Archives of Environmental Protection
Vol. 50 no. 2 pp. 65–74

PL ISSN 2083-4772
DOI.10.24425/aep.2024.150553

Predicting air quality trends in Malaysia's largest cities: the role of urban population dynamics and COVID-19 effects

Wong Ming Wong[1], Shian-Yang Tzeng[2], Hao-Fan Mo[3], Wunhong Su[4]*

[1]International College, Krirk University, Thailand
[2]School of Economics and Management, Quanzhou University of Information Engineering, China
[3]JinWen University of Science and Technology, Taiwan
[4]School of Accounting, Hangzhou Dianzi University, China

* Corresponding author's e-mail: whsu@hdu.edu.cn

Keywords: Air quality index, COVID-19, Malaysia, Cluster Analysis, Generalized Additive Mixed Model

Abstract: This paper aims to explore the relationship between the Air Quality Index (AQI), COVID-19 incidence rates, and population density within Malaysia's ten most populous cities from January 2018 to December 2021. Data were sourced from the Department of Statistics Malaysia, the World Air Quality Index Project, and Our World in Statistics. The methodology integrated population-based city classification and AQI assessment, cluster analysis through SPSS, and Generalized Additive Mixed Model (GAMM) analysis using R Studio despite encountering a data gap in AQI for five months in 2019. Cities were organized into three clusters based on their AQI: Cluster One included Ipoh, Penang, Kuala Lumpur, and Melaka, Cluster Two comprised Kuantan, Seremban, Johor Bahru, and Kota Bharu, Cluster Three featured Kota Kinabalu and Kuching. GAMM analysis revealed prediction accuracies for AQI variations of 58%, 60%, and 41% for the respective clusters, indicating a notable impact of population density on air quality. AQI variations remained unaffected by COVID-19, with a forecasted improvement in air quality across all clusters. The paper presents novel insights into the negligible impact of COVID-19 on AQI variations and underscores the predictive power of population dynamics on urban air quality, offering valuable perspectives for environmental and urban planning.

Introduction

SARS-CoV-2, a novel coronavirus termed COVID-19, was first discovered in Wuhan, China, in December 2019 and was identified as the pandemic's causal virus by the World Health Organization on January 20, 20202 (Tang et al. 2020, Yang et al. 2020, Zhang & Ma, 2020). The global expansion of COVID-19 impacts customers' beliefs, attitudes, intentions, actions, and health (Li et al. 2020, Wang et al. 2020). Specifically, during the lockdown period, consumers make more purchases online or while working from home (Chenarides et al. 2021, Li et al. 2020).

COVID-19 studies are currently categorized into three themes. The first theme explores the impact of COVID-19 on consumer behavior, including purchase decisions (Li et al. 2020, Wang et al. 2020). The second theme is public health-related, examining issues such as social isolation and diseases associated with the COVID-19 pandemic (Barouki et al. 2021, Liao et al. 2020, Wong et al. 2023). Finally, the third theme concerns COVID-19-related issues such as air pollution or COVID-19 vaccination and their impact on daily confirmed COVID-19 cases or death cases (Lim et al. 2021, Meo et al. 2021, Wetchayont, 2021). Consequently, this paper focuses on the third theme, which examines the association between AQI variation in Malaysian cities before and after the COVID-19 period.

Previous research has examined the elements of air pollution, such as PM2.5 and CO, in cities during the COVID-19 pandemic or before and after it, as well as the association between COVID-19 and meteorological characteristics, such as climate, air pollution, and the environment (Liu et al. 2021, Valdés Salgado et al. 2021). For instance, Liu et al. (2021) analyzed the effects of air pollution elements (PM2.5, Pm10, SO_2, CO, NO_2, and O_3) on COVID-19 in China, Japan, Korea, Canada, the United States of America, Russia, the United Kingdom, Germany, and France. Their findings indicate that increased levels of air pollution, when combined with levels of air pollution, correlate with an increase in newly confirmed COVID-19 cases. In Italy, PM2.5 negatively affects new COVID-19 cases (Kotsiou et al. 2021). However, the death rate of COVID-19 does not correlate with PM2.5 in Chile (Valdés Salgado et al. 2021).

On the other hand, numerous studies have compared air pollution levels before and during COVID-19 lockdowns. These studies show that during the COVID-19 shutdowns, air pollution decreased significantly, particularly in transportation-related pollution and business and industrial activities, notably PM2.5 (Gkatzelis et al. 2021, Kaewrat and Janta, 2021, Lee and Finerman, 2021, Wetchayont, 2021).

During the COVID-19 pandemic, South Korea experienced a reduction in air pollution emissions by 10% to 20%. Notably, a mere 1% reduction in commute flows led to a decrease in air pollutants by approximately 0.08 to 0.17%, including PM2.5, PM10, NO_2, CO, and SO_2 (Lee and Finerman, 2021). In Greater Bangkok, Thailand, the effect of the COVID-19 lockdown on air pollution was examined across three distinct periods: before, during, and after the lockdown. The restrictions imposed during the lockdown, which affected traffic, commercial, and industrial activities, resulted in a noticeable reduction in anthropogenic emissions and subsequently improved air quality, particularly in PM2.5, PM10, O_3, and CO levels (Wetchayont, 2021). Furthermore, Kaewrat and Janta (2021) found that air quality significantly improved during the COVID-19 lockdown, with CO and NO_2 transportation emissions showing reductions ranging from 20% to 50% across all of Thailand's metropolitan, industrial, and suburban cities.

However, the scope of these studies is limited due to (1) the lack of integration of a single index of AQI variation and trends in a single country, (2) the short data collection period, typically less than 36 months, and (3) the association between AQI and COVID-19 in a single country. Therefore, this paper aims to address these gaps by investigating the association between cities, COVID-19 cases, time series data, and AQI in Malaysia.

The research question of this paper is to examine the association between: (1) the top ten Malaysian population cities, (2) the presence or absence of COVID-19, (3) the time series covering the period from January 2018 to December 2021, and (4) AQI variation. To address this research gap, this paper utilizes Generalized Additive Mixed Models (GAMM) (Augustin et al. 2009, Chen, 2000, Wood, 2006, 2011).

First of all, this paper focuses on Malaysia due to ongoing air pollution issues. The primary cause of Malaysia's air pollution is haze resulting from farmers burning forest to make way for palm oil plantations, exacerbated by winds from Indonesia. For instance, the 2015 haze incident in Malaysia (Jenkins, 2015).

Secondly, the air quality index (AQI) is a nonlinear, dimensionless assessment of air quality conditions affecting a city's health and environment (Plaia and Ruggieri, 2011). Additionally, the AQI is available to the public to determine the air pollution level in a given city (Chaudhuri and Chowdhury, 2018, Li et al. 2015). Thus, this paper chooses a single index to combine the AQIs of all cities to describe one country's AQI variation.

Numerous countries utilize various components of AQI indicators to assess air quality or pollution. For example, Canada's Air quality health index is determined using PM2.5, O_3, and NO_2 (Environment and Climate Change Canada, 2021). In Malaysia, the government names AQI the Air Pollutant Index (API), which indicates the air quality status in any particular area. The API value is calculated based on the average concentration of air pollutants, namely PM2.5, PM10, SO_2, NO_2, CO, and O_3. Therefore, the air pollutant with the highest concentration (dominant pollutant) will determine the API value. Usually, the concentration of particulate matter, PM2.5, is the highest among other pollutants and determines the API value (Department of Environment, 2013).

Thirdly, the paper utilizes secondary data, which often presents challenges related to (1) nonlinearity, (2) normality, (3) independence of errors, and (4) missing values. Consequently, the analysis in this paper employs data mining technology, specifically GAMM. Data mining technology can extract crucial information and aid in decision-making through various applications such as categorization, visualization, predictive modeling, correlation analysis, bias detection, relational modeling, and data overview (Hormozi and Giles, 2004).

Fourthly, GAMM represents a machine learning approach used in data mining for comparing time series analysis methods, such as ARIMA models. GAMM offers advantages over traditional methods by combining the strengths of Generalized Linear Model (GLM) and Generalized Additive Model (GAM). Specifically, GAMM treats variables added to the GAM as smooth curves, while incorporating mixed model elements as random effects (Constantinescu, 2019).

Table 1. Malaysia's Top Ten Cities.

	City	Abbreviation of City Name	States	The sum of the population ('000)
1	Kuala Lumpur	KL	Kuala Lumpur, the capital city of Malaysia	1790.1
2	Johor Bahru	JB	Johor	1579.5
3	Ipoh	Ipoh	Perak	829.7
4	Kuching	Kuching	Sarawak	693.8
5	Seremban	Seremban	Negeri Sembilan	619.1
6	Kota Bharu	KB	Kelantan	596.9
7	Timur Laut	Penang	Pulau Pinang	577.9
8	Melaka	Melaka	Melaka	571.3
9	Kota Kinabalu	KK	Sabah	563.3
10	Kuantan	Kuantan	Pahang	522.3

Table 2. Descriptive Statistics of AQI in Ten Malaysian Cities.

City	Minimum	Maximum	Mean	Std. Deviation
Ipoh	22.520	90.700	50.042	11.898
JB	16.480	78.770	46.484	12.185
KB	22.480	65.630	45.995	10.771
KK	15.610	62.430	32.103	10.036
KL	29.450	112.270	54.208	13.308
Kuantan	21.230	77.900	43.117	10.847
Kuching	10.550	135.230	37.991	19.382
Melaka	22.420	97.230	50.694	13.827
Penang	21.840	76.470	50.611	12.254
Seremban	19.360	99.100	45.194	13.261

N=43

Terms

COVID-19

COVID-19 represents newly confirmed cases due to COVID-19 infection, where 0 indicates no cases and 1 indicates the presence of cases, essentially functioning as a dummy variable.

Time Series (STOL)

The STOL refers to an ordinal variable representing monthly time series data spanning from January 2018 to December 2021.

Materials and Methods

Research Design

Four phases were proposed in this paper. Firstly, to determine Malaysia's top ten cities based on their population. Secondly, these ten cities were analyzed using cluster analysis by SPSS. Thirdly, each cluster's cities have been analyzed by GAMM via R Studio. As a result, the frequency of GAMM analysis is proportional to the number of clusters.

Data Extraction

This paper utilized three datasets: (1) cities' population in Malaysia accessed from the Department of Statistics Malaysia Office portal under the title "Population by State, Administrative District, and Sex, 2016 – 2018" on April 3, 2022 (Department of Statistics Malaysia Official Portal, 2020), (2) AQI dataset for each city accessed from the Air Quality Open Data Platform on April 8, 2022 (The World Air Quality Index Project, 2022), and (3) COVID-19 dataset from the "Our World in Statistics" website on April 5, 2022 (Mathieu et al. 2021). The sample size includes 48 months of AQI and COVID-19 data for these ten cities, with five months of missing data for each city's AQI from January to May 2019.

Process

This paper's two-step analysis utilized three datasets from three independent sources. The first step involves modifying all relevant data to create research variables. Initially, the paper compiles monthly average data by calculating the daily AQI for each city. Secondly, the paper employs a dummy variable to indicate the presence or absence of COVID-19 cases in a given month, denoted as COVID-19 (where 0 indicates no COVID-19 cases and 1 indicates the presence of COVID-19 cases). Thirdly, the paper defines the monthly time series as the ordinal variable "STOL" from January 2018 to December 2021. This variable assigns a numerical value starting from one according to the chronological order of the year and month. Fourthly, the paper identifies Malaysia's ten most populous cities for cluster analysis.

In the second step, the paper adopted cluster analysis based on each city's AOI. Initially, the cluster analysis identified a group of cities with comparable AQI, from which these cities were chosen. After that, the paper examined the associations among research variables, which included (1) the ten cities, (2) AQI for each city, (3) COVID-19, and (4) STOL via GAMM analysis.

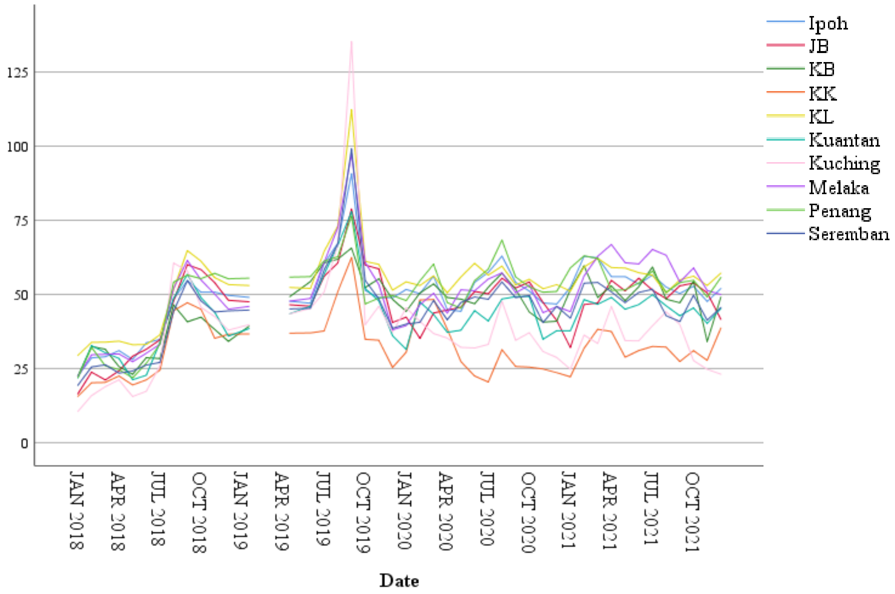

Figure 1. The AQI for Each City from January 2018 to December 2021.

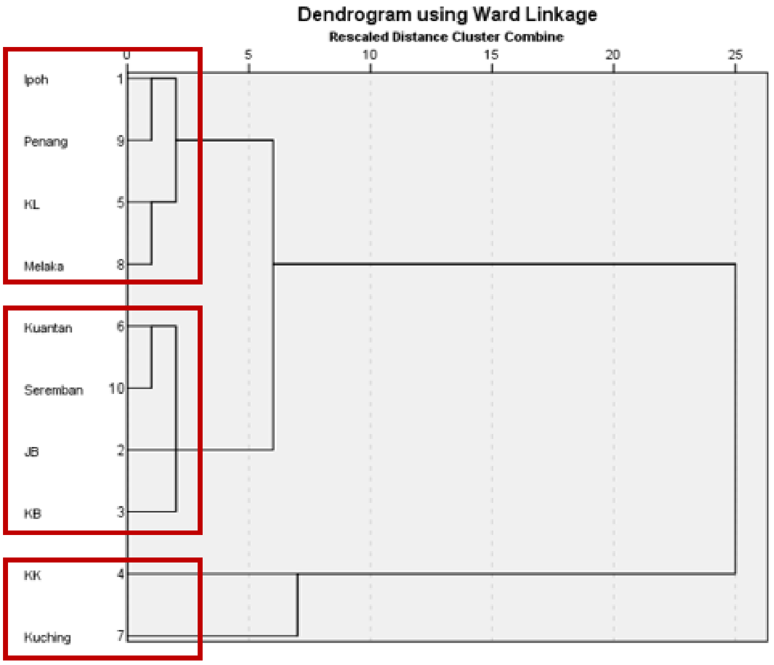

Figure 2. Three Malaysian City Clusters.

Data Analysis

Descriptive Statistics of Top Ten Population Cities in Malaysia

Malaysia's top ten cities with the highest population are summarized in Table 1. Additionally, Table 2 presents the monthly AQI for each city. However, due to missing data, the AQI data only covers 43 months. Therefore, the AQI for each city is shown in Figure 1 from January 2018 to December 2021. As illustrated in Figure 1, the lines representing the AQI have been interrupted due to missing data between January and May 2019.

Cluster Analysis

The initial phase of this paper involved conducting a hierarchical cluster analysis through SPSS, categorizing ten Malaysian cities based on their AQI data spanning from January 2018 to December 2022. The analysis resulted in dividing these cities into three clusters, as outlined in Figure 2. Figure 3 visually represents the classification, organizing Malaysia's ten most populous cities into three clusters based on population density. Specifically, Cluster One is depicted in dark red and consists of Ipoh, Penang, KL, and Melaka. Cluster Two, depicted in deep brown, comprises Kuantan, Seremban, JB, and KB. Cluster Three, shown in dark blue, comprises KK and Kuching. This structured approach to clustering underscores the study's methodical examination of urban AQI variations within Malaysia's diverse urban landscapes.

Descriptive Statistics of Three Clusters' AQI

Each cluster of AQI is depicted in Figures 4, 5, and 6. These lines exhibit interruptions attributed to missing data between January and May 2019. Figure 4 describes the AQI of Cluster

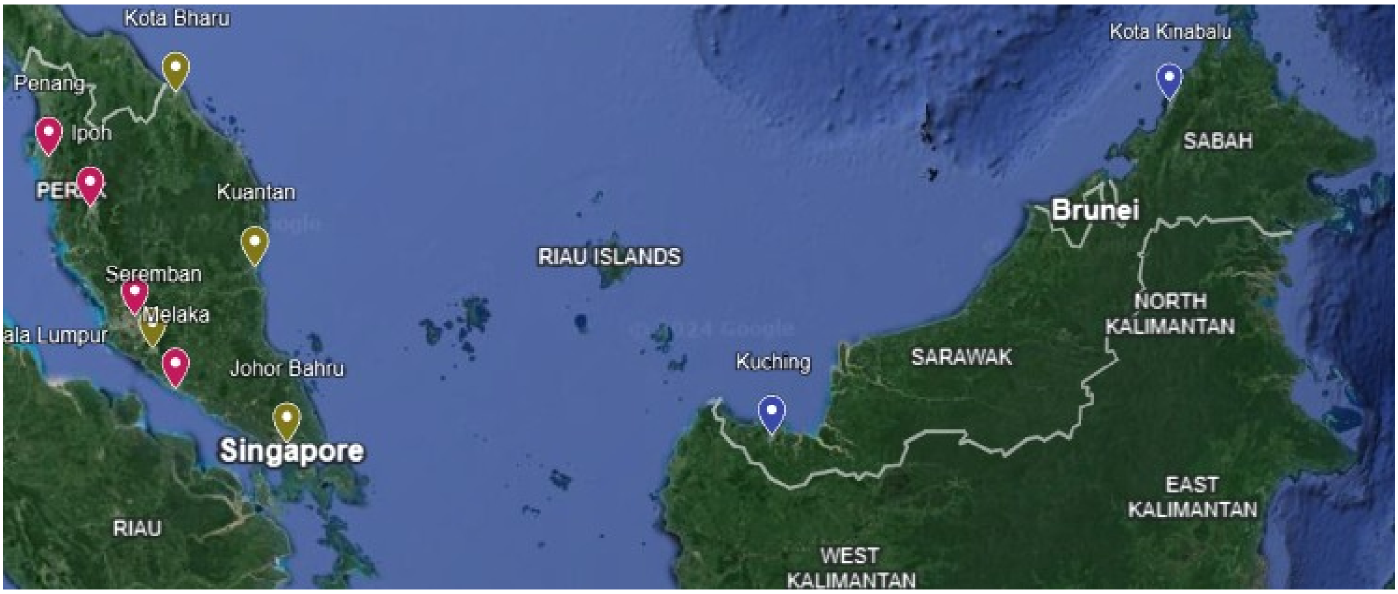

Figure 3. Three Clusters of Malaysia Cities' Locations.

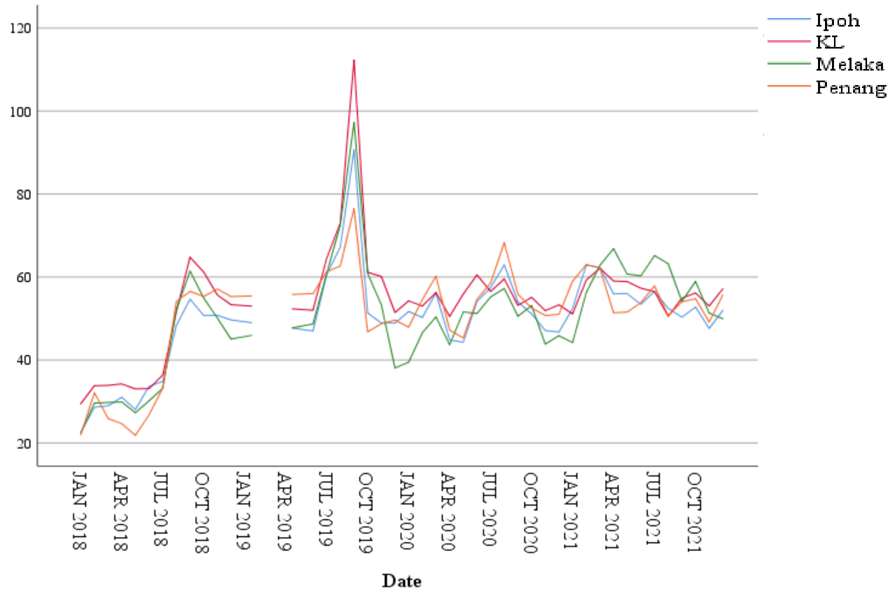

Figure 4. AQI of Cluster One.

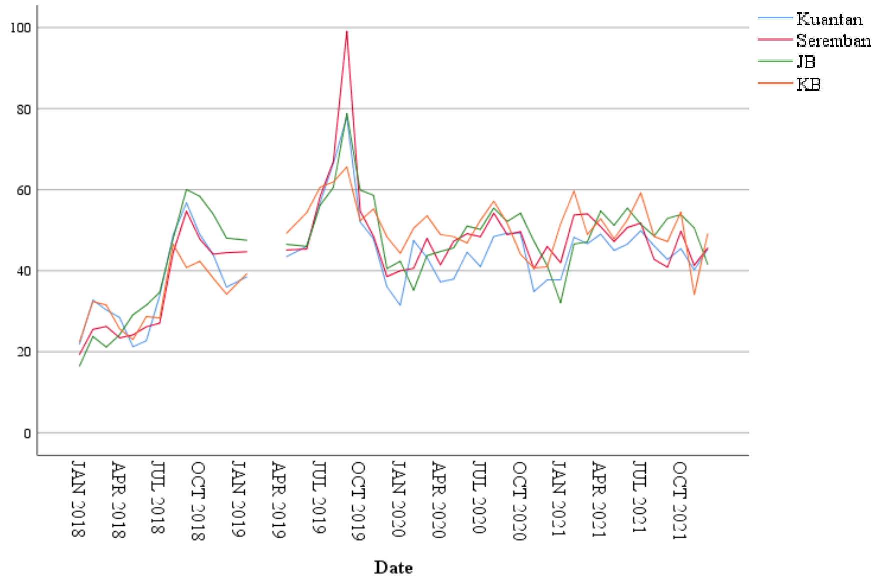

Figure 5. AQI of Cluster Two.

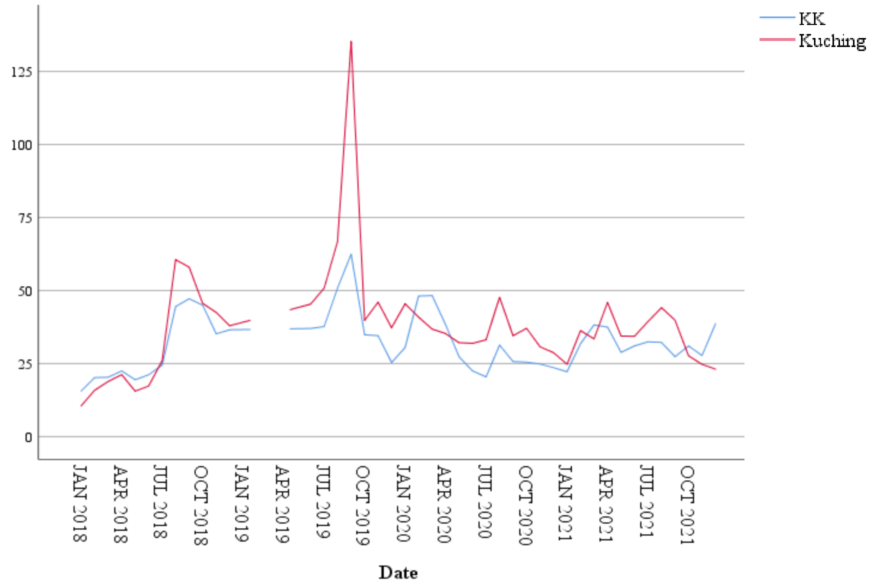

Figure 6. AQI of Cluster Three.

One. Figure 5 presents Cluster Two, which comprises the AIP of Kuantan, Seremban, JB, and KB. Finally, the AQI of KK and Kuching is placed in Cluster Three, as depicted in Figure 6. These figures illustrate the AQI variations of each city, showing a nearly uniform trend, particularly in the rising and falling patterns.

Generalized Additive Mixed Model Analysis

In this paper, the GAMM package (version 0.2-6) within R Studio was employed to conduct GAMM analysis, focusing on the three clusters identified earlier. GAMM emerges as a comprehensive statistical framework that enhances the capabilities of GLM and GAM by integrating both fixed and random effects (Hormozi and Giles, 2004). This integration proves instrumental in analyzing datasets characterized by nonlinear dependencies between variables and the presence of correlated observations - phenomena often encountered in longitudinal or hierarchical data settings (Wood, 2017). By leveraging smooth functions, GAMM adeptly models the complexity inherent in these relationships, thereby circumventing the necessity for a predefined relationship model and facilitating the identification of intricate data patterns (Hastie and Tibshirani, 1990).

A notable feature of GAMM is its provision for random effects, which are crucial for analyzing grouped or hierarchical data structures where observations within a group are correlated. This capability is particularly valuable in longitudinal studies or nested datasets, as it allows for the accommodation of variability within and across groups, ensuring a nuanced analysis (Pinheiro and Bates, 2009). Moreover, the versatility of GAMM extends to accommodating a wide range of response variable types, including continuous, binary, count, and time-to-event data. This versatility renders GAMM applicable across a broad spectrum of research domains.

Several compelling attributes justify the adoption of GAMM in data analysis. Primarily, the model's flexibility in addressing nonlinear trends enables the exploration of complex data structures without predetermining the relationship form. The model's adaptability and ability to handle correlations enhance estimation accuracy and support informed decisions. Furthermore, the adaptability of GAMM to diverse data types solidifies their position as an indispensable tool in statistical analysis, particularly for researchers navigating complex datasets (Zuur et al. 2009).

In summary, GAMM stands as an essential instrument within the statistical analysis toolkit, offering a detailed exploration of complex, nonlinear relationships in hierarchical or longitudinal data (Hastie and Tibshirani, 1990, Pinheiro and Bates, 2009, Wood, 2017, Zuur et al. 2009). Its capacity to handle various data types and structures while accurately addressing correlation and variability establishes GAMM as a pivotal methodology for researchers seeking in-depth insights from their data analyses (Hastie and Tibshirani, 1990, Pinheiro and Bates, 2009, Wood, 2017, Zuur et al. 2009).

Table 3. GAMM's Fixed-Effect Model on Cluster One's Cities.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	52.9322	2.8629	18.489	<2e-16***
KL	4.1656	1.7946	2.321	0.0215*
Melaka	0.652	1.7946	0.363	0.7169
Penang	0.5694	1.7946	0.317	0.7514
COVID19	-5.1785	4.5979	-1.126	0.2617

Note: *** p< 0.001, ** p<0.01, * p< 0.05

Table 4. GAMM's Smoothing Fixed-Effect Model on STOL.

Variable	edf	Ref. df	F	p-value
s(STOL)	6.171	7.356	28.490	<2e-16***

Note: *** p< 0.001, ** p<0.01, * p< 0.05
Adj R2 = 0.580, Deviance Explained = 60.5%
GCV = 74.055, Scale est. = 69.245, n = 172

GAMM Analysis of Cluster One

Cluster One includes Ipoh, Penang, KL, and Melaka, with Ipoh serving as the analysis equation's benchmark (1). GAMM generated two reports to assess the effects of smoothing and non-smoothing on the three variables under consideration. Table 3 ouylines the fixed-effect impact of GAMM on the significant relationship between KL, Ipoh, and the AQI, while Table 4 highlights the significant influence of the STOL on AQI.

Additionally, the model employs two metrics to elucidate its explanatory power: the adjusted R^2 and the deviance explained in Table 4. The adjusted R^2 value, standing at 0.580, signifies that the monthly time series data of KL and Ipoh can account for 58% of the variability in AQI. On the other hand, the deviance explained, which amounts to 60.5%, demonstrates the proportion of AQI variation elucidated within this research model.

$$\text{AQI} \sim \text{City}+s(\text{STOL})+\text{COVID19} \qquad (1)$$

Figure 7 presents the relationship between AQI and STOL within KL and Ipoh, showcasing six inflection points that indicate fluctuations in AQI, whether increasing or decreasing. The depicted AQI trends in Figure 7 show a subtle yet consistent reduction in the AQI levels for both cities. Specifically, the figure marks two notable peaks in the expected value, *E(X)*, along a red line within a zone, signifying a smoothed trend derived from STOL (which encompasses monthly time series data from January 2018 to December 2021). These peaks appear to maintain a consistent level within the figure.

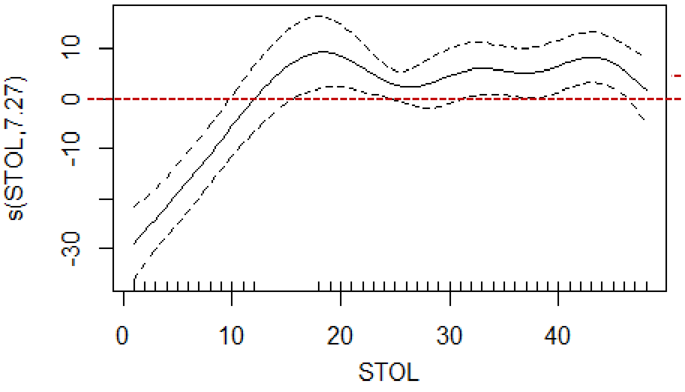

Figure 7. Smoothing Fixed-Effect Model of KL and Ipoh on AQI variation.

Table 5. GAMM's Fixed-Effect Model on Cluster One's Cities

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	50.9877	2.8641	17.803	<2e-16***
KB	-0.4897	1.6627	-0.295	0.7687
Kuantan	-3.3671	1.6627	-2.025	0.0445*
Seremban	-1.2905	1.6627	-0.776	0.4388
COVID19	-8.0689	4.6792	-1.724	0.0866

Note: *** p< 0.001; ** p<0.01; * p< 0.05

Table 6. GAMM's Smoothing Fixed-Effect Model on STOL

Variable	edf	Ref. df	F	p-value
s(STOL)	7.269	8.291	26.64	<2e-16***

Note: *** p< 0.001; ** p<0.01; * p< 0.05
R-sq.(adj) = 0.572 Deviance explained = 60%
GCV = 64.001 Scale est. = 59.436 n = 172

Furthermore, the second peak suggests a slight improvement in AQI, indicating marginally better air quality compared to earlier observations.

GAMM Analysis of Cluster Two

In the configuration for Cluster Two, JB serves as the benchmark, encompassing Kuantan, Seremban, JB, and KB. As per equation (2), Table 5 presents the fixed-effect influence of GAMM on the significant correlation between JB, Kuantan, and AQI. Table 6 displays the significant impact of the STOL variable on AQI. The analytical model is elucidated through two key metrics: the adjusted R^2 and the deviance explained in Table 6. An adjusted R^2 value of 0.572 indicates that the model successfully predicts 58% of the AQI variation based on monthly time series data of Kuantan and JB. Additionally, the model's explained deviance, reported at 60%, quantifies the proportion of variance in AQI captured by this specific study model.

$$\text{AQI} \sim \text{City}+s(\text{STOL})+\text{COVID19} \quad (2)$$

Equation Formula: AQI ~ City+s(STOL)+COVID19 (2)

Figure 8 illustrates the relationship between JB, Kuantan, and AQI, highlighting discernible inflection points in AQI alongside variations influenced by the STOL variable, encompassing monthly time series data from January 2018 to December 2021. Forecasts indicate an imminent decline in AQI variation for JB and Kuantan. However, the anticipated value, denoted as *E(X)*, exhibits three prominent peaks represented by the fitted red line within a zone in Figure 7. As illustrated, the third peak in the expected value matches the previous levels, indicating stable AQI and suggesting the maintenance of consistent air quality standards.

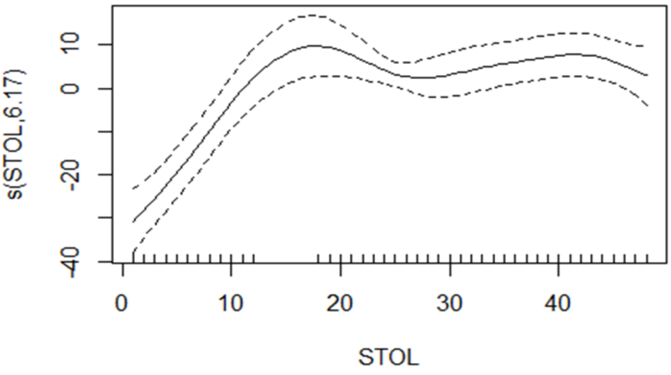

Figure 8. Smoothing Fixed-Effect Model of JB and Kuantan on AQI variation.

Table 7. GAMM's Fixed-Effect Model on Cluster One's Cities.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	35.119	4.931	7.122	0.000***
Kuching	5.888	2.693	-0.295	0.0318*
COVID19	-5.403	8.149	-1.724	0.5092

Note: *** p< 0.001, ** p<0.01, * p< 0.05

Table 8. GAMM's Smoothing Fixed-Effect Model on STOL.

Variable	edf	Ref. df	F	p-value
s(STOL)	4.46	5.506	7.964	0.000***

Note: *** p< 0.001, ** p<0.01, * p< 0.05
R-sq.(adj) = 0.361 Deviance explained = 41%
GCV = 170.79 Scale est. = 155.98 n = 86

GAMM Analysis of Cluster Three

The analysis of Cluster Three, comprising KK and Kuching, with KK serving as the benchmark in equation (3), reveals significant findings regarding the relationship between KK, Kuching, and AQI. As detailed in Table 7, GAMM's fixed effect demonstrates a significant association between KK, Kuching, and AQI. Further analysis, as presented in Table 8, indicates that STOL significantly impacts AQI.

The model's efficacy in explaining AQI variation is quantified through two metrics: adjusted R-squared and deviance explained, as shown in Table 8. The adjusted R^2 value of 0.361 indicates that the model accounts for 36% of the variance in AQI across STOL (the monthly time series)

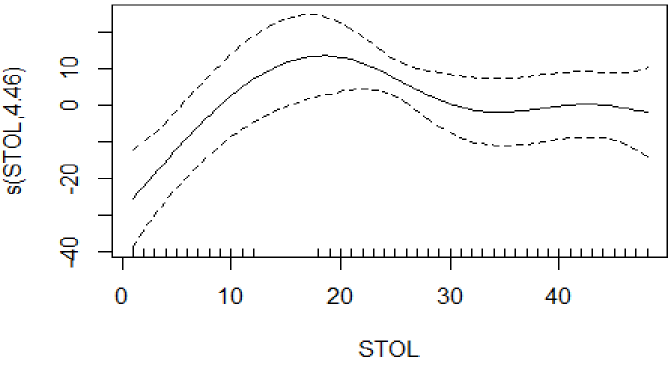

Figure 9. Smoothing Fixed-Effect Model of KK and Kuching on AQI variation

within this cluster. Moreover, the model's deviance explained stands at 41%, further affirming its capability to elucidate the variance in AQI within the context of this paper. These figures are instrumental in demonstrating the model's performance in predicting AQI variation in KK and Kuching as part of Cluster Three.

$$\text{AQI} \sim \text{City}+\text{s(STOL)}+\text{COVID19} \qquad (3)$$

Figure 9 showcases the relationship between the AQI in KK, Kuching, and the STOL, identifying four critical inflection points that demonstrate fluctuations in AQI, with directions either ascending or descending. This visualization suggests a gentle downward trend in the variability of AQI within KK and Kuching, hinting at a gradual improvement in air quality.

Moreover, the expected value, *E(X)*, is highlighted by two pronounced peaks along the highlighted red line within a zone in Figure 9. It is of particular interest that the second peak is observed to be lower than the initial peak, suggesting a decrease in the expected AQI value. Such a finding, as illustrated in Figure 9, serves as evidence of generally improved air quality across Cluster Three.

Conclusion and Discussion

This paper adopts data mining analysis to analyze gathered data from various sources regarding AQI variation in a few cities in one country. Similar studies have limitations, such as the absence of multiple cities in a single country and more than a 36-month data collection period. On the other hand, because the data are derived from secondary sources, data are frequently (1) non-parametric, (2) nonlinear, and (3) missing (Hastie and Tibshirani, 1990, Pinheiro and Bates, 2009, Wood, 2017, Zuur et al. 2009). As a result, the paper utilized cluster analysis and GAMM analysis to address the research gap.

This paper utilized cluster analysis to categorize Malaysia's ten largest cities by population into three clasters. Cluster one comprises Ipoh, Penang, KL, and Melaka. Cluster two includes Kuantan, Seremban, JB, and KB. KK and Kuching are included in cluster three. Additionally, the paper utilized GAMM analysis, with the number of analyses proportional to the number of clusters.

The result of cluster one indicate that the GAMM analysis reveals a significant association between KL, Ipoh, monthly time series, and AQI. Moreover, the model predicts that 58% of AQI variation can be accounted for by the monthly time series of KL and Ipoh. In cluster two, it is concluded that the GAMM analysis shows a significant association between JB, Kuantan, monthly time series, and AQI. Furthermore, the model predicts 60% of AQI variation using monthly time series data from JB and Kuantan. Finally, in cluster three, the GAMM analysis demonstrates a significant association between KK, Kuching, and AQI. The model predicts 41% of AQI variation based on the monthly time series of KK and Kuching.

Discussion

This paper uncovers three noteworthy findings. Firstly, the research suggests that the AQI variation in Malaysian cities is not significantly influenced by their geographic locations, Instead, the analysis focused on the AQI data for each city to form clusters. Despite this, it is interesting to note that the three clusters align with distinct geographical regions. Cities in cluster one, including Ipoh, Penang, KL, and Melaka, are situated along the western coast of the Malay Peninsula. Cluster two comprises two cities, Kuantan and KB, located on the east coast of the Malay Peninsula, along with Seremban and JB, located on the Malay Peninsula's west coast. Finally, KK and Kuching from cluster three are found on Borneo Island in East Malaysia.

Secondly, to compare time series and GAMM analysis, the time series analysis provides a detailed examination of AQI variation for each city, as illustrated in Figures 4, 5, and 6. In addition, Figure 1 offers an overview of AQI variation across all ten cities. However, a current limitation is that time series analysis cannot effectively integrate all data into a single series that accurately describes the variation. Conversely, GAMM analysis addresses this issue by handling missing data and assuming AQI variation for each cluster. Figures 7, 8, and 9 illustrate the AQI variation for the three clusters. Using existing data from January 2018 to December 2021, GAMM is used to forecast future AQI variation within each cluster.

Thirdly, utilizing GAMM analysis, the expected value E(X) represented by the flitted line in Figures 7, 8, and 9 predicts the future air quality for each cluster. Cluster one shows a slight improvement in air quality. Cluster Two appears to have maintained its previous air quality level. In contrast, the overall air quality for Cluster Three is generally better than before.

In conclusion, this paper highlights a significant finding: the AQI variation among three clusters in Malaysia remained unaffected by the presence of COVID-19 cases, both before and during the COVID-19 pandemic. This observation can be attributed to the calculation method of the AQI format, in which the air pollutant components (such as SO_2, NO_2, CO, O_3, PM2.5, and PM10) with the highest concentration (known as the dominant pollutant) determine the API value. In Malaysia, the API value is primarily determined by the PM2.5 concentration, which typically remains the highest among all other pollutants (Department of Environment, 2013).

This paper differs from previous studies by utilizing AQI to explore its association with COVID-19, rather than focusing solely on AQI components. Consequently, the findings diverge from those of earlier research. For example, prior studies observed a decrease in air pollution during COVID-19 lockdowns due to restricted transportation, business and industrial activities (Lee and Finerman, 2021, Wetchayont, 2021). Specifically, during COVID-19 lockdowns, air quality deteriorated in Seoul and Daegu, South Korea, because PM2.5, PM10, NO_2, CO, and SO_2 decreased (Lee and Finerman, 2021). Similarly, air pollution in Bangkok, Thailand, characterized by PM2.5, PM10, O_3, and CO, experienced reductions during this period (Wetchayont, 2021).

Limitations and Further Research

This paper, employing data mining technology for analysis, has two limitations. Firstly, the data gathered are secondary, resulting in missing data due to inherent characteristics of such data sources. Consequently, the statistical methods utilized are constrained by these limitations. Secondly, the research scope is limited to analyzing AQI variation in Malaysia's top ten most populous cities before and after the COVID-19 pandemic. In other words, the paper focuses on exploring the association among these variables and elucidating the underlying reasons behind the findings.

Two recommendations are proposed for future studies. Firstly, researchers should expand their focus beyond Malaysia to countries like the United States and Australia. This broader scope will provide comparative insights into AQI variations across different regions. Secondly, future studies should continue to utilize data mining technology. This approach allows for the analysis of a wider range of secondary data, including datasets spanning more than 48 months of time series data.

Statements & Declarations

Declaration of Interest Statement: All authors declare that no conflict of interest exists.

Data Availability Statement: The data supporting this paper's findings are available from the corresponding author upon reasonable request.

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests: The authors have no relevant financial or non-financial interests to disclose.

Author Contributions: All authors contributed to the study's conception and design.

Ethics declarations

Ethical approval: Not applicable.

Consent to participate: Not applicable.

Consent for publication: Not applicable.

Competing interests: Not applicable.

Reference

Augustin, N. H., Musio, M., von Wilpert, K., Kublin, E., Wood, S. N. & Schumacher, M. (2009). Modeling Spatiotemporal Forest Health Monitoring Data. Journal of the American Statistical Association, 104(487), pp. 899-911. DOI:10.1198/jasa.2009.ap07058

Barouki, R., Kogevinas, M., Audouze, K., Belesova, K., Bergman, A., Birnbaum, L. & Vineis, P. (2021). The COVID-19 pandemic and global environmental change: Emerging research needs. *Environment International*, 146, 106272. DOI:10.1016/j.envint.2020.106272

Chaudhuri, S. & Chowdhury, A. R. (2018). Air quality index assessment prelude to mitigate environmental hazards. *Natural Hazards*, 91(1), pp. 1-17.DOI:10.1007/s11069-017-3080-3

Chen, C. (2000). Generalized additive mixed models. *Communications in Statistics - Theory and Methods*, 29(5-6), pp. 1257-1271. DOI:10.1080/03610920008832543

Chenarides, L., Grebitus, C., Lusk, J. L. & Printezis, I. (2021). Food consumption behavior during the COVID-19 pandemic. *Agribusiness*, 37(1), pp. 44-81. DOI:,DOI:10.1002/agr.21679

Constantinescu, C. (2019, April 25). Using generalised additive mixed models (gamms) to predict visitors to edinburgh and craigmillar castles. Technical blog from our data science team. https://thedatalab.com/tech-blog/using-generalised-additive-mixed-models-gamms-to-predict-visitors-to-edinburgh-and-craigmillar-castles/

Department of Environment. (2013). General Information of Air Pollutant Index. Retrieved May 6 from http://www.doe.gov.my/webportal/en/info-umum/bahasa-inggeris-general-information-of-air-pollutant-index/

Department of Statistics Malaysia Official Portal. (2020). Population by state, administrative district and sex, 2016-2018. Retrieved April 25 from https://www.dosm.gov.my/v1/index.php?r=column3/accordion&menu_id=aHhRYUpWS3B4VXlYaVBOeUF0WFpWUT09

Environment and Climate Change Canada. (2021, April 28, 2021). About the Air Quality Health Index. Retrieved May 6 from https://www.canada.ca/en/environment-climate-change/services/air-quality-health-index/about.html

Gkatzelis, G. I., Gilman, J. B., Brown, S. S., Eskes, H., Gomes, A. R., Lange, A. C. & Kiendler-Scharr, A. (2021). The global impacts of COVID-19 lockdowns on urban air pollution: A critical review and recommendations. *Elementa: Science of the Anthropocene*, 9(1). DOI:10.1525/elementa.2021.00176

Hastie, T. J. & Tibshirani, R. J. (1990). Generalized Additive Models. Taylor & Francis. https://books.google.co.th/books?id=qa29r1Ze1coC

Hormozi, A. M. & Giles, S. (2004). Data Mining: A Competitive Weapon for Banking and Retail Industries. *Information Systems Management*, 21(2), pp. 62-71. DOI:10.1201/1078/44118.21.2.20040301/80423.9

Jenkins, N. (2015, October 4, 2015). The current haze over Southeast Asia could be among the worst ever. Time. https://time.com/4060786/haze-singapore-indonesia-malaysia-pollution/

Kaewrat, J. & Janta, R. (2021). Effect of COVID-19 Prevention Measures on Air Quality in Thailand. Aerosol and Air Quality Research, 21(3), 200344. DOI:10.4209/aaqr.2020.06.0344

Kotsiou, O. S., Kotsios, V. S., Lampropoulos, I., Zidros, T., Zarogiannis, S. G. & Gourgoulianis, K. I. (2021). PM2.5 Pollution Strongly Predicted COVID-19 Incidence in Four High-Polluted Urbanized Italian Cities during the Pre-Lockdown and Lockdown Periods. *International Journal of Environmental Research and Public Health*, 18(10), 5088. DOI:10.3390/ijerph18105088

Lee, M. & Finerman, R. (2021). COVID-19, commuting flows, and air quality. *Journal of Asian Economics*, 77, 101374. DOI:10.1016/j.asieco.2021.101374

Li, J., Hallsworth, A. G. & Coca-Stefaniak, J. A. (2020). Changing Grocery Shopping Behaviours Among Chinese Consumers At The Outset Of The COVID-19 Outbreak. *Journal of Economic and Human Geography,* 111(3), pp. 574-583. DOI:10.1111/tesg.12420

Li, L., Lin, G.-Z., Liu, H.-Z., Guo, Y., Ou, C.-Q. & Chen, P.-Y. (2015). Can the Air Pollution Index be used to communicate the health risks of air pollution? *Environmental Pollution*, 205, pp. 153-160. DOI:,DOI:10.1016/j.envpol.2015.05.038

Liao, Q., Yuan, J., Dong,M., Yang,L., Fielding,R. & Lam, W.W.T. (2020). Public Engagement and Government Responsiveness in the Communications About COVID-19 During the Early Epidemic Stage in China: Infodemiology Study on Social Media Data. *J Med Internet Res,* 22(5), e18796. DOI:10.2196/18796

Lim, Y. K., Kweon, O. J., Kim, H. R., Kim, T.-H. & Lee, M.-K. (2021). The impact of environmental variables on the spread of COVID-19 in the Republic of Korea. *Scientific Reports*, 11(1), 5977. DOI:10.1038/s41598-021-85493-y

Liu, Q., Xu, S. & Lu, X. (2021). Association between air pollution and COVID-19 infection: evidence from data at national and municipal levels. *Environ Sci Pollut Res Int*, 28(28), pp. 37231-37243. DOI:10.1007/s11356-021-13319-5

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., Appel, C. & Rodés-Guirao, L. (2021). A global database of COVID-19 vaccinations. *Nature Human Behaviour*, 5(7), pp. 947-953. DOI:10.1038/s41562-021-01122-8

Meo, S. A., Abukhalaf, A. A., Alessa, O. M., Alarifi, A. S., Sami, W. & Klonoff, D. C. (2021). Effect of Environmental Pollutants PM2.5, CO, NO2, and O3 on the Incidence and Mortality of SARS-CoV-2 Infection in Five Regions of the USA. *International Journal of Environmental Research and Public Health*, 18(15), 7810. DOI:10.3390/ijerph18157810

Pinheiro, J. C. & Bates, D. (2009). Mixed-Effects Models in S and S-PLUS. Springer. https://books.google.co.th/books?id=y54QDUTmvDcC

Plaia, A. & Ruggieri, M. (2011). Air quality indices: a review. *Reviews in Environmental Science and Bio/Technology*, 10(2), pp. 165-179. DOI:10.1007/s11157-010-9227-2

Tang, W., Hu, T., Yang, L. & Xu, J. (2020). The role of alexithymia in the mental health problems of home-quarantined university students during the COVID-19 pandemic in China. *Pers Individ Dif,* 165, 110131. DOI:10.1016/j.paid.2020.110131

The World Air Quality Index Project. (2022). Air Quality Historical Data Platform. https://aqicn.org/data-platform/register

Valdés Salgado, M., Smith, P., Opazo, M. A. & Huneeus, N. (2021). Long-Term Exposure to Fine and Coarse Particulate Matter and COVID-19 Incidence and Mortality Rate in Chile during 2020. *International Journal of Environmental Research and Public Health*, 18(14), 7409. DOI:10.3390/ijerph18147409

Wang, J., Wang, J. X. & Yang, G. S. (2020). The Psychological Impact of COVID-19 on Chinese Individuals. *Yonsei Med J*, 61(5), pp. 438-440. DOI:10.3349/ymj.2020.61.5.438

Wetchayont, P. (2021). Investigation on the Impacts of COVID-19 Lockdown and Influencing Factors on Air Quality in Greater Bangkok, Thailand. *Advances in Meteorology*, 6697707. DOI:10.1155/2021/6697707

Wong, W. M., Wang, X. & Wang, Y. (2023). The intersection of COVID-19 and air pollution: A systematic literature network analysis and roadmap for future research. *Environ Res,* 237(Pt 2), 116839. DOI:10.1016/j.envres.2023.116839

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4), pp. 1025-1036. DOI:10.1111/j.1541-0420.2006.00574.x

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology*), 73(1), pp. 3-36. DOI:,DOI:10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2017). Generalized Additive Models: An Introduction with R (2nd ed.). CRC Press. https://books.google.co.th/books?id=HL-PDwAAQBAJ

Yang, A., Qiu, Q., Kong, X., Sun, Y., Chen, T., Zuo, Y. & Peng, A. (2020). Clinical and Epidemiological Characteristics of COVID-19 Patients in Chongqing China. *Front Public Health*, 8, 244. DOI:10.3389/fpubh.2020.00244

Zhang, Y. & Ma, Z. F. (2020). Impact of the COVID-19 Pandemic on Mental Health and Quality of Life among Local Residents in Liaoning Province, China: A Cross-Sectional Study. *Int J Environ Res Public Health,* 17(7). DOI:10.3390/ijerph17072381

Zuur, A., Ieno, E. N., Walker, N., Saveliev, A. A. & Smith, G. M. (2009). Mixed Effects Models and Extensions in Ecology with R. Springer. https://books.google.co.th/books?id=vQUNprFZKHsC