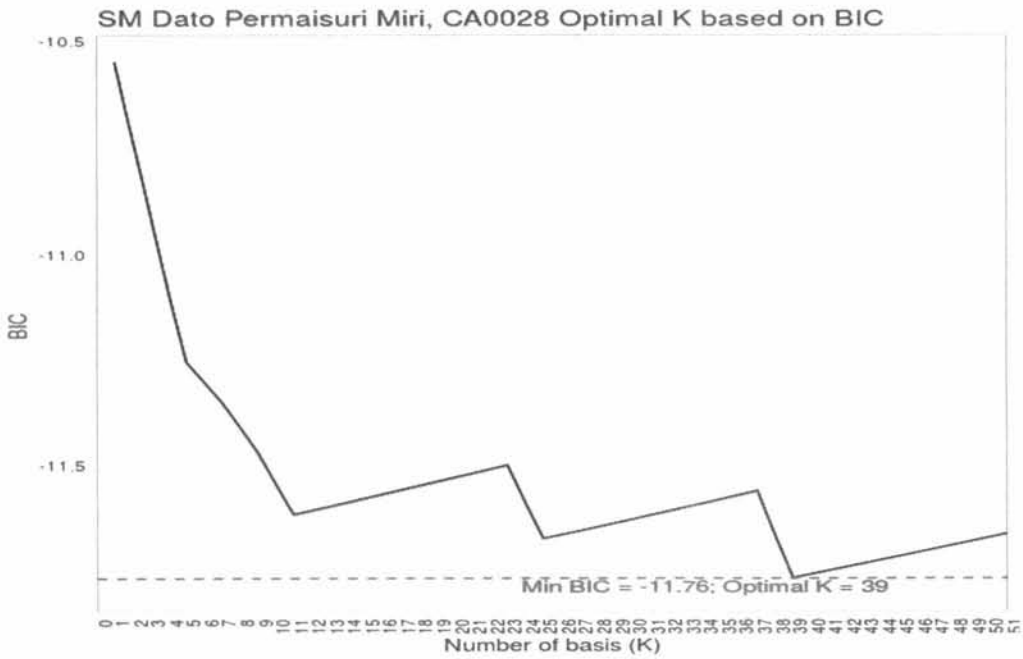

Figure 4.2: BIC Value with different number of basis (K) for Sekolah Menengah Dato Permaisuri monitoring station

Figure 4.3 shows the transformation of a fourier basis with optimal K (K=39). In this study, the basis function was determined using a fourier basis because the observed data is periodic data. At first, with K=39, the readings of O_3 were very high, but after a certain time, the level of O_3 keep dropping and stable.

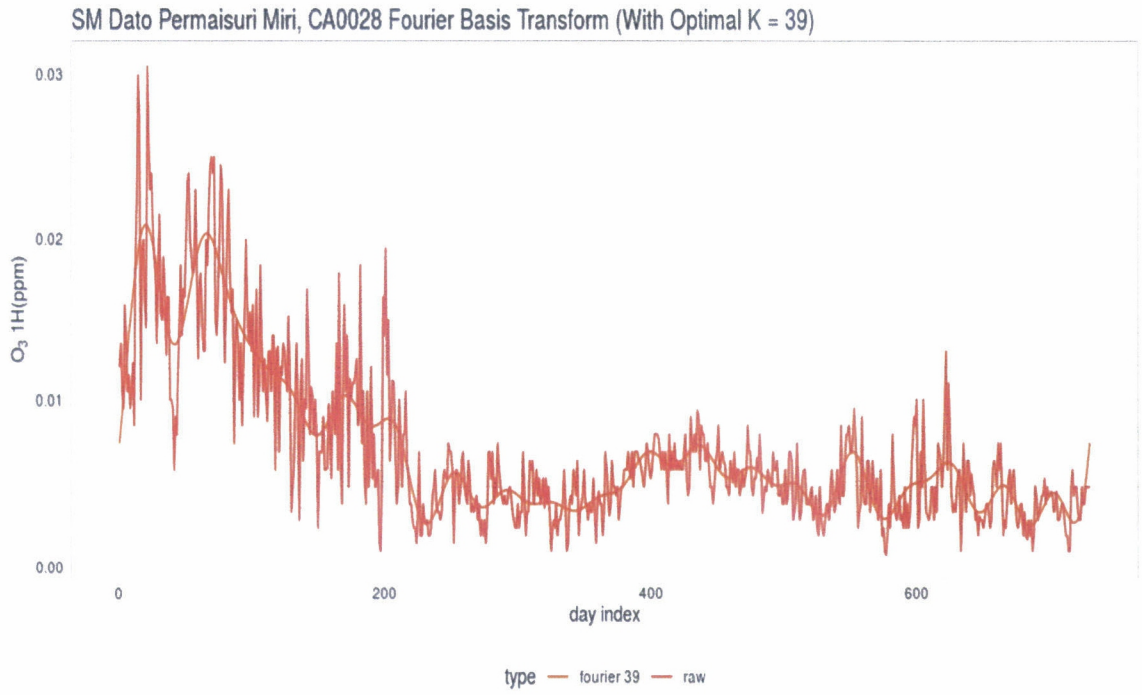

Figure 4.3: Fourier basis transform with optimal $K = 39$

4.2 Anomaly Detection

To detect anomalies, a Multivariate robust Mahalanobis distance method is used (Hyndman & Shang, 2010). Shaadan, Jemain, Latif, and Deni (2015) clarified that this approach consists of two sub-procedures, the first sub-procedure is the method for the projection analysis and the second is the calculation of the steps for the procedures for the outward curve. Figure 4.4 below shows the detection anomalies of O_3 hourly daily data at Sekolah Menengah Dato Permaisuri in Miri, Sarawak for years 2014 and 2015. Based on the figure below, anomalies were observed in the first half of 2014, while anomalies were not identified in 2015.

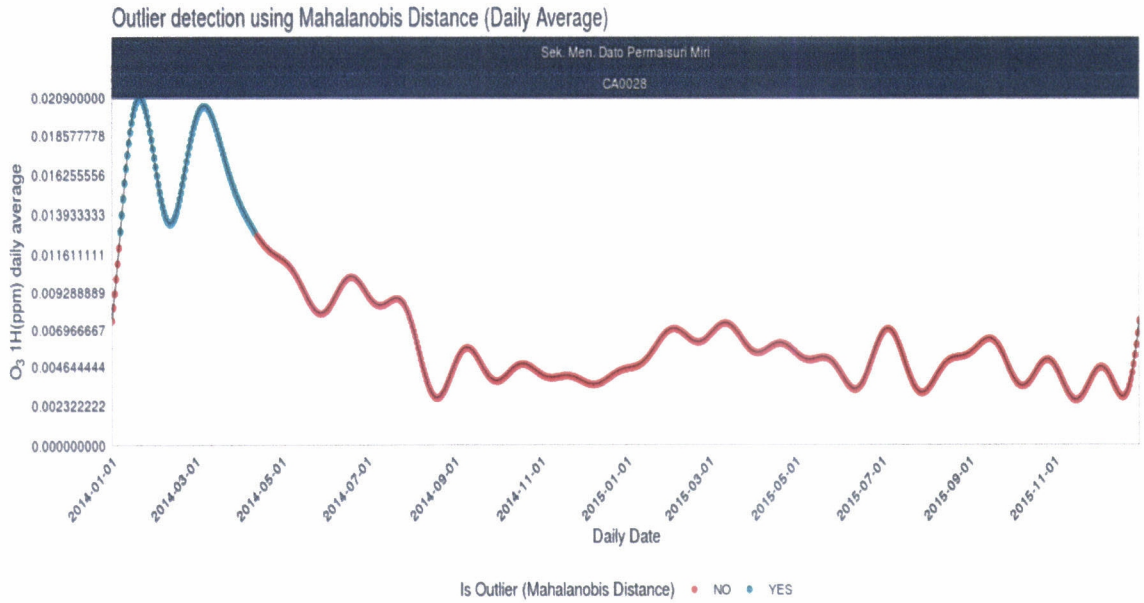

Figure 4.4: Anomalies detection using Mahalanobis Distance

4.3 Statistical Technique

Figure 4.5 shows the principal component of O_3 in Sekolah Menengah Dato Permaisuri at Miri monitoring station.

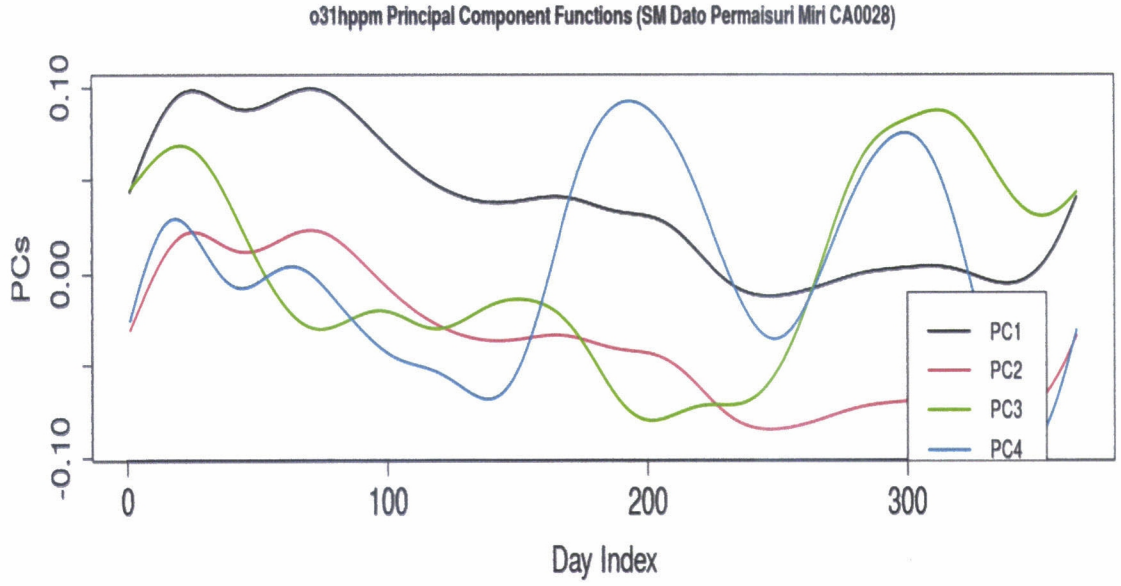

Figure 4.5: Principal component functions of O_3

Based on the figure, there are four principal components obtained from FPCA. PC1 represents the most important variation contained in the data set, the PC2 represents the second most important variation, then followed by the PC3 and PC4.

CHAPTER 5
CONCLUSION AND RECOMMENDATION

5.1 Conclusion

This study based on the assessment of the ozone, O_3 functional curve at the monitoring station in Miri, Sarawak for years 2014 and 2015. The results from the analysis revealed implicit information on the existence of two significantly different from the O_3 behaviors between 2014 and 2015. The result also provided that anomalies were detected in the first half of 2014 using the Mahalanobis Distance method. This showed that the diurnal behavior was influenced by the various dominant emission sources and other methodological conditions that existed in those years.

5.2 Recommendation

It is hoped that this statistical approach will be introduced as a new model in the methodology and procedure for future air quality assessment. This is because there are not many O_3 reports that can be used as sources. The findings obtained were based on the suggested statistics employed, which could be used as an alternative tool for further analysis not only for O_3 but also for PM10, NO_2, and CO. This study also can be done using another method so that the result from those methods can be compared with the result from our method. From that, we can see which methods are more effective to use in solving air quality assessment regarding O_3. Lastly, other than using different methods, conducting the study at different locations can be recommended for future study.

REFERENCES

Air quality in Malaysia's Sarawak has hit a hazardous level of 367 - and 2 people in Indonesia have died, Business Insider - Business Insider Malaysia. (n.d.). Retrieved December 3, 2019, from https://www.businessinsider.my/air-quality-in-malaysias-sarawak-has-hit-a-hazardous-level-of-367-and-2-people-in-indonesia-have-died/

Azmi, S. Z., Latif, M. T., Ismail, A. S., Juneng, L., & Jemain, A. A. (2010). Trend and status of air quality at three different monitoring stations in the Klang Valley, Malaysia. *Air Quality, Atmosphere and Health*, *3*(1), 53–64. https://doi.org/10.1007/s11869-009-0051-1

BERNAMA.com - Haze: 25 schools in Putrajaya will be closed tomorrow if API reading exceeds 200. (n.d.). Retrieved December 8, 2019, from http://www.bernama.com/en/news.php?id=1767928

Chen, T., DeJuan, J., & Tian, R. (2018). Distributions of GDP across versions of the Penn World Tables: A functional data analysis approach. *Economics Letters*, *170*, 179–184. https://doi.org/10.1016/j.econlet.2018.05.038

Das, P., Lahiri, A., & Das, S. (2018). Understanding sea ice melting via functional data analysis. *Current Science*, *115*(5), 920–929. https://doi.org/10.18520/cs/v115/i5/920-929

Department of Environment. (2010). *Air quality: Continuous air quality monitoring (CAQM)*. 1.

Fontana, M., Tavoni, M., & Vantini, S. (2019). Functional Data Analysis of high-frequency load curves reveals drivers of residential electricity consumption. *PLoS ONE*, *14*(6), 1–17. https://doi.org/10.1371/journal.pone.0218702

Gentleman, R., Hornik, K., & Parmigiani, G. (2009). *Functional data analysis with R and MATLAB- James Ramsay, Giles Hooker, Spencer Graves (auth.)-Springer-Verlag New York (2009)*. https://doi.org/10.1007/978-0-387-98185-7

Hyndman, R. J., & Shang, H. L. (2010). Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*, *19*(1), 29–45. https://doi.org/10.1198/jcgs.2009.08158

Jamaludin, S., & Zulkifli, Y. (2017). Spatial and temporal variabilities of rainfall data using functional data analysis. *Theoretical and Applied Climatology*, *129*(1–2), 229–242. https://doi.org/10.1007/s00704-016-1778-x

Levitin, D. J., Nuzzo, R. L., Vines, B., & Ramsay, J. O. (2007). Introduction to functional data analysis. *Canadian Psychology*, *48*(3), 135–155. https://doi.org/10.1037/cp2007014

Martínez, J., Saavedra, Á., García-Nieto, P. J., Piñeiro, J. I., Iglesias, C., Taboada, J., … Pastor, J. (2014). Air quality parameters outliers detection using functional data analysis in the Langreo urban area (Northern Spain). *Applied Mathematics and Computation*, *241*(2), 1–10. https://doi.org/10.1016/j.amc.2014.05.004

More schools closed in S'wak as haze spreads, 138,384 students affected | The Star Online. (n.d.). Retrieved December 4, 2019, from https://www.thestar.com.my/news/nation/2019/09/18/more-schools-closed-in-s039wak-as-haze-spreads-138384-students-affected#cxrecs_s

Morris, J. S. (2015). Functional Regression. *Annual Review of Statistics and Its Application*, *2*(1), 321–359. https://doi.org/10.1146/annurev-statistics-010814-020413

Official Website of Natural Resources And Environment Board Sarawak. (n.d.). Retrieved December 3, 2019, from https://www.nreb.gov.my/page-0-0-254-Air-Quality.html

Rahman, N. H. A., Lee, M. H., Suhartono, & Latif, M. T. (2016). Evaluation performance of time series approach for forecasting air pollution index in Johor, Malaysia. *Sains Malaysiana*, *45*(11), 1625–1633.

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

Rowland, C. M., Shiffman, D., Caulfield, M., Garcia, V., Melander, O., & Hastie, T. (2019). Association of cardiovascular events and lipoprotein particle size: Development of a risk score based on functional data analysis. *PLoS ONE*, *14*(3), 1–17. https://doi.org/10.1371/journal.pone.0213172

Ruth, M., & Franklin, R. S. (2014). Livability for all? Conceptual limits and practical implications. *Applied Geography*, *49*, 18–23. https://doi.org/10.1016/j.apgeog.2013.09.018

Shaadan, N., Deni, S. M., & Jemain, A. A. (2012). Assessing and comparing PM10 pollutant behaviour using functional data approach. *Sains Malaysiana*, *41*(11), 1335–1344.

Shaadan, N., Deni, S. M., & Jemain, A. A. (2015). Application of functional data analysis for the treatment of missing air quality data. *Sains Malaysiana*, *44*(10), 1531–1540. https://doi.org/10.17576/jsm-2015-4410-19

Shaadan, N., Jemain, A. A., Latif, M. T., & Deni, S. M. (2015). Anomaly detection and assessment of PM10 functional data at several locations in the Klang Valley, Malaysia. *Atmospheric Pollution Research*, *6*(2), 365–375. https://doi.org/10.5094/APR.2015.040

Shaadan, N., Rosslan, R. R., Deni, S. M., & Ismail, A. (2018). The diurnal Ozone dynamics profile at several locations in Selangor, Malaysia: A functional data analysis approach. *AIP Conference Proceedings*, *2013*(October). https://doi.org/10.1063/1.5054210

Ullah, S., & Finch, C. F. (2013). Applications of functional data analysis: A systematic review. *BMC Medical Research Methodology*, *13*(1). https://doi.org/10.1186/1471-2288-13-43

World Health Organization, W. (2005). *WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide: Global update 2005*. 1–21. https://doi.org/10.1016/0004-6981(88)90109-6

APPENDICES

APPENDIX A: DATASET FOR OZONE, O_3 IN MIRI MONITORING STATION

STATION ID	STATION ID	DATE TIME	OZONE (ppm)	DATE TIME	OZONE (ppm)
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 0:00	0.007	1/01/2015 0:00	0.004
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 1:00	0.008	1/01/2015 1:00	0.003
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 2:00	0.008	1/01/2015 2:00	0.003
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 3:00	0.007	1/01/2015 3:00	0.004
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 4:00	0.007	1/01/2015 4:00	0.003
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 5:00	0.006	1/01/2015 5:00	N/A
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 6:00	0.007	1/01/2015 6:00	0.002
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 7:00	N/A	1/01/2015 7:00	0.001

CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 8:00	0.004	1/01/2015 8:00	0.001
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 9:00	0.01	1/01/2015 9:00	0.003
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 10:00	0.017	1/01/2015 10:00	0.004
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 11:00	0.022	1/01/2015 11:00	0.007
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 12:00	0.024	1/01/2015 12:00	0.01
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 13:00	0.025	1/01/2015 13:00	0.011
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 14:00	0.027	1/01/2015 14:00	0.011
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 15:00	0.028	1/01/2015 15:00	0.011
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 16:00	0.03	1/01/2015 16:00	0.009
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 17:00	0.03	1/01/2015 17:00	0.007
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 18:00	0.029	1/01/2015 18:00	0.008

CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 19:00	0.021	1/01/2015 19:00	0.008
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 20:00	0.014	1/01/2015 20:00	0.008
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 21:00	0.006	1/01/2015 21:00	0.009
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 22:00	0.009	1/01/2015 22:00	0.008
CA0028	Sek. Men. Dato Permaisuri Miri	1/01/2014 23:00	0.014	1/01/2015 23:00	0.008
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 0:00	0.014	2/01/2015 0:00	0.008
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 1:00	0.015	2/01/2015 1:00	0.008
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 2:00	0.012	2/01/2015 2:00	0.008
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 3:00	0.012	2/01/2015 3:00	0.01
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 4:00	0.013	2/01/2015 4:00	N/A
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 5:00	0.011	2/01/2015 5:00	0.009

CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 6:00	N/A	2/01/2015 6:00	0.007
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 7:00	0.006	2/01/2015 7:00	0.005
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 8:00	0.006	2/01/2015 8:00	0.006
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 9:00	0.009	2/01/2015 9:00	0.006
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 10:00	0.012	2/01/2015 10:00	0.005
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 11:00	0.015	2/01/2015 11:00	0.006
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 12:00	0.015	2/01/2015 12:00	0.007
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 13:00	0.015	2/01/2015 13:00	0.007
CA0028	Sek. Men. Dato Permaisuri Miri	2/01/2014 14:00	0.018	2/01/2015 14:00	0.007

**Table above the only consist of first 40 datasets of ozone, O_3 concentration for 2014 and 2015.

APPENDIX B: SAMPLE COMMAND IN R SOFTWARE

1. Define Functional Principal Component Analysis (FPCA)

```
# 1. Import Dataset ------------------------------------------------------
-
# libraries
library(tidyverse)
── Attaching packages ─────────────────────────────────── tidyverse 1.3.0 ──
✓ ggplot2 3.3.2     ✓ purrr   0.3.4
✓ tibble  3.0.1     ✓ dplyr   1.0.0
✓ tidyr   1.1.0     ✓ stringr 1.4.0
✓ readr   1.3.1     ✓ forcats 0.5.0
── Conflicts ────────────────────────────────────── tidyverse_conflicts() ──
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()

ozoneppm <- fs::dir_ls(path = '.', regexp = '.csv') %>%
+   furrr::future_map_dfr(readr::read_csv, col_names = F, skip = 1, col_ty
pes = cols(.default = "c"), trim_ws = T) %>% janitor::clean_names()

# define col type
ozoneppm$x4 <- as.numeric(ozoneppm$x4)
ozoneppm$x5 <- lubridate::mdy(ozoneppm$x5)
ozoneppm$x6 <- paste0(str_pad(ozoneppm$x6, 2, pad = "0"), ':00')

# create date time column
ozoneppm$datetime <- lubridate::ymd_hm(paste(ozoneppm$x5, ozoneppm$x6))

#ozoneppm$lbyear <- lubridate::year(ozoneppm$datetime)
#ozoneppm$lbmonth <- lubridate::month(ozoneppm$datetime)

# convert excel date number to proper datetime
# ozoneppm$x3 <- as.POSIXct(ozoneppm$x3 * (60*60*24)
#               , origin="1899-12-30"
#               , tz="GMT")
# clean date time to nearest hour
# ozoneppm$x3 <- format(round(ozoneppm$x3), format="%D %H:%M")

# lubridate
# ozoneppm$x3 <- lubridate::mdy_hm(ozoneppm$x3)

# create actual grid
hourlyGrid <- seq(min(ozoneppm$datetime), max(ozoneppm$datetime), by = "ho
urs")
tdf <- tibble(datetime = hourlyGrid)
tdf$dateyearmonth <- tsibble::yearmonth(tdf$datetime)
tdf$dateyear <- lubridate::year(tdf$datetime)
tdf$datedaymonth <- lubridate::day(tdf$datetime)
```

```
tdf$datedayyear <- as.numeric(strftime(tdf$datetime, format = "%j"))
tdf$dateday <- as.Date(tdf$datetime)

ozoneppm <- ozoneppm %>% dplyr::select(-x3)
colnames(ozoneppm) <- c('station', 'location','o31hppm', 'mydate', 'myhour
', 'datetime')
ozoneppm <- tdf %>% left_join(ozoneppm) %>% dplyr::select(station, locatio
n, o31hppm, everything())
Joining, by = "datetime"
ozoneppm %>% group_by(station,dateyear) %>% tally()
# A tibble: 7 x 3
# Groups:   station [3]
  station dateyear     n
  <chr>      <dbl> <int>
1 CA0028      2014  8760
2 CA0028      2015  8760
3 CA0028      2016  8784
4 CA56Q       2017  4344
5 CA56Q       2018  8759
6 NA          2017  4416
7 NA          2018     1

# clean location
CapStr <- function(y) {
+   y <- tolower(y)
+   c <- strsplit(y, " ")[[1]]
+   paste(toupper(substring(c, 1,1)), substring(c, 2),
+         sep="", collapse=" ")
+ }
# capitalize only first letter of a word
ozoneppm$location <- unlist(lapply(ozoneppm$location, CapStr))
glimpse(ozoneppm)
Rows: 43,824
Columns: 11
$ station       <chr> "CA0028", "CA0028", "…
$ location      <chr> "Sek. Men. Dato Perma…
$ o31hppm       <dbl> 0.007, 0.008, 0.008, …
$ datetime      <dttm> 2014-01-01 00:00:00,…
$ dateyearmonth <mth> 2014 Jan, 2014 Jan, 2…
$ dateyear      <dbl> 2014, 2014, 2014, 201…
$ datedaymonth  <int> 1, 1, 1, 1, 1, 1, 1, …
$ datedayyear   <dbl> 1, 1, 1, 1, 1, 1, 1, …
$ dateday       <date> 2014-01-01, 2014-01-…
$ mydate        <date> 2014-01-01, 2014-01-…
$ myhour        <chr> "00:00", "01:00", "02…


# check dups
#sum(duplicated(ozoneppm))
```

```
ggplot(data = ozoneppm %>% filter(), aes(x = datetime, y = o31hppm)) + geo
m_line() + theme_bw() +
+   facet_wrap(location~station~dateyear, scales = 'free_x') #+
#labs(y = expression(paste(O[3],"  ", "1H(ppm)")), x = 'day
index', x = 'Date')
# subset to CA0028, 2014 & 2015
ozoneppm <- ozoneppm %>% filter(dateyear %in% c(2014:2015) & station %in%
c('CA0028'))
# visualize filtered result
ggplot(data = ozoneppm, aes(x = datetime, y = o31hppm)) + geom_line() + th
eme_bw() +
+   facet_wrap(location~station~dateyear, scales = 'free_x') +
+   labs(y = expression(paste(O[3],"  ", "1H(ppm)")), x = 'day index', x =
'Date')
# remove duplicates if any
ozoneppm <- ozoneppm %>% distinct()
# check length size
ozoneppm %>% group_by(station, dateyear) %>% tally()
# A tibble: 2 x 3
# Groups:   station [1]
  station dateyear     n
  <chr>      <dbl> <int>
1 CA0028      2014  8760
2 CA0028      2015  8760
# note that each year has different length
table(ozoneppm$station) # check length by station

CA0028
 17520

# interpolate missing values
sum(is.na(ozoneppm)) / nrow(ozoneppm) * 100 # 1.6% missing value
[1] 6.392694
ozoneppm$o31hppm <- forecast::na.interp(ozoneppm$o31hppm)
sum(is.na(ozoneppm)) / nrow(ozoneppm) * 100 # 0% missing value
[1] 0
# visualize interpolated result
ggplot(data = ozoneppm, aes(x = datetime, y = o31hppm)) + geom_line() + th
eme_bw() +
+   facet_wrap(location~station~dateyear, scales = 'free_x') +
+   labs(y = expression(paste(O[3],"  ", "1H(ppm)")), x = 'day index', x =
'Date')
> # check length
> table(ozoneppm$station)

CA0028
 17520
range(ozoneppm$dateday)
[1] "2014-01-01" "2015-12-31"
# calculate daily average, uses median as average calculation
```

```
daily <- ozoneppm %>% group_by(station, location, dateday) %>%
+   summarise(o31hppmavg = median(o31hppm, na.rm = T)) %>% ungroup() %>%
+   dplyr::select(station, location, dateday, o31hppmavg)
`summarise()` regrouping output by 'station', 'location' (override with `.
groups` argument)

# check dimensions
dim(daily)
[1] 730   4

# date variables
daily$dateyearmonth <- tsibble::yearmonth(daily$dateday)
daily$dateyear <- lubridate::year(daily$dateday)
daily$datedaymonth <- lubridate::day(daily$dateday)
daily$datedayyear <- as.numeric(strftime(daily$dateday, format = "%j"))

# daily plot
p <- ggplot(data = daily, aes(x = dateday, y = o31hppmavg, color = locatio
n)) +
+   geom_line() + tidyquant::theme_tq() +
+   facet_wrap(location~station~dateyear, scales = 'free') +
+   labs(y = expression(paste(O[3]," ", "1H(ppm)")), x = 'day index', x =
'Date')+
+   scale_x_date(expand = c(0,0),breaks = seq(as.Date('2014-01-01'), as.Da
te('2016-12-31'), by = "1 month"), date_labels = "%Y %b") +
+   theme(axis.text.x = element_text(angle = 90))
p


# 2. Data Conversion - Fourier Basis ----------------------------
--------------
# load script fourier.R
source(file = 'fourier.R')
# use 39 nbasis for CA20B (optimal K based on BIC)

# 2. Data Conversion - Fourier Basis ----------------------------------
----
# load script fourier.R
source(file = 'fourier.R')
K= 1 ;BIC= -10.5473292661031 ;RMSE= 0.005 ;SSE= 0.019 ;GCV= 0
K= 3 ;BIC= -10.9087557520981 ;RMSE= 0.004 ;SSE= 0.013 ;GCV= 0
K= 5 ;BIC= -11.2584173965134 ;RMSE= 0.003 ;SSE= 0.009 ;GCV= 0
K= 7 ;BIC= -11.3581372964598 ;RMSE= 0.003 ;SSE= 0.008 ;GCV= 0
K= 9 ;BIC= -11.4736055533744 ;RMSE= 0.003 ;SSE= 0.007 ;GCV= 0
K= 11 ;BIC= -11.6096930974916 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 13 ;BIC= -11.5916299617817 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 15 ;BIC= -11.5735668260717 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 17 ;BIC= -11.5555036903617 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 19 ;BIC= -11.5374405546517 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 21 ;BIC= -11.5193774189417 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
```

```
K= 23 ;BIC= -11.5013142832318 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 25 ;BIC= -11.6655727043157 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 27 ;BIC= -11.6475095686058 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 29 ;BIC= -11.6294464328958 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 31 ;BIC= -11.6113832971858 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 33 ;BIC= -11.5933201614758 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 35 ;BIC= -11.5752570257658 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 37 ;BIC= -11.5571938900559 ;RMSE= 0.002 ;SSE= 0.005 ;GCV= 0
K= 39 ;BIC= -11.7622743056601 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 41 ;BIC= -11.7442111699501 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 43 ;BIC= -11.7261480342401 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 45 ;BIC= -11.7080848985301 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 47 ;BIC= -11.6900217628202 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 49 ;BIC= -11.6719586271102 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 51 ;BIC= -11.6538954914002 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
```

```
daily <- bind_cols(daily, generateFourierDaily(daily,39) %>% dplyr::select
(value))
K= 39 ;BIC= -11.7622743056601 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
colnames(daily)[9] <- 'fourierOptimal'
ggplot(daily, aes(x = dateday,y = fourierOptimal)) + geom_line()
```

```
# 3. Detection of Anomalies - Mahalanobis Distance ----------------------
----
library(StatMatch)
md <- as_tibble(mahalanobis(x = daily %>% dplyr::select(fourierOptimal),
+                            center = colMeans(daily %>% dplyr::select(fo
urierOptimal)),
+                            cov = cov(daily %>% dplyr::select(fourierOpt
imal))))
colnames(md) <- 'MahalanobisDist'
md$dateday <- daily$dateday
md <- md %>% dplyr::select(dateday, MahalanobisDist, everything()) %>% arr
ange(desc(MahalanobisDist))
outliers <- boxplot(md$MahalanobisDist)
```

```
dfMahalanobisDis <- tibble(MahalanobisDist = outliers$out,
+                           outlierMahalanobisDis = 'YES')
md <- md %>% left_join(dfMahalanobisDis) %>%
+   mutate(outlierMahalanobisDis = ifelse(is.na(outlierMahalanobisDis),'NO
', 'YES')) %>% arrange(dateday)
Joining, by = "MahalanobisDist"
table(md$outlierMahalanobisDis)

 NO YES
633  97
```

```
daily <- daily %>% dplyr::select(station, location, dateday, o31hppmavg, f
ourierOptimal) %>% left_join(md) %>% distinct()
Joining, by = "dateday"
daily$outlierMahalanobisDis <- factor(daily$outlierMahalanobisDis)
table(daily$outlierMahalanobisDis)/ nrow(daily) * 100 # 13.29% of daily ou
tliers

      NO      YES
86.71233 13.28767
max(daily$o31hppmavg, na.rm = T)
[1] 0.0305
min(daily$o31hppmavg, na.rm = T)
[1] 0.00075
max(daily$fourierOptimal, na.rm = T)
[1] 0.02089111
min(daily$fourierOptimal, na.rm = T)
[1] 0.002649759


# date variables
daily$dateyearmonth <- tsibble::yearmonth(daily$dateday)
daily$dateyear <- lubridate::year(daily$dateday)
daily$datedaymonth <- lubridate::day(daily$dateday)
daily$datedayyear <- as.numeric(strftime(daily$dateday, format = "%j"))
p <- ggplot(data = daily, aes(x = dateday, y = fourierOptimal, color = out
lierMahalanobisDis)) +
+   geom_point() + geom_line(color = 'black', alpha = 0.5) + facet_wrap(lo
cation~station, scales = 'free') +
+   tidyquant::theme_tq() +
+   scale_x_date(expand = c(0,0), breaks = seq(as.Date("2014-01-01"), as.D
ate("2015-12-31"), by = "2 month")) +
+   scale_y_continuous(expand = c(0,0), limits = c(0, 0.0209), breaks = se
q(from = 0,to = 0.0209,length.out = 10)) +
+   theme(axis.text.x=element_text(angle=45, hjust=1)) +
+   labs(color = "Is Outlier (Mahalanobis Distance)",
+        title = "Outlier detection using Mahalanobis Distance (Daily Aver
age)",
+        x = "Daily Date",
+        y = expression(paste(O[3]," ", "1H(ppm) daily average")))
p


# add dateday col to ozoneppm
colnames(ozoneppm)
 [1] "station"       "location"      "o31hppm"       "datetime"
 [5] "dateyearmonth" "dateyear"      "datedaymonth"  "datedayyear"
 [9] "dateday"       "mydate"        "myhour"
ozoneppm <- ozoneppm %>% mutate(hour = lubridate::hour(datetime))


# only plot the anomaly days
```

```
outlierdays <- daily %>% filter(outlierMahalanobisDis == 'YES' & station =
= 'CA0028') %>% ungroup() %>% dplyr::select(station, dateday, outlierMahal
anobisDis) %>% distinct()
length(outlierdays) / 365 * 100
[1] 0.8219178
max(ozoneppm$o31hppm, na.rm = T)
[1] 0.053
> min(ozoneppm$o31hppm, na.rm = T)
[1] 0
p <- ggplot(data = ozoneppm %>% filter(dateday %in% outlierdays$dateday &
station == 'CA0028')
+             ,aes(x = hour, y = o31hppm, group = dateday)) +
+   geom_point() + geom_line() +
+   tidyquant::theme_tq() +
+   scale_y_continuous(expand = c(0,0), limits = c(0, 0.055), breaks = seq
(from = 0,to = 0.055,length.out = 6)) +
+   scale_x_continuous(expand = c(0,0), limits = c(0, 23), breaks = seq(fr
om = 0,to = 23,length.out = 24)) +
+   theme(axis.text.x=element_text(angle=45, hjust=1)) +
+   labs(title = "SM Dato Permaisuri Miri, CA0028 Behavior of the maximum
level of anomalies (days with hourly anomalies)",
+        x = "Hour",
+        y = expression(paste(O[3]," ", "1H(ppm)")), x = 'day index') +
+   theme(legend.position = "none") +
+   geom_hline(yintercept = median(ozoneppm$o31hppm),  linetype="dashed",
color = "#6A1C2B", size = 2) +
+   annotate("text", x=14, y=median(ozoneppm$o31hppm)-0.002, label= 'Media
n', color = "#6A1C2B", size = 4) +
+   theme(plot.title = element_text(size=7))
P

# 4. Functional Principal Component Analysis ----------------------------
----
library(fda)
daybasis365 = create.fourier.basis(c(0,365),365)
harmLfd = vec2Lfd(c(0,(2*pi/365)^2,0), c(0, 365))
daily %>% group_by(dateyear) %>% tally()
# A tibble: 2 x 2
  dateyear     n
     <dbl> <int>
1     2014   365
2     2015   365
CA0028_2014 <- daily %>% filter(station == 'CA0028' & dateyear == '2014')
%>% dplyr::select(o31hppmavg) %>% rename(CA0028_2014 = o31hppmavg) %>% hea
d(365)
CA0028_2015 <- daily %>% filter(station == 'CA0028' & dateyear == '2015')
%>% dplyr::select(o31hppmavg) %>% rename(CA0028_2015 = o31hppmavg) %>% hea
d(365)
CA0028_daily <- cbind(as.matrix(CA0028_2014), as.matrix(CA0028_2015))
tempfdPar = fdPar(daybasis365,harmLfd,1e4)
```

```
tempfd_CA0028 = smooth.basis(1:365,CA0028_daily,tempfdPar)
plot(tempfd_CA0028$fd,xlab='Day Index',ylab=expression(paste(O[3]," ", "1H
(ppm)")),cex.lab=1.5,cex.axis=1.5,col=4)
[1] "done

tempfdPar2 = fdPar(daybasis365,harmLfd,1e-2)
tempfd2_CA0028 = smooth.basis(1:365,CA0028_daily,tempfdPar2)
plot(tempfd2_CA0028$fd,xlab='Day Index',ylab=expression(paste(O[3]," ", "1
H(ppm)")),cex.lab=1.5,cex.axis=1.5,col=4)
[1] "done"

tempfdPar3 = fdPar(daybasis365,harmLfd,1e6)
tempfd3_CA0028 = smooth.basis(1:365,CA0028_daily,tempfdPar3)
plot(tempfd3_CA0028$fd,xlab='Day Index',ylab=expression(paste(O[3]," ", "1
H(ppm)")),cex.lab=1.5,cex.axis=1.5,col=4)
[1] "done"

# do FPCA with a roughness penalty on FPCs.
# We use the curves estimated with the smoothing spline
# with the smoothing parameter lambda = 10^{-2}
# Here we define the roughness penalty by the harmonic acceleration differ
ential operator
# with the smoothing parameter lambda = 10^6
ptemppca_CA0028 = pca.fd(tempfd2_CA0028$fd,nharm=4,harmfdPar=tempfdPar3)
options(scipen = 999)
round(ptemppca_CA0028$varprop,4) * 100
[1] 100   0   0   0

# Get FPCs
pharmfd_CA0028 = ptemppca_CA0028$harmonics
pharmvals_CA0028 = eval.fd(1:365,pharmfd_CA0028)
matplot(1:365,pharmvals_CA0028,xlab='Day Index',ylab='PCs',
+         lwd=2,lty=1,cex.lab=1.5,cex.axis=1.5,type='l')
> legend(300,-0.01,c('PC1','PC2','PC3','PC4'),col=1:4,lty=1,lwd=2)
> title('o31hppm Principal Component Functions (SM Dato Permaisuri Miri CA
0028)', cex.main=1)

# hourly chart
ozoneppm <- ozoneppm %>% mutate(hour = hour(datetime))
p <- ggplot(data = ozoneppm, aes(x = hour, y = o31hppm, group = factor(dat
eday))) +
+   geom_line() + facet_wrap(location~station~dateyear) +
+   tidyquant::theme_tq() +
+   scale_y_continuous(expand = c(0,0), limits = c(0, 0.06), breaks = seq
(from = 0,to = 0.06,length.out = 7)) +
+   scale_x_continuous(expand = c(0,0), limits = c(0, 23), breaks = seq(fr
om = 0,to = 23,length.out = 24)) +
+   theme(axis.text.x=element_text(angle=45, hjust=1)) +
+   labs(title = "SM Dato Permaisuri Miri CA0028 o31hppm Behavior by year
",
```

```
+         x = "Hour",
+         y = expression(paste(O[3]," ", "1H(ppm)")), x = 'day index') +
+    theme(legend.position = "none") +
+    theme(plot.title = element_text(size=9))
p

# convert to factor for summary
ozoneppm$station <- factor(ozoneppm$station)
ozoneppm$location <- factor(ozoneppm$location)
summary(ozoneppm %>% dplyr::select(-dateyearmonth))
station                                location        o31hppm
 CA0028:17520   Sek. Men. Dato Permaisuri Miri:17520   Min.   :0.000000
                                                      1st Qu.:0.003000
                                                      Median :0.006000
                                                      Mean   :0.007953
                                                      3rd Qu.:0.010000
                                                      Max.   :0.053000
    datetime                       dateyear     datedaymonth    datedayyear

 Min.   :2014-01-01 00:00:00   Min.   :2014   Min.   : 1.00   Min.   :  1

 1st Qu.:2014-07-02 11:45:00   1st Qu.:2014   1st Qu.: 8.00   1st Qu.: 92

 Median :2014-12-31 23:30:00   Median :2014   Median :16.00   Median :183

 Mean   :2014-12-31 23:30:00   Mean   :2014   Mean   :15.72   Mean   :183

 3rd Qu.:2015-07-02 11:15:00   3rd Qu.:2015   3rd Qu.:23.00   3rd Qu.:274

 Max.   :2015-12-31 23:00:00   Max.   :2015   Max.   :31.00   Max.   :365

    dateday               mydate              myhour               hour

 Min.   :2014-01-01   Min.   :2014-01-01   Length:17520       Min.   : 0.0
0
 1st Qu.:2014-07-02   1st Qu.:2014-07-02   Class :character   1st Qu.: 5.7
5
 Median :2014-12-31   Median :2014-12-31   Mode  :character   Median :11.5
0
 Mean   :2014-12-31   Mean   :2014-12-31                      Mean   :11.5
0
 3rd Qu.:2015-07-02   3rd Qu.:2015-07-02                      3rd Qu.:17.2
5
 Max.   :2015-12-31   Max.   :2015-12-31
Max.   :23.00
```

2. Define Fourier basis

```
library(fda)
# define Fourier Daily function, with nbasis as parameter
generateFourierDaily <- function(df, nbasis) {
+   tt <- 1:nrow(df)
+   range(tt)
+   dayindex <- seq(from = range(tt)[1], to = range(tt)[2])
+   basisobj = create.fourier.basis(rangeval = range(tt),nbasis = nbasis)
+
+   ys = smooth.basis(argvals = tt, method = 'qr', y = df$o31hppmavg, fdPa
robj = basisobj)
+   #str(ys)
+   fourier.fd = ys$fd
+   RMSE = round(sqrt(mean((eval.fd(tt, fourier.fd) - df$o31hppmavg)^2)),
3)
+   SSE = round(ys$SSE,3)
+   gcv = as.vector(round(ys$gcv,3))
+   BIC = log(SSE/nrow(df)) + (nbasis/nrow(df))*log(nrow(df))
+
+   # progress report:
+   cat(paste('K=',nbasis,';BIC=',BIC,';RMSE=',RMSE,';SSE=',SSE,';GCV=',gc
v,"\n"))
+
+   res <- tibble(dayindex = dayindex,
+                 value = as.vector(eval.fd(tt,fourier.fd)),
+                 type = paste0('fourier ',nbasis),
+                 BIC = BIC,
+                 RMSE = RMSE,
+                 SSE = SSE,
+                 gcv = gcv)
+   return(res)
+ }

# define K vector
Kseries <- seq(from = 1, to = 51, by = 2)[-1] # exclude 1
#Kseries <- seq(from = 1, to = 341, by = 2)[-1] # exclude 1
#Kseries <- seq(from = 359, to = 341, by = 2)[-1] # exclude 359
# init with K = 1
tt <- 1:nrow(daily)
range(tt)
[1]   1 730
dayindex <- seq(from = range(tt)[1], to = range(tt)[2])
dailyfourier <- generateFourierDaily(daily,1)
K= 1 ;BIC= -10.5473292661031 ;RMSE= 0.005 ;SSE= 0.019 ;GCV= 0
#dailyfourier <- generateFourierDaily(daily,359)
# calculate remaining K
for (i in Kseries) {
+   dailyfourier <- bind_rows(dailyfourier,generateFourierDaily(daily,i))
+ }
K= 3 ;BIC= -10.9087557520981 ;RMSE= 0.004 ;SSE= 0.013 ;GCV= 0
```

```
K= 5 ;BIC= -11.2584173965134 ;RMSE= 0.003 ;SSE= 0.009 ;GCV= 0
K= 7 ;BIC= -11.3581372964598 ;RMSE= 0.003 ;SSE= 0.008 ;GCV= 0
K= 9 ;BIC= -11.4736055533744 ;RMSE= 0.003 ;SSE= 0.007 ;GCV= 0
K= 11 ;BIC= -11.6096930974916 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 13 ;BIC= -11.5916299617817 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 15 ;BIC= -11.5735668260717 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 17 ;BIC= -11.5555036903617 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 19 ;BIC= -11.5374405546517 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 21 ;BIC= -11.5193774189417 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 23 ;BIC= -11.5013142832318 ;RMSE= 0.003 ;SSE= 0.006 ;GCV= 0
K= 25 ;BIC= -11.6655727043157 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 27 ;BIC= -11.6475095686058 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 29 ;BIC= -11.6294464328958 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 31 ;BIC= -11.6113832971858 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 33 ;BIC= -11.5933201614758 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 35 ;BIC= -11.5752570257658 ;RMSE= 0.003 ;SSE= 0.005 ;GCV= 0
K= 37 ;BIC= -11.5571938900559 ;RMSE= 0.002 ;SSE= 0.005 ;GCV= 0
K= 39 ;BIC= -11.7622743056601 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 41 ;BIC= -11.7442111699501 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 43 ;BIC= -11.7261480342401 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 45 ;BIC= -11.7080848985301 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 47 ;BIC= -11.6900217628202 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 49 ;BIC= -11.6719586271102 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
K= 51 ;BIC= -11.6538954914002 ;RMSE= 0.002 ;SSE= 0.004 ;GCV= 0
```

```
# add in raw data
dfraw <- tibble(dayindex = dayindex,
+                 value = daily$o31hppmavg,
+                 type = 'raw',
+                 RMSE = NA,
+                 SSE = NA,
+                 gcv = NA)
# combine with fourier transformed data
dailyfourier <- bind_rows(dfraw, dailyfourier)
# set colors
cols <- c(as.vector(inlmisc::GetColors(n = length(Kseries) + 1)), '#808080
')

# visualize
p <- ggplot(dailyfourier, aes(x = dayindex,y = value, color = type)) + geo
m_line() +
+   scale_fill_manual(values=cols) +
+   scale_color_manual(values=cols) + tidyquant::theme_tq() +
+   labs(y = expression(paste(O[3]," ", "1H(ppm)")), x = 'day index', titl
e = 'SM Dato Permaisuri Miri, CA0028 Fourier Basis Transform (K from 1 to
51)')
p

> ggplot(dailyfourier, aes(x = dayindex,y = value, color = type)) + geom_l
ine() +
```

```
+   scale_fill_manual(values=cols) +
+   scale_color_manual(values=cols) + tidyquant::theme_tq() +
+   labs(y = expression(paste(O[3],"  ", "1H(ppm)")), x = 'day index', titl
e = 'SM Dato Permaisuri Miri, CA0028 Fourier Basis Transform (K from 1 to
51)') +
+   theme(legend.position = "none")

dailyfourierSummary <- dailyfourier %>% filter(type != 'raw') %>% dplyr::s
elect(type, BIC, RMSE, SSE, gcv) %>% distinct()
dailyfourierSummary$K <- c(1,Kseries)
minBICindex <- which.min(dailyfourierSummary$BIC)
minBICvalue <- min(dailyfourierSummary$BIC)
minBIC_K <- dailyfourierSummary$K[minBICindex]

p <- ggplot(dailyfourierSummary, aes(K, BIC)) + geom_line() +
+   scale_x_continuous(expand = c(0,0), limits = c(0, max(Kseries)), break
s = seq(from = 0,to = max(Kseries),length.out = max(Kseries) + 1)) +
+   tidyquant::theme_tq() +
+   labs(y = 'BIC', x = 'Number of basis (K)', title = 'SM Dato Permaisuri
 Miri, CA0028 Optimal K based on BIC') +
+   theme(legend.position = "none") +
+   geom_hline(yintercept = minBICvalue,  linetype="dashed", color = "#FF0
000") +
+   annotate("text", x=minBIC_K-5, y= minBICvalue -0.02, label= paste0('Mi
n BIC = ',round(minBICvalue,2),'; Optimal K = ',minBIC_K), color = "#FF000
0", size = 4) +
+   theme(axis.text.x=element_text(angle=90, hjust=1))
p


# visualize with optimal K fourier
minBICindex <- which.min(dailyfourierSummary$BIC)
minBIC_K <- dailyfourierSummary$K[minBICindex]
p1 <- ggplot(dailyfourier %>% filter(type %in% c('raw','fourier 39')), aes
(x = dayindex,y = value, color = type)) + geom_line() +
+   scale_fill_manual(values=cols[21:22]) +
+   scale_color_manual(values=cols[21:22]) + tidyquant::theme_tq() +
+   labs(y = expression(paste(O[3],"  ", "1H(ppm)")), x = 'day index', x =
'Day Index', title = 'SM Dato Permaisuri Miri, CA0028 Fourier Basis Transf
orm (With Optimal K = 39)')
p1
```