

Air Pollution Index Analysis Using Functional Data Analysis

Khoo Chun Kai, Muhammad Fauzee Hamdan*

Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia

*Corresponding author: mfauzee@utm.my

Abstract

This study aims to analyze the trends and patterns of Malaysia's air pollution index in depth by applying Functional Data Analysis (FDA). Functional data analysis is a widely used and dynamic method in analyzing statistical data. In this paper, B-spline and Fourier basis functions are applied in analyzing the Air Pollution Index (API) data of Malaysia for all states in 2017. Root Mean Square Error (RMSE) is utilized to show a clearer picture in comparing both methods. The results of the functional fitting plots were plotted with the aid of R software together with computing the RMSE. The plots were then being analyzed by identifying the trend and patterns. The results show that the overall API in Malaysia tends to peak around the beginning one-third of the year 2017 then shows a clear decreasing trend. However, the state Pulau Pinang exhibits a relatively difference trend and pattern among all states. The API tends to continue rising in the first half of the year and then gradually decreases till end of the year. Overall, API Pulau Pinang achieved a higher peak compared with other states. Lastly, B-spline function is a better fitting on the data as the RMSE computed is lower.






Keywords: functional data analysis; b-spline; fourier

Introduction

This study seeks to study deep into the trends and patterns of Malaysia's air pollution index by employing Functional Data Analysis (FDA). The primary objective is to conduct a thorough examination and analysis of how the air pollution index in Malaysia fluctuates over time, with a focus on identifying and understanding the various trends and curves present within the data.

The formulation of air quality specifications for air pollutants traces back to the year 1989 when the Malaysian Department of Environment (DOE) took the initiative. These specifications, delineated in the Recommended Malaysian Air Quality Guidelines (RMG), set permissible limits for various air pollutants. The primary objective behind establishing these limits was to mitigate potential adverse effects on the health and well-being of the populace. To quantify air quality in Malaysia, the Air Pollution Index (API) emerged as a widely embraced indicator. This index not only serves as a vital tool for informing government policies but also plays a crucial role in raising public awareness regarding air quality issues. In 1993, the DOE introduced the Malaysian Air Quality Index (MAQI), a system designed to provide real-time updates on air quality status, ranging from good to emergency levels. Recognizing the effectiveness of such indices in monitoring air quality and safeguarding public health, the DOE in Malaysia introduced a modified index system in 1996. As part of this revamped system, Malaysia adopted the Air Pollution Index (API), which bears striking resemblance to the Pollutant Standard Index (PSI) utilized in the United States [1].

Table 1: API Status Indicator

Color	Descriptor	API
	Good	≤ 50
	Moderate	51 – 100
	Unhealthy	101 – 200
	Very Unhealthy	201 – 300
	Hazardous	> 300

Source: AIPMS, DOE Malaysia (2022) [2]

PSI is recognized as one of the initial synthetic indices that received approval from the United States Environmental Protection Agency (USEPA). This index was established by Ott and Hunt [3]. Yet in 1999, the USEPA made the decision to substitute the PSI with the Air Quality Index (AQI), while Malaysia opted to retain the API. The status indicator of the API was segmented into several distinct groups. The categories of air quality levels indicated in Table 1.1 include good, moderate, unhealthy, very unhealthy, and hazardous [2]. These categories are relevant for air quality management and decision-making processes related to data interpretation.

Functional data analysis (FDA) encompasses various methods adapted from multivariate analysis to handle data represented as curves rather than vectors, often derived from observations of stochastic processes over continuous time intervals. FDA extends beyond time-based scenarios, finding applications in fields like chemometrics. Researchers commonly employ B-spline and Fourier basis functions for analyzing diverse datasets. B-splines are constructed by joining polynomial segments at defined points (knots), with algorithms ensuring numerical stability. Proper selection of knots is critical to avoid overfitting or underfitting data, especially in non-penalized spline regression [4]. Fourier analysis, or harmonic analysis, decomposes series into sinusoidal components to measure variations in time series data broadly [5].

Literature Review

In [6], the study has applied Time Series Transformation method accessed in Excel utilizing the XLSTAT add-on statistical software to analyze the hourly readings air pollution index in Malaysia. The results obtained show that API value hiked greater than 500 on 23rd June 2013 due to the haze episodes in Malaysia [7].

The utilization of clustering techniques enables the detection of significant patterns and distributions within a dataset, hence offering potential insights into the fundamental structure of the data [8]. Clustering has been extensively used in atmospheric science data, particularly climate and meteorological data, for the past 50 years [9].

There are researchers that applied k-means and clustering approach to analyze the air pollution distribution in Makassar City, Indonesia [10]. The first clustering technique used was k-Means, and the results were visualized using Self-Organizing Maps (SOM).

Studies from [11] have tested the hourly and daily API data obtained and collected by 37 monitoring stations with Mann-Kendall (MK) Test and Sen's Slope Test. However, the experiments encompassed daily, monthly, and seasonal time series. Malaysia experiences two primary seasons that can impact air quality: the Southwest Monsoon (SWM) from May to August and the Northeast Monsoon (NEM) from November to February.

In [12], the study predicts air pollution data using semi-experimental regression model and Adaptive Neuro-Fuzzy Inference System (ANFIS) method. The study conducted utilized existing data on significant pollutants to forecast their future conditions by employing time-series modelling techniques. Their objective was to address this constraint of traditional time series forecasting on non-linear and complex components by enhancing the precision of daily pollutant predictions. They compare results among semi-experimental model and ANFIS and showed that ANFIS provides a more accurate forecasting results [12].

Researchers have applied FDA together with Ordinary Kriging Approach in estimating the rainfall curve. The study aims to observe and interpret the rainfall patterns that happen around the year. Functional data analysis techniques are competent to transform discrete data into a function that can effectively represent the rainfall pattern, unveiling concealed patterns within the rainfall data [13].

In previous studies, researchers have used functional data analysis to perform acoustic seascape partitioning. The study employs acoustic backscatter as a function of depth, simultaneously at three frequencies to numerically depict the vertical arrangement and constitution of sound-scattering organisms in the water column [14].

There is a study that applied functional data analysis to analyze dynamic Positron Emission Tomography (PET) data. The approach in this study, using functional data analysis, treats each

subject's IRF as the fundamental unit of analysis, models multiple subjects collectively, and estimates the IRF in a nonparametric manner [15].

In a study, functional data analysis (FDA) was used to model winter daily mean temperatures in Detroit using data from Chicago. The study hypothesized significant correlations between the temporal scales and temperatures of both cities. FDA techniques, including B-spline curve fitting with smoothing via least squares and roughness penalty, were applied to transform collected data into predictive functions. Linear models for functional responses demonstrated a significant relationship between the variables. Prediction utilized time warping functions in landmark registration for accurate results [16].

In 2021, a study by Ishmael Amartey, a comparative analysis of hierarchical clustering performance was conducted using simulated functional data with a mixed structure. Previous clustering research predominantly relied on the inverse weight technique and B-spline smoothing. This study will employ the Fourier basis smoothing method and assess clustering algorithm performance using the Rand index and adjusted Rand index [17].

Methodology

This study incorporates an analysis of the API in Malaysia for the year 2017. The API index data is sourced from the Air Quality Division of the DOE. The analysis involves examining the hourly API data from each monitoring station, with the API determined by the average concentrations of PM_{10} , O_3 , CO , SO_2 , and NO_2 . This study will be mainly analyzing the API of 13 states and 3 federal territories in Malaysia. Table 3.1 below shows the list of stations in each state capital.

Table 2: The List of Stations Located in each State Capital or City

State	State Capital / City	Location
Perlis	Kangar	Institut Latihan Perindustrian Kangar
Kedah	Alor Setar	SM Agama Kedah
Pulau Pinang	Minden	Universiti Sains Malaysia (USM)
Perak	Ipoh	Sek. Men. Keb. Jalan Tasek
W.P. Kuala Lumpur	Cheras	Sek. Men. Keb. Seri Permaisuri
W.P. Putrajaya	Putrajaya	Sek. Keb. Presint 18
Selangor	Shah Alam	Sek. Keb. TTDI Jaya
Negeri Sembilan	Seremban	Sek. Men. Teknik Tuanku Jaafar
Melaka	Bandaraya Melaka	Sekolah Tinggi Melaka
Johor	Larkin	Institut Perguruan Temenggong Ibrahim
Pahang	Kuantan	Sek. Keb. Indera Mahkota
Terengganu	Kuala Terengganu	Sek. Keb. Chabang Tiga
Kelantan	Kota Bharu	Sek. Men. Keb. Tanjong Chat
Sarawak	Kuching	Depot Ubat Kementerian Kesihatan Malaysia
Sabah	Kota Kinabalu	Sek. Men. Keb. Tansau
W.P. Labuan	Labuan	Taman Perumahan Majlis Perbandaran

Mean Imputation Method is applied to fill the unavailable or missing data of API in 2017. The mean function, denoted as

$$\mu(t) = E[y(t)] \tag{1}$$

represents the average trajectory of the API over time t . The missing data is filled with the computed $\mu(t)$ to proceed with the analysis of API.

The primary approach in FDA involves transforming the discrete observed from the data (y_i, t_i) into a functional form of $x(t_i)$ by utilizing a suitable basis function called regression modelling. The model is expressed as

$$y_i = x(t_i) + \varepsilon_i \tag{2}$$

where ε_i represents an error with a mean of zero and a finite variance, σ^2 . In this case, it is assumed that the variable t_i is observed without error so t_i is not treated as a random variable [6].

Despite potential observational errors in these discrete data, the initial stage of FDA involves converting raw data into functional objects. This transformation includes fitting a curve to the discrete observations, effectively capturing the continuous underlying process. The representation of these smooth functions is indicated by the linear combination of the following basis function

$$x(t) = \sum_{k=1}^K c_k \varphi_k(t) \tag{3}$$

where c_k are basis coefficients and φ_k is approximation of the functional form $x(t_i)$. K is the size of the maximum basis required [18].

The widely used B-splines are formed through linear combinations of spline functions characterized by a specified order and a predetermined number of breakpoints. To estimate $x(t)$, it is crucial in determining the basis functions, conduct estimations on the coefficient's values c_k by roughness penalty approach, least squared techniques, root mean square, etc. Note the $\varphi_{i,k}$ for the k^{th} order piecewise spline within the interval $[t_i, t_{i+1})$. Thus, the B-spline basis functions are built as

$$\varphi_{i,0}(t) = \begin{cases} 1, & \text{if } t_i \leq t \leq t_{i+1} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

and

$$\varphi_{i,k}(t) = \frac{t - t_i}{t_{i+k-1} - t_i} \varphi_{i,k-1}(t) + \frac{t_{i+k} - t}{t_{i+k} - t_{i+1}} \varphi_{i+k,k-1}(t) \tag{5}$$

where k indicates the order of the B-spline.

Besides B-splines, Fourier basis function is also widely used on functional data analysis. The choice of Fourier basis function in FDA is because of its high-speed computing and flexibility characteristics that fits periodic data. The Fourier basis is always composed of an odd number of functions. Thus, a Fourier basis function can be stated as below

$$x(t) = c_0 + c_1 \sin \omega t + c_2 \cos \omega t + c_3 \sin 2\omega t + c_4 \cos 2\omega t + \dots \tag{6}$$

The value of k must be selected wisely, as too many basis functions will over fit the data, while too less basis functions will be unable to determine the pattern of curves.

In this study, RMSE is utilized to determine the suitable number of basis for the basis function that fits well onto the data. Besides, it is also used to determine which type of basis function fits better on the API data. The RMSE can be stated as below

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{n}} \tag{7}$$

where \hat{y}_i are the predicted values and y_i are the observed values. The n is the number of observations.

Results and Discussion

The dataset records a total of 140160 data which is 8760 data from 13 locations. for 365 days with 1-hour interval of air pollution index recorded in 2017. In 2017, Malaysians generally breathed easy, with most states recording air quality that was either good (API ≤ 50) or moderate (API 51-100). Perlis,

Kedah, Perak, Negeri Sembilan, Pahang, Terengganu, Sarawak, and Sabah enjoyed the cleanest air, logging upwards of 7700 days each with healthy air. Pulau Pinang, with its 5815 days of moderate air, followed closely behind. But the result wasn't entirely encouraging. Selangor, with 30 days exceeding the "unhealthy" threshold (API 101-200), stood out as the air quality is worst among the states in Malaysia. Thankfully, unhealthy for sensitive groups (API 201-300) and hazardous (API > 300) air quality days presented zero in the data, offering some relief.

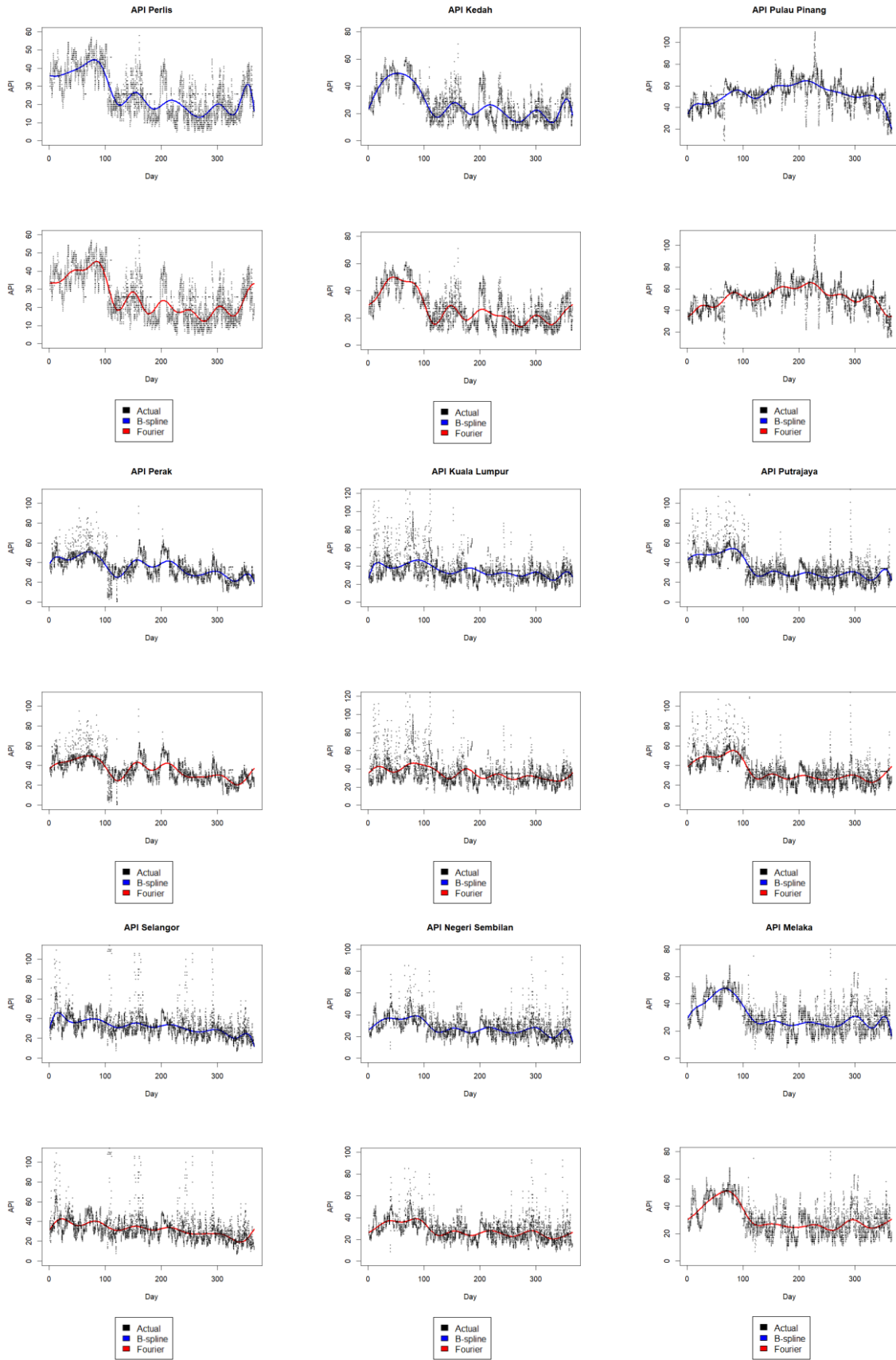
Table 3: Summary of API data in each State and Federal Territory

State	Mean	Minimum	Median	Maximum
Perlis	25.67	5	25	58
Kedah	27.07	6	24	71
Pulau Pinang	51.98	9	53	110
Perak	35.73	1	35	97
W.P. Kuala Lumpur	35	12	33	132
W.P. Putrajaya	34.34	8	33	131
Selangor	32.13	7	31	204
Negeri Sembilan	28.07	9	27	93
Melaka	31.09	7	29	122
Johor	32.5	6	32	104
Pahang	24.99	6	24.99	58
Terengganu	27.75	2	27	68
Kelantan	31.49	4	27	79
Sarawak	22.14	5	21	61
Sabah	24.89	1	24.89	58
W.P. Labuan	29.42	6	25	61

Malaysians enjoyed generally good to moderate air throughout 2017, as reflected by a mean Air Pollution Index (API) score of 32.13 across the nation. Pahang claimed its title for cleanest air, recording a minimum API of 1, while Selangor, at the other end of the spectrum, recorded the highest peak of 204. Perlis was named as the state with the healthiest lungs, averaging a low API of 25.67, while Selangor recorded the highest average API at 35.73. Besides, Selangor that hits beyond an index of 200, Pulau Pinang, W.P. Kuala Lumpur, W.P. Putrajaya, Melaka and Johor also hit a pollution record of API over 100. Overall, 2017 present a picture of generally breathable air in Malaysia, with a few localized concerns demanding attention.

To determine the number of basis, an approach of finding the errors to do comparison is conducted by calculating the RMSE. RMSE of the plot and the mean of data is calculated for 5 basis to 15 basis. This is because the number of basis determines the fitting and the lower the number of basis will gives less accurate output and may be underfitting. However, the larger the number of basis, the complexity of curve increases and it will be difficult to interpret. So considering there are 12 months in a year, the range is prior fixed within 5 to 15. The error is computed between the mean of actual values and the fitting value on the curve by varying the number of basis.

The RMSE computed resulted that number of basis should be 15 as error is smaller. B-spline and Fourier with the number of basis 15 is used as FDA approach in this paper for API analysis. The plotting of API data is then plotted with the aid of R programming using FDA. Each plot draws the scatter plot for hourly index recorded every day in 2017 and two smooth curves of functional graph by B-spline and Fourier functions that represent the API data.



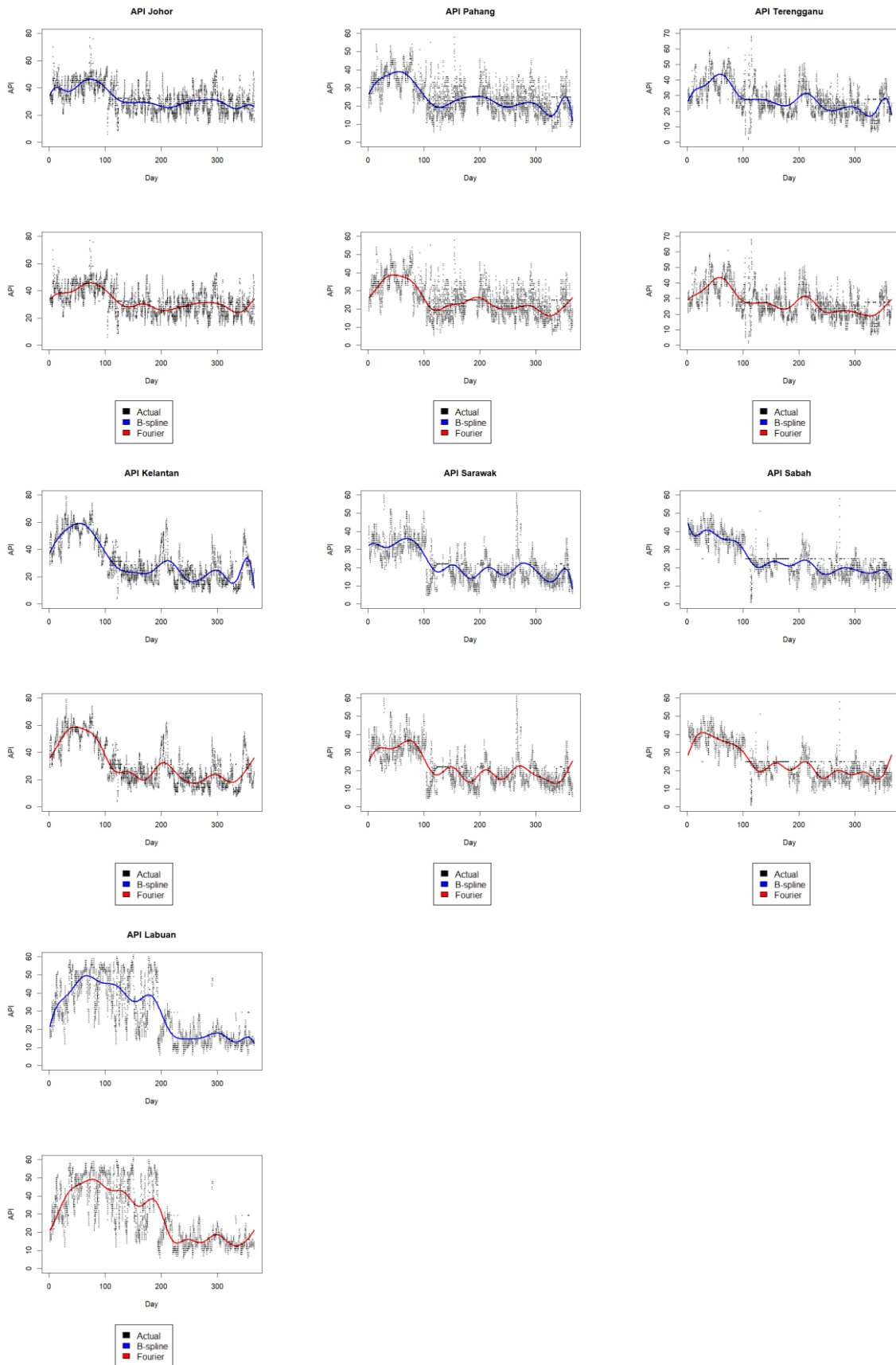


Figure 1: Functional Data Plots of API Malaysia 2017

Generally in 2017, Perlis people could breathe comfortably due to little air pollution. Pollution peaks before day 100 and then drops in the plot of API values 0–60. Kedah's air quality was similar to Perlis'. API scores range from 0 to 60, indicating healthy air. Visual indicates two peaks, one around day 50 and another near 350. At the end of the graph, API values dropped and the mean API never exceeded 60. API Pulau Pinang scored higher than the previous state. First-half year growth was remarkable. Two substantial mean peaks around day 100 and 200 require attention. Pollution gradually declined after the second peak and dropped substantially at year's conclusion. The 2017 Perak air pollution index fluctuated throughout the year. It peaked before 100 days. Two peaks occur between days 150 and 250. Then it declined gradually approaching year-end. API Kuala Lumpur functional plots usually slope steadily. However, API range exceeds 100 but the functional plot fluctuations are steady between 30 and 40. The year-round decline is small. The Putrajaya API functional plot was 40–60 in the first 100 days, then dropped to 30 for the balance of the year. The functional plot peaks before day 100, almost reaching 60 API. The scattering reveals many records over 80. The functional plot of Selangor resembles Kuala Lumpur. The scatter suggests some records exceeded 100. The plot trended downward. API Negeri Sembilan's scatter plot barely dropped below 80. Since the functional plot curve runs between 20 and 40. Melaka peaks in the first third of the year. API Melaka fluctuates significantly by year's end. The API Melaka plot ends with two lower peaks. The curves stay between 20 and 50. The API trends in Johor and Melaka are comparable. It peaked before 100 days. Unlike the API of Melaka, the trend is quieter year-round. Next is API Pahang's functional curve, which peaks at 40 and falls at 10. The peak is about day 50 and dips below 20 around day 100. The spectrum drops at year's end. The plots in API Terengganu peak after day 50 and then steadily decline over the year. The functional plot climbs somewhat at year's end and drops. The functional plot of API Kelantan displays a stronger falling. Later in the curve, API dips to 10. Lastly, the East Malaysian states. Similar diminishing curves exist in Sarawak and Sabah. Sarawak peaks at 35 before day 100, while Sabah starts at 45 and drops to 15 throughout the year. API Sarawak fluctuated till it dropped below 10. The API of Labuan has a striking functional curve that rises from day 1 to 50 around day 70. The curve then stays at 40 before plummeting to 15 till its end.

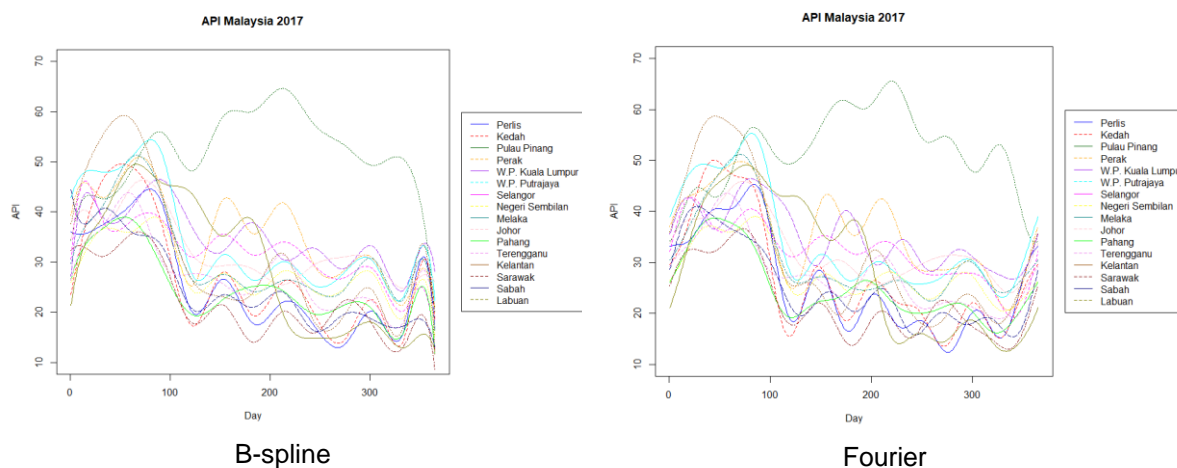


Figure 2: Summary of API Malaysia in 2017

From the visuals above, notice that most of the states undergo the same fluctuation patterns of air pollution index. They all peak in between the beginning to 100th day of the year. Generally, the air pollution index in Malaysia peaks first then drop to a lower range throughout the year except for the state Pulau Pinang. Pulau Pinang has an exceptional case in which the spectrum from the diagram shows a different pattern compared to the other states. Overall, the air pollution index in Malaysia possesses a decreasing trend throughout the year of 2017. This generally shows that the air pollution mainly happens seriously at the beginning of the year and then the situation turns better by the end of the year.

Both B-spline and Fourier provide good results in curve smoothing when fitting the API data. By comparing the results obtained, RMSE is computed to measure the accuracy for both methods. From the calculated result of RMSE, there are 13 results that show B-spline is a better option to fit the API data as they exhibit smaller RMSE. Thus, B-spline is a better function in FDA to analyze this API data.

Conclusion

This study uses Functional Data Analysis (FDA) to evaluate and monitor Malaysia's Air Pollution Index (API) to meet the first chapter's research goals. The study examines Malaysian air pollution by monitoring patterns and trends using FDA and R programming. The statistical approach to API data analysis is thought to provide policymakers and scholars with significant insights. The ultimate purpose of this research is to improve air quality solutions and promote Malaysia's sustainable growth. R program handles curve fitting and graphing. API Malaysia has comparable trends for both functionalities in 2017. Visuals suggest a steady decline with year-round swings. Pulau Pinang is the only state with a year-round high curve. After the first peak, Malaysia's air pollution index has steadily decreased. Malaysia's air pollution is usually worse in the start of the year. It becomes better all year. RMSE determines the optimal function for API Malaysia data in this research. When comparing curve errors and daily mean indexes, lesser errors indicate better fitting. R programming calculates RMSE quicker and more accurately. The 2017 API Malaysia data fits better with the B-spline basis function. This also means API data has no periodic pattern. Patterns in API data are irregular. In conclusion, the B-spline basis functional curve fits API data effectively and visualizes its trend and pattern. Functional Data Analysis (FDA) effectively analyzes as shown above. Successful discoveries might benefit society and the environment. The findings may also help improve Malaysia's air quality and avoid it.

Acknowledgement

The researcher would like to convey my sincere appreciation to my supervisor, Dr Muhammad Fauzee Bin Hamdan, for her generous guidance and patience in supervising me during this research. In addition, he has shared a significant amount of his expertise and provided me with support in finishing this project. I am deeply appreciative of his guidance. I express my sincere gratitude to my parents, who provide me with both emotional and financial assistance. They consistently have faith in me and provide support when I encounter difficulties. They are the reason for my current identity. Additionally, I would like to express my gratitude to my friends who have provided me with encouragement and motivation during this process. They are supportive and willing to lend a hand to me when I faced barriers or challenges in completing this project. I deeply value their presence and support.

References

- [1] DEPARTMENT OF ENVIRONMENT MALAYSIA (DOE). A guide to air pollutant index (API) in Malaysia. Department of Environment. Kuala Lumpur, Malaysia, 1. 2000.
- [2] DEPARTMENT OF ENVIRONMENT MALAYSIA (DOE). APIMS. Wilayah Persekutuan Putrajaya, Malaysia. 2022. Retrieved from: <https://apims.doe.gov.my/home.html>
- [3] Ott WR, Hunt Jr WF. A quantitative evaluation of the pollutant standards index. *Journal of the Air Pollution Control Association*. 1976 Nov 1;26(11):1050-4.
- [4] Aguilera AM, Aguilera-Morillo MC. Comparative study of different B-spline approaches for functional data. *Mathematical and Computer Modelling*. 2013 Oct 1;58(7-8):1568-79.
- [5] Bloomfield P. *Fourier analysis of time series: an introduction*. John Wiley & Sons; 2004 Apr 5.
- [6] Jain AK, Murty MN, Flynn PJ. Data clustering: a review. *ACM computing surveys (CSUR)*. 1999 Sep 1;31(3):264-323.
- [7] Haliza AR. Haze phenomenon in Malaysia: domestic or transboundary factor. In 3rd international journal conference on chemical engineering and its applications 2013 (pp. 590-599).
- [8] Halkidi M, Batistakis Y, Vazirgiannis M. On clustering validation techniques. *Journal of intelligent information systems*. 2001 Dec;17:107-45.

- [9] Govender P, Sivakumar V. Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric pollution research*. 2020 Jan 1;11(1):40-56.
- [10] Annas S, Uca U, Irwan I, Safei RH, Rais Z. Using k-Means and Self Organizing Maps in Clustering Air Pollution Distribution in Makassar City, Indonesia. *Jambura Journal of Mathematics*. 2022;4(1):167-76.
- [11] Alyousifi Y, Ibrahim K, Zin WZ, Rathnayake U. Trend analysis and change point detection of air pollution index in Malaysia. *International Journal of Environmental Science and Technology*. 2021:1-22.
- [12] Zeinalnezhad M, Chofreh AG, Goni FA, Klemeš JJ. Air pollution prediction using semi-experimental regression model and Adaptive Neuro-Fuzzy Inference System. *Journal of Cleaner Production*. 2020 Jul 10;261:121218.
- [13] Hamdan MF, Jemain AA, Jamaludin SS. Estimation of Rainfall Curve by using Functional Data Analysis and Ordinary Kriging Approach. *MATEMATIKA: Malaysian Journal of Industrial and Applied Mathematics*. 2018 Dec 31:167-77.
- [14] Ariza A, Lebourges-Dhaussy A, Nerini D, Pauthenet E, Roudaut G, Assunção R, Tosetto E, Bertrand A. Acoustic seascape partitioning through functional data analysis. *Journal of Biogeography*. 2023 Sep;50(9):1546-60.
- [15] Chen Y, Goldsmith J, Ogden RT. Functional data analysis of dynamic PET data. *Journal of the American Statistical Association*. 2018 Oct 26.
- [16] Wu Y. Predictions of Temperatures in Winter Months in Detroit based on Temperatures in a nearby city Chicago: An FDA Regression (Doctoral dissertation). 2016.
- [17] Amartey I. Functional Mixed Data Clustering with Fourier Basis Smoothing (Master's thesis, East Tennessee State University). 2021 Dec.
- [18] Carey M, Gath EG, Hayes K. Generalised smoothing in functional data analysis (Doctoral dissertation, Ph. D. Thesis, University).