VOL. 6, NO. 2 (August 2026) 192-221
https://jagst. utm.my

JAGST
Journal of Advanced Geospatial Science & Technology

Regional Spatiotemporal Variability of Multi-Pollutant Ambient Air Quality Across Major Regions of Malaysia

Siti Haslina Mohd Shafie[1], Nurul Amalin Fatihah Kamarul Zaman[2,3*]
[1]Geography Section, School of Humanities, Universiti Sains Malaysia, 11800, USM, Penang, Malaysia
[2]Geography Section, School of Distance Education, Universiti Sains Malaysia, 11800, USM, Penang, Malaysia
[3]Transdisciplinary Research on Environmental Science and Occupational Health (TREES), School of Distance Education, Universiti Sains Malaysia, 11800, USM, Penang, Malaysia

*Corresponding author: amalinkz@usm.my

Abstract – Rapid urbanisation and intensified anthropogenic activities have contributed to increasing spatial–temporal variability in ambient air pollutant concentrations across Malaysia. This study applied pollutant-specific K-means clustering to compare the spatial and temporal concentration patterns of PM_{10}, $PM_{2.5}$, CO, and NO_2 across Malaysia. Daily- and monthly-averaged data from the Department of Environment (DOE) Continuous Air Quality Monitoring (CAQM) network from 2018 to 2021 were analysed following data preprocessing procedures including mean imputation, Min–Max normalization, and Z-score outlier removal. Descriptive statistical analysis revealed significant regional disparities, with the Centre region recording the highest mean concentrations of CO (0.76 ppm), PM_{10} (30.02 μg/m³), and $PM_{2.5}$ (21.73 μg/m³), while the lowest levels were observed in the Borneo region. Following data verification and unit harmonisation, nitrogen dioxide (NO_2) showed regional variability, with the Centre region recording the highest mean concentration (0.0126 ppm). Clustering analysis identified distinct pollution regimes, with high-concentration clusters predominantly associated with urbanised zones. Daily PM_{10} concentrations in the Centre region exceeded the WHO 24-hour guideline during extreme cluster events, reaching values above 100 μg/m³. Monthly clustering further demonstrated seasonal influences, with elevated pollutant concentrations during the Southwest Monsoon associated with regional biomass burning and transboundary haze episodes. These findings indicate that severe pollution events in Malaysia are episodic rather than persistent, yet capable of generating substantial short-term exposure risks in densely populated metropolitan regions. The integration of unsupervised machine learning enables improved identification of hidden pollution regimes and supports the development of region-specific air quality management strategies aligned with Sustainable Development Goal 11.

Keywords *–Air quality; K-means clustering; Spatiotemporal variability; Criteria pollutants; Malaysia*

Article History: Received 8 April 2026, Accepted 18 August 2026, Published 31 August 2026

How to cite: Mohd Shafie, S.H. and Kamarul Zaman, N.A.F (2026). Regional Spatiotemporal Variability of Multi-Pollutant Ambient Air Quality Across Major Regions of Malaysia. Journal of Advanced Geospatial Science and Technology. 6(2), 192-221. *DOI: https://doi.org/10.11113/jagst.v6n2.129*

1.0 Introduction

Rapid urbanisation, industrial expansion, and intensified transportation activities have significantly altered the atmospheric composition of urban environments, particularly across rapidly developing nations in Southeast Asia. In Malaysia, the escalation of anthropogenic emissions from vehicular traffic, industrial combustion, and residential energy consumption has contributed to the persistent deterioration of ambient air quality, raising substantial concerns regarding environmental sustainability and public health exposure risks (Ariff et al., 2023). Among the commonly monitored criteria pollutants, particulate matter (PM_{10} and $PM_{2.5}$), nitrogen dioxide (NO_2), and carbon monoxide (CO) are widely recognised as primary indicators of urban atmospheric degradation due to their strong associations with respiratory morbidity, cardiovascular diseases, and premature mortality (WHO, 2021).

Globally, ambient air pollution is estimated to contribute to approximately 4.2 million premature deaths annually, predominantly in low- and middle-income countries where urban growth often outpaces environmental governance mechanisms (Health Effects Institute [HEI], 2022). Within the Southeast Asian region, recurring transboundary haze episodes, driven largely by biomass burning and peatland fires, further exacerbate pollutant accumulation under stagnant meteorological conditions, particularly during inter-monsoonal and dry-season transitions. These episodic events frequently elevate pollutant concentrations beyond recommended thresholds, resulting in acute exposure risks, especially in densely populated urban corridors.

In the Malaysian context, spatial disparities in air pollutant distribution are strongly influenced by heterogeneous land use patterns, population density, traffic intensity, and regional economic development. Highly urbanised zones such as the Centre and North regions of Peninsular Malaysia consistently exhibit elevated concentrations of particulate and gaseous pollutants due to concentrated emission sources and urban heat island–induced atmospheric stagnation (Mohd Shafie et al., 2022; Naidin et al., 2024). Conversely, regions with lower industrialisation levels, such as Borneo, typically demonstrate comparatively lower baseline concentrations, although episodic increases linked to long-range pollutant transport remain evident. These spatially heterogeneous patterns suggest that air pollution dynamics across Malaysia are governed by a complex interplay between localised emission intensity, meteorological variability, and regional atmospheric transport mechanisms.

In tropical climates such as Malaysia, atmospheric processes including monsoonal wind circulation, convective rainfall patterns, and boundary layer dynamics play a crucial role in regulating pollutant dispersion and accumulation. During dry inter-monsoon periods, reduced precipitation and lower wind speeds can limit atmospheric mixing, resulting in the accumulation of particulate and gaseous pollutants near ground level. Conversely, wet monsoon seasons typically facilitate pollutant removal through wet deposition processes. Such meteorological variability introduces substantial temporal fluctuations in pollutant concentration profiles, which may obscure underlying emission-driven patterns when analysed using traditional statistical approaches alone.

Although continuous monitoring networks established by environmental agencies have enabled the generation of high-resolution air quality datasets, conventional statistical approaches such as descriptive trend analysis and regression modelling often provide limited insight into the underlying temporal and spatial structures embedded within atmospheric time series data. Recent studies have highlighted the limitations of traditional analytical frameworks in detecting pollution regimes influenced by seasonal variability, transboundary transport, and emission-driven anomalies. Consequently, there is an increasing need for analytical approaches capable of capturing hidden patterns and non-linear relationships within large environmental datasets.

In response to these challenges, machine learning–based clustering techniques have emerged as effective tools for the exploratory analysis of environmental time series data. Unsupervised algorithms, particularly K-means clustering, have been widely applied to classify pollution episodes into distinct regimes based on similarities in pollutant concentration patterns across both spatial and temporal dimensions. Such approaches enable the identification of high-pollution anomalies, transitional episodes, and background concentration regimes that may not be readily observable through conventional statistical analyses. Furthermore, clustering-based segmentation offers a robust framework for associating pollution regimes with potential emission sources and meteorological drivers, thereby supporting improved air quality governance and early-warning strategies in rapidly urbanising regions.

Despite the growing application of machine learning techniques in air quality assessment, empirical investigations examining the integrated spatial–temporal clustering of multiple criteria pollutants across Malaysia remain limited. Existing studies predominantly focus on individual pollutants or single monitoring locations, which may not adequately capture the regional variability in atmospheric pollution dynamics across Peninsular and East Malaysia. This limitation underscores the need for comprehensive analytical frameworks capable across diverse geographical regions.

Moreover, understanding the spatial–temporal variability of pollutant concentrations is essential for developing targeted mitigation strategies that account for region-specific emission sources and climatic influences. Data-driven identification of pollution regimes may assist policymakers in formulating adaptive air quality management plans, particularly in urban areas experiencing rapid infrastructure development and increasing vehicular dependency. Integrating clustering-based analytical frameworks into air quality assessment may also enhance the predictive capability of environmental monitoring systems by enabling the detection of recurring pollution episodes and transitional atmospheric states.

In light of the aforementioned research gaps, this study is designed to achieve several primary objectives. First, it aims to examine the spatial variability of daily- and monthly-averaged concentrations of PM_{10}, $PM_{2.5}$, CO, and NO_2 across five major geographical regions of Malaysia, namely the Centre, North, South, East, and Borneo. Second, the study seeks to identify and classify distinct pollution regimes through the application of K-means clustering techniques, thereby capturing temporal fluctuations in pollutant concentration patterns that may be associated with anthropogenic activities or seasonal meteorological influences. Third, this study intends to evaluate inter-regional similarities and divergences in pollutant behaviour to better understand the spatial coherence of air pollution dynamics within a tropical urban context. The scope of this study encompasses both daily- and monthly-averaged pollutant concentration datasets obtained from multiple monitoring stations distributed across the five aforementioned regions. By adopting a multi-pollutant analytical framework, the study focuses on identifying recurring pollution regimes and transitional atmospheric conditions that may influence short-term exposure risk. Unlike conventional trend-based assessments, this approach facilitates the segmentation of large-scale air quality datasets into meaningful clusters, thereby enabling a more nuanced interpretation of pollutant variability across different spatial and temporal scales.

Ultimately, the findings of this study are expected to contribute towards the advancement of data-driven air quality management strategies in Malaysia. By enhancing the understanding of regional pollution regimes and their temporal behaviour, this research may support the formulation of targeted mitigation measures and adaptive policy interventions, particularly in rapidly urbanising zones where emission intensity and population exposure remain critically high. In doing so, the study provides an empirical foundation for strengthening environmental monitoring frameworks aligned with national sustainable development priorities and the broader objectives of Sustainable Development Goal (SDG) 11.

2.0 Study Area

The geographical scope of this study encompasses all thirteen states and three Federal Territories of Malaysia, which were grouped into five analytical regions: North, Centre, South, East Coast, and Borneo (Figure 1). Sabah and Sarawak were combined as the Borneo region for regional aggregation. This classification is based on administrative boundaries while also reflecting broad differences in topography, climatic conditions, and dominant land-use characteristics that may influence ambient air-quality patterns.

The North region is primarily characterised by agricultural land use, whereas the Centre region represents the most densely urbanised and industrialised zone in the country, with substantial anthropogenic emission sources. In contrast, the Borneo region, comprising Sabah and Sarawak, is largely covered by tropical rainforest ecosystems and extensive peatland areas that are susceptible to seasonal biomass burning activities and transboundary haze intrusion (Gaveau et al., 2014).

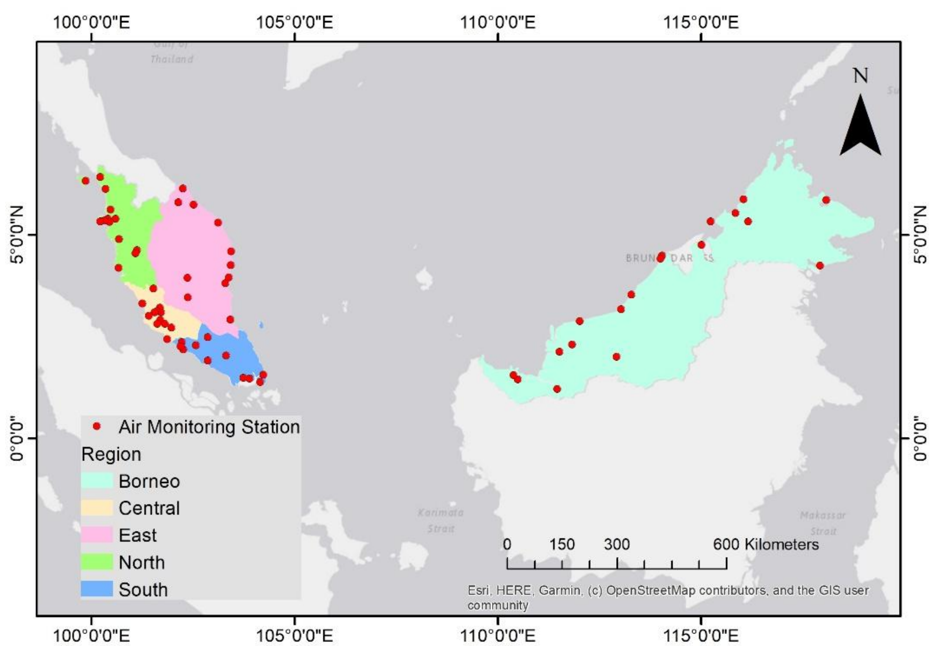

Figure 1: $PM_{2.5}$ Monitoring stations (red dots) are grouped into five regions.

2.1 Topographic and Climatic Influences

The spatial distribution and variability of air pollutant concentrations in Malaysia are strongly influenced by the country's topographical features and prevailing climatic conditions. In Peninsular Malaysia, the Titiwangsa Mountain Range serves as a natural climatic boundary separating the East and western coastal regions. This orographic formation alters regional wind circulation, precipitation patterns, and atmospheric mixing processes, consequently shaping pollutant dispersion behaviour across both coastal zones. As a result, differences in local microclimatic conditions may contribute to spatial inconsistencies in ambient pollutant accumulation (Dominick et al., 2012).

Urbanised areas such as the Klang Valley, which includes Kuala Lumpur and its surrounding municipalities, are characterised by relatively low-lying terrain that may restrict vertical atmospheric mixing under stable meteorological conditions. Such geographical settings can facilitate the retention of pollutants, particularly during periods of weak wind circulation or temperature inversion events (Dominick et al., 2012). In East Malaysia, complex terrain features such as Mount Kinabalu and the Crocker Range in Sabah further influence regional airflow and rainfall distribution through orographic lifting mechanisms, potentially limiting pollutant dispersion within enclosed valley systems.

Additionally, the presence of peatland ecosystems in Sarawak and certain parts of Sabah introduces further atmospheric complexity. Degraded or drained peatlands are highly susceptible to combustion during prolonged dry conditions, releasing substantial quantities of fine particulate matter ($PM_{2.5}$), carbon monoxide (CO), and volatile organic compounds (VOCs) into the atmosphere. These emissions often contribute to large-scale haze events during the Southwest Monsoon season, particularly when transboundary biomass burning occurs in neighbouring regions (Gaveau et al., 2014).

Malaysia's equatorial rainforest climate, classified as Af under the Köppen–Geiger system, is characterised by consistently high temperatures ranging from approximately 26°C to 32°C, relative humidity levels between 70% and 90%, and mean annual precipitation exceeding 2,500 mm. Seasonal monsoonal cycles further influence atmospheric pollutant dynamics. The Northeast Monsoon period, typically occurring from November to March, is associated with increased rainfall that promotes wet deposition and pollutant removal. In contrast, the drier Southwest Monsoon season from May to September is often linked to reduced precipitation, thereby facilitating pollutant accumulation and increasing the likelihood of regional haze episodes originating from biomass burning activities in Sumatra and Kalimantan. These conditions may be further intensified during El Niño events, which are known to exacerbate drought frequency and fire occurrences within the region.

3.0 Methodology

3.1 Data sources

Malaysia's DOE maintains a comprehensive air quality surveillance system through its Continuous Air Quality Monitoring (CAQM) network, which consists of automated monitoring stations strategically distributed across urban, suburban, industrial, and rural settings. These stations are designed to measure ambient concentrations of key air pollutants, which include $PM_{2.5}$, PM_{10}, CO, and NO_2, on an hourly basis from 1 January 2018 to 31 December 2021 with a primary focus on emissions resulting from transportation, domestic fuel use, and industrial sources. All monitoring activities adhere to internationally recognized standards, including protocols established by the U.S. Environmental Protection Agency (USEPA) and the European Committee for Standardization (CEN).

The management, calibration, and quality control of these monitoring stations are entrusted to Alam Sekitar Malaysia Sdn. Bhd. (ASMA), a DOE-appointed technical agency. ASMA oversees daily operations, including system diagnostics, calibration routines, and data verification procedures. This ensures data consistency, minimizes measurement uncertainties, and enhances the scientific reliability of the datasets for temporal trend analysis and advanced forecasting models.

The monitoring stations were grouped into five regions: North, Centre, South, East Coast, and Borneo. For each pollutant, the daily concentration for each region was calculated by averaging the available monitoring-station data within that region. Monthly regional concentrations were subsequently calculated from the daily regional averages. All available stations within each region were given equal weight. The number of stations used in each region was as follows: North (14 stations), Centre (11 stations), South (11 stations), East Coast (11 stations), and Borneo (18 stations).

3.2 Data preprocessing

This study investigates the clustering of daily and monthly-averaged air pollutant concentrations, specifically PM_{10}, $PM_{2.5}$, CO, and NO_2, recorded at various air quality monitoring stations under the DOE in Malaysia using the K-means clustering algorithm. The methodology involves data collection, preprocessing, and clustering analysis to identify distinct pollution patterns and explore the relationship between air quality and time. Before performing the clustering analysis, data preprocessing was necessary to ensure quality and consistency.

Following regional aggregation, the daily dataset contained complete observations for PM_{10}, $PM_{2.5}$, CO, and NO_2 throughout the study period (2018–2021). Therefore, no missing-value imputation was applied in the clustering analysis. Monthly values were derived from the complete daily regional dataset.

Prior to analysis, NO_2 observations were checked to ensure consistency of measurement units across the study period. During data verification, the North-region NO_2 records for 2021 were identified as being reported in ppb and had not been converted to ppm before regional aggregation. These values were corrected using the conversion of 1,000 ppb = 1 ppm. All subsequent descriptive and clustering analyses were conducted using the corrected NO_2 dataset in ppm. Data scaling was applied only for the K-means clustering procedure. For reporting and environmental interpretation, cluster means were calculated and presented using the original pollutant concentration units.

Prior to K-means clustering, each pollutant was analysed separately at the daily and monthly temporal resolutions. To minimise the influence of differences in concentration magnitude among the five regions, the regional concentration variables were normalised separately using Min–Max normalisation. The transformation rescaled each regional variable to a range between 0 and 1:

$$X = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \text{(Equation 1)}$$

Where:

X = original value,

X_{min} = minimum value in the feature,

X_{max} = maximum value in the feature,

X_{norm} = normalized value.

The normalised data were used only to determine cluster membership in the K-means procedure. For reporting and environmental interpretation, the mean concentrations for each cluster were calculated from the original, unnormalised pollutant concentrations and are reported in their respective physical units, namely μg/m^3 for PM_{10} and $PM_{2.5}$, and ppm for CO and NO_2. The study adopted a comparative multi-pollutant approach involving PM_{10}, $PM_{2.5}$, CO, and NO_2. However, K-means clustering was conducted separately for each pollutant rather than jointly across all pollutants. This pollutant-specific approach enabled the identification and comparison of distinct spatial and temporal concentration patterns for each pollutant at the daily and monthly scales.

NoeExtreme pollutant concentrations were retained in the analysis because they may represent genuine high-pollution episodes, including haze events. No observations were removed solely based on their statistical magnitude. The study used the final regional dataset derived from the DOE monitoring records, which had undergone routine quality-control procedures by the relevant monitoring authority.

3.3 Clustering analysis using K-means

3.3.1 K-Means algorithm overview

The K-means clustering algorithm was applied to group the daily pollutant concentrations into distinct clusters based on their similarity. The K-means method is a widely used unsupervised machine learning technique for partitioning data into clusters where each observation belongs to the cluster with the nearest mean. The algorithm works by minimizing the within-cluster sum of squares (WCSS),

$$j = \sum_{i}^{k} \sum_{xj \in C_i} ||x_j - \mu_i||^2 \qquad \text{(Equation 2)}$$

Where:

J = is the objective function (within-cluster sum of squares),

K = is the number of clusters,

C_i = set of data points in the (i)-th cluster,

X_j = data point in cluster (C_i),

μ_i = centroid of the (i)-th cluster,

$\| x_j - \mu_i \|^2$ = squared Euclidean distance between data point (x_j) and the centroid μ_i

This minimization results in the assignment of data points to the nearest cluster centroid. The algorithm iteratively updates the cluster centroids and reassigns data points until convergence is achieved, ensuring that the clusters are as compact and distinct as possible. To determine the appropriate number of clusters (k), the Cubic Clustering Criterion (CCC) was used. The CCC compares the observed clustering structure with that expected under a uniform distribution. Higher positive CCC values indicate stronger clustering tendency, whereas negative values indicate weak or less distinct cluster separation. Candidate clustering solutions were assessed across different values of k, and the final solution was selected by considering the CCC value, cluster size distribution, and the interpretability of the resulting concentration patterns.

3.3.2 Cluster validation techniques

The K-means clustering algorithm was executed on the normalized dataset with the number of clusters (k), selected based on the criteria mentioned above. The algorithm was run for a maximum of 100 iterations with a tolerance level of 0.0001, ensuring convergence to a stable solution. The clustering output was then analyzed to identify patterns in the concentrations of PM_{10}, $PM_{2.5}$, CO, and NO_2, across the different clusters.

Once the optimal number of clusters was determined, the daily pollutant concentration profiles were examined for each cluster. Descriptive statistics such as mean, median, and standard deviation were computed for each cluster, providing insights into the levels of air pollution and potential sources. Temporal patterns of pollution were also assessed by analyzing the frequency of high-concentration days and identifying periods of extreme pollution (e.g., monsoon or dry season) (Zoran et al., 2023). Cluster solutions were evaluated using the Cubic Clustering Criterion (CCC), cluster-size distribution, and visual inspection of the scatterplot matrices. Clusters with very small memberships were interpreted cautiously as potentially isolated or episodic concentration patterns rather than stable pollution regimes.

3.3.3 Cluster Robustness Assessment

To assess the robustness of the clustering solutions, K-means clustering was evaluated across alternative values of *k*. For the monthly analysis, solutions that produced singleton or very small clusters were not considered stable representations of recurring pollution patterns. The final clustering solution was selected by considering the CCC value, cluster-size distribution, and the interpretability of the resulting concentration patterns. Clusters with fewer than three monthly observations were interpreted cautiously as isolated episodes rather than stable pollution regimes.

4.0 Descriptive Statistical Distribution of Pollutants

The descriptive analysis revealed pronounced spatial variability in pollutant concentrations across the selected study regions (Table 1). In the case of carbon monoxide (CO), the Centre region recorded the highest average concentration at approximately 0.76 ppm, whereas the lowest mean level was observed in the East region at 0.38 ppm. The remaining regions, namely the North, South, and Borneo zones, exhibited intermediate concentration levels ranging between 0.55 and 0.59 ppm.

A similar spatial pattern was observed for particulate matter. The Centre region demonstrated the highest mean PM_{10} concentration (30.02 $\mu g/m^3$), which may be attributed to the intensity of urbanisation and industrial activities in this area. Conversely, the Borneo region recorded the lowest PM_{10} concentration at 19.49 $\mu g/m^3$, indicating comparatively improved ambient air quality conditions. In terms of $PM_{2.5}$, a declining concentration gradient was observed from the Centre region (21.73 $\mu g/m^3$) towards Borneo (11.33 $\mu g/m^3$), suggesting the influence of population density and urban development on fine particulate emissions.

Nitrogen dioxide (NO_2) concentrations showed regional variation across the study regions. Following unit harmonisation and data verification, the Centre region recorded the highest mean NO_2 concentration (0.0126 ppm), while the other regions showed lower mean concentrations. The corrected mean NO_2 concentration for the North region was 0.0055 ppm. The observed variability in standard deviation and confidence interval values across all pollutants indicates differing levels of concentration dispersion among the regions, with the Centre region generally demonstrating greater fluctuation, likely due to the complexity of urban emission sources and atmospheric interactions. Collectively, these results underscore the existence of regional differences in air pollutant distribution, which may be influenced by anthropogenic activities, geographical characteristics, and prevailing meteorological conditions. These findings establish a preliminary basis for subsequent spatial–temporal assessment of air quality dynamics.

Table 1: Descriptive statistical distribution of criteria pollutants

Item	Centre	East	North	South	Borneo
CO					
Mean	0.756586	0.381195	0.571889	0.589982	0.546317
Median	0.742665	0.4672	0.560004	0.57776	0.53805
Standard Deviation	0.140297	0.229835	0.108537	0.10119	0.088334
Minimum	0.428539	0.001872	0.315232	0.363687	0.331671
Maximum	1.570819	1.083071	1.350364	1.219812	1.341072
Confidence Level (95.0%)	0.0072	0.011795	0.00557	0.005193	0.004533
PM_{10}					
Mean	30.02188	24.11029	23.39425	23.53759	19.49422
Median	27.47938	22.18151	21.49376	20.99227	17.07957
Standard Deviation	13.2794	9.392372	9.425763	10.7049	10.22811
Minimum	10.57138	9.433416	8.151811	8.141709	10.21204
Maximum	152.0813	107.7574	120.3862	123.0269	119.733
Confidence Level (95.0%)	0.681493	0.482013	0.483726	0.549371	0.524902
$PM_{2.5}$					
Mean	21.72954	15.33029	15.7273	15.98895	11.33173
Median	19.37283	13.52602	13.90069	13.68314	9.161258
Standard Deviation	11.66731	8.219983	8.232179	9.125585	8.978459
Minimum	7.469545	4.839432	4.461352	4.780595	4.674812
Maximum	135.6866	93.20254	107.6958	102.7865	99.85785
Confidence Level (95.0%)	0.598761	0.421846	0.422472	0.468321	0.460771
NO_2					
Mean	0.012628	0.004428	0.005450	0.005792	0.004181
Median	0.012681	0.004162	0.006693	0.005744	0.003935
Standard Deviation	0.003461	0.001396	0.003129	0.002445	0.001324
Minimum	0.003504	0.001461	0.000258	0.001852	0.001278

Maximum	0.023706	0.010926	0.012092	0.013385	0.009211
Confidence Level (95.0%)	0.000178	7.16E-05	0.000160	0.000125	6.79E-05

4.2 K-means clustering analysis

The clustering procedure was performed using Min–Max normalised data; however, all cluster mean concentrations presented in the following sections were calculated using the original concentration values. Therefore, the reported values retain their original physical units and can be interpreted environmentally. The clustering results are presented separately for PM_{10}, $PM_{2.5}$, CO, and NO_2 because each pollutant was analysed independently. This approach allows pollutant-specific regional concentration patterns to be compared without implying that the four pollutants were jointly used to define a single clustering solution.

4.1 Spatial-temporal clustering of daily-averaged pollutants

4.1.1 PM_{10} Concentration

The clustering analysis of daily-averaged PM_{10} concentrations across five major regions of Malaysia, which is Borneo, Centre, East, North, and South, produced three distinct clusters based on the K-means algorithm output. The daily PM_{10} dataset was grouped into three clusters (K = 3) based on the selected K-means solution. The final clustering solution was evaluated using the CCC, cluster-size distribution, and interpretability of the regional concentration patterns. The selected clustering solution was evaluated using the Cubic Clustering Criterion (CCC), cluster-size distribution, and the interpretability of regional concentration patterns. The negative CCC value (−10.231) indicates relatively weak cluster separation; therefore, the identified clusters are interpreted as exploratory concentration groupings rather than strongly distinct and stable pollution regimes (Figure 2a).

The cluster summary reveals an asymmetric distribution of observations, with Cluster 3 comprising 1,135 daily records, Cluster 2 with 308, and Cluster 1 with only 18. This indicates that the majority of days across all regions are categorized under relatively low PM_{10} levels (Cluster 3), followed by moderate pollution episodes (Cluster 2), while only a small fraction of days represent high-pollution outliers (Cluster 1). The clustering analysis identifies recurring concentration patterns and regional differences but does not independently establish the causal drivers of these patterns. Therefore, the potential influences of traffic emissions, industrial activities, meteorological conditions, biomass burning, transboundary haze, and movement restrictions should be interpreted as plausible explanations rather than confirmed causal relationships. Further analysis incorporating meteorological variables, emission inventories, hotspot data, mobility indicators, and air-mass trajectory information would be required to confirm these relationships. This distribution pattern reflects typical urban air quality patterns in tropical Southeast Asia, where clean-air days dominate, punctuated by episodic spikes due to anthropogenic or meteorological anomalies (Latif et al., 2020).

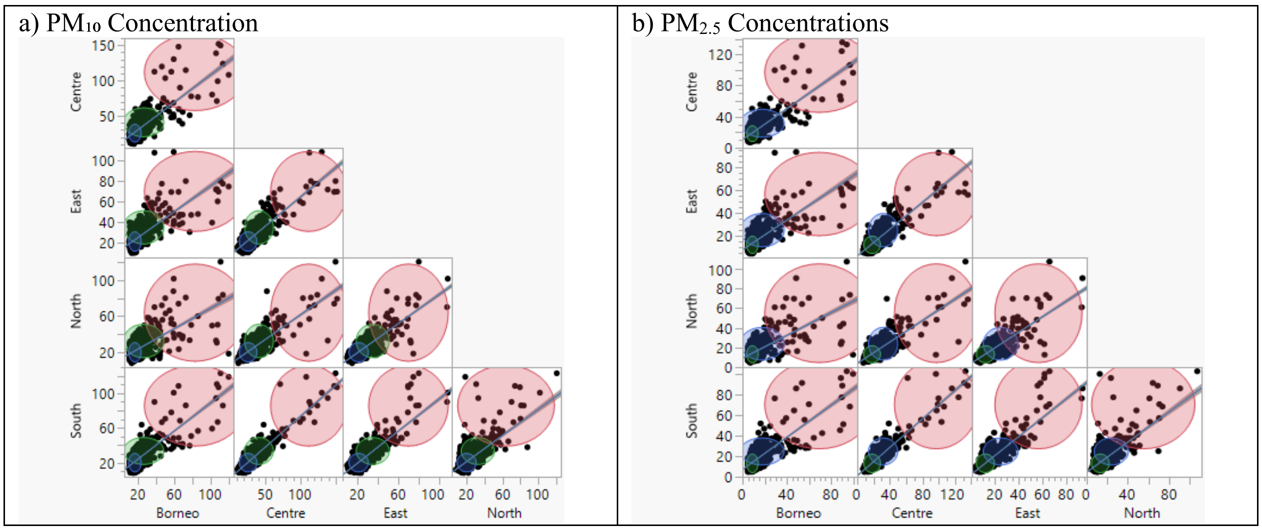

Figure 2: Scatter plot matrix of inter-regional daily-averaged a) PM_{10} and b) $PM_{2.5}$ concentrations across five regions of Malaysia

The mean concentrations across clusters and regions show striking differences. In Cluster 1 (high PM_{10}), the Centre region recorded the highest mean daily concentration at 111.75 µg/m³, followed by the South (85.62 µg/m³) and Borneo (82.88 µg/m³). These values exceed the WHO 24-hour guideline for PM_{10}, which recommends levels below 25 µg/m³ (WHO, 2021), highlighting serious short-term exposure risks. This cluster likely corresponds to periods of intense anthropogenic activity such as traffic congestion, industrial combustion, or transboundary haze episodes originating from peat and forest fires in the Indonesian archipelago. The inclusion of Borneo, a region typically perceived as less industrialized, in the high pollution cluster suggests the influence of long-range transport or localized sources such as diesel-powered generators and open burning.

In contrast, Cluster 3, comprising the majority of the dataset, represents background or clean conditions. The mean PM_{10} concentrations in this cluster are significantly lower, ranging from 16.64 µg/m³ in Borneo to 25.71 µg/m³ in the Centre. These values are within acceptable air quality limits and likely correspond to periods of favorable meteorological dispersion (e.g., high wind speeds, rainfall) or reduced human activity, such as public holidays or pandemic-related lockdowns (Abdullah et al., 2023). Cluster 2, representing moderate pollution levels, exhibits transitional concentrations and may be influenced by seasonal variability or intermediate emissions from urban sources.

The scatterplot matrix illustrates positive correlations among all regional PM_{10} concentrations, especially between the Centre and South, where elliptical clustering patterns reflect linear relationships. This suggests that air pollution trends in these two regions are synchronized, possibly due to shared economic activities, urban development corridors, and continuous vehicular emissions. The East and North regions exhibit weaker correlations with Borneo, reinforcing their different pollution dynamics and source compositions. Nevertheless, some overlap exists in the ellipses of Clusters 2 and 3, suggesting transitional days that could shift clusters depending on slight meteorological changes or variations in emission intensity. These findings align with previous studies that observed the highest PM_{10} concentrations in Peninsular Malaysia's Centre urban belt, including Kuala Lumpur and Selangor, due to high vehicle density, industrial zones, and energy consumption (Ariff et al., 2023). The use of K-means clustering effectively segments air quality data, identifying both persistent pollution zones and episodic high-exposure events. This clustering approach supports the detection of hidden structures within time series pollution data, which conventional trend analysis might overlook.

Furthermore, the cluster-based classification provides a useful temporal framework to associate PM_{10} exposure patterns with potential health outcomes. Short-term exposure to elevated particulate matter concentrations, particularly during episodic peaks exceeding air quality guideline thresholds, has been consistently associated with increased risks of respiratory and cardiovascular morbidity, especially among vulnerable populations such as children and the elderly (World Health Organization, 2021). Hence, early-warning systems based on cluster identification could be instrumental in mitigating public health risks through pre-emptive advisories.

4.1.2 $PM_{2.5}$ Concentration

K-means clustering of daily $PM_{2.5}$ concentrations produced three groups (K = 3). However, the negative CCC value (−8.684) indicates relatively weak cluster separation. Therefore, the identified groups are interpreted as exploratory concentration patterns rather than strongly distinct and stable pollution regimes. Cluster 1, which contained only 18 observations, represents rare high-concentration episodes and should be interpreted cautiously. Cluster 2 dominated the dataset with 1,172 observations, representing the baseline $PM_{2.5}$ conditions, while Cluster 1 (18 observations) denoted periods of extreme pollution, and Cluster 3 (271 observations) reflected moderate pollution scenarios. Analysis of cluster means revealed that the Centre region consistently recorded the highest $PM_{2.5}$ concentrations across all clusters, particularly in Cluster 1, where levels peaked at 97.01 µg/m³. This reflects the substantial influence of dense traffic networks, industrial activities, and urban infrastructure. Cluster 1, which represents high-pollution episodes, also showed elevated values in the South (70.34 µg/m³) and Borneo (69.18 µg/m³), suggesting widespread haze conditions or long-range pollutant transport, often exacerbated by transboundary biomass burning during dry spells. In contrast, Cluster 2 displayed comparatively lower $PM_{2.5}$ levels, with values such as 8.95 µg/m³ in Borneo and 12.42 µg/m³ in the East, possibly corresponding to cleaner air periods aided by meteorological factors like rainfall or favorable wind conditions. Cluster 3 captured intermediate pollution levels, likely representing transitions or localized emission events (Figure 2b).

The scatterplot matrix validates these clusters, particularly showing compact grouping for Cluster 2 (low $PM_{2.5}$), while Clusters 1 and 3 displayed wider dispersion in higher concentration ranges. The plots also suggest moderate interregional correlations, especially between the Centre and South, and Centre and North, indicating synchronicity in pollution episodes across interconnected urban zones in Peninsular Malaysia. Although Borneo generally enjoys better air quality, its high $PM_{2.5}$ values in Cluster 1 indicate vulnerability during regional haze events, likely due to its proximity to biomass burning hotspots in Kalimantan and Sumatera.

4.1.3 CO Concentration

The optimal clustering solution consisted of six clusters (K = 6), supported by a robust CCC = 0.91121, indicating statistically meaningful group separation. To account for regional disparities in baseline pollution, the regional CO variables were normalised separately using Min–Max normalisation before clustering, ensuring comparability across spatial zones. Cluster sizes ranged significantly, from as few as 3 to as many as 446 observations, indicating varied temporal frequency and intensity of CO pollution events. Cluster 1 (n = 33) and Cluster 5 (n = 298) corresponded to episodes of elevated and moderate CO concentrations, respectively. Meanwhile, Cluster 3 (n = 17) represented rare occurrences, potentially indicative of unique pollution conditions or atypical meteorological events (Figure 3a).

Mean CO levels across clusters revealed notable regional variability. Cluster 1, which represented the highest pollution levels, exhibited elevated CO concentrations in Borneo (1.26 ppm), Centre (1.27 ppm), and East (0.72 ppm). These values suggest sporadic pollution surges likely associated with biomass burning, transboundary haze, or localized combustion sources. Conversely, Cluster 2, which included some of the lowest recorded values (e.g., Borneo: 0.48 ppm; East: 0.30 ppm; North: 0.49 ppm), may correspond to periods of diminished emissions possibly linked to reduced anthropogenic activity such as the COVID-19 lockdowns (Mohd Nadzir et al., 2020). Cluster 6 displayed moderate CO concentrations with relatively uniform values across regions, indicating a stable pollution regime, potentially driven by recurring seasonal patterns and consistent vehicular emissions.

Visual inspection through scatterplot matrices reinforced these findings, with clear and distinct clustering observed, with ellipse contours delineating well-separated groupings, underscoring the validity of the clustering structure. Furthermore, a strong positive correlation between CO concentration in the Centre and East regions was identified, possibly reflecting concurrent urban activities or shared meteorological influences. Borneo often diverged from peninsular trends, a distinction likely due to differences in land use, industrial development, and lower traffic volume (DOE, 2022).

The results underscore critical implications for regional air quality governance. The concentration of high CO levels in urban zones during specific periods signals the need for targeted control measures, including stricter vehicle emission standards, enhanced transboundary haze monitoring, and localized air quality interventions. Moreover, the identification of distinct spatial-temporal pollution clusters reinforces the necessity of regionally tailored policies that consider both emission sources and climatic conditions. This study further validates the application of K-Means clustering as an effective tool for extracting meaningful patterns from complex environmental datasets, in alignment with growing evidence supporting the integration of machine learning in atmospheric monitoring.

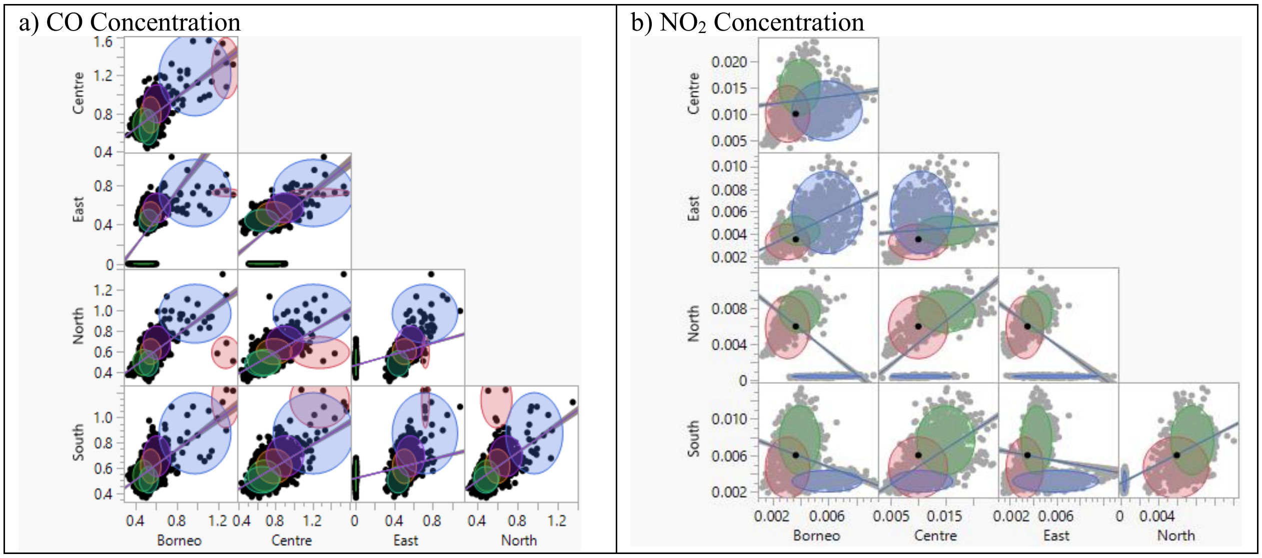

Figure 3: Scatter plot matrix of inter-regional daily-averaged a) CO and b) NO_2 concentrations across five regions of Malaysia

4.1.4 NO_2 Concentration

The analysis was conducted using the corrected NO_2 dataset, in which all observations were harmonised and reported in ppm. The three-cluster solution revealed distinct daily NO_2 concentration patterns across the five Malaysian regions, indicating substantial spatial and temporal variability in ambient NO_2 levels with supported by CCC = 11.4456) (Figure 3b). Cluster sizes ranged from 355 to 725 daily observations, indicating that all identified clusters represented recurring NO_2 concentration conditions rather than isolated events. Cluster 1 (n = 381) was characterised by relatively lower NO_2 concentrations, with mean values of 0.0031 ppm in Borneo, 0.0100 ppm in the Centre region, 0.0033 ppm in the East Coast, 0.0059 ppm in the North region, and 0.0047 ppm in the South region. Cluster 2 (n = 355) showed comparatively higher concentrations in Borneo (0.0059 ppm) and the East Coast (0.0059 ppm), while the North region recorded the lowest mean concentration (0.0005 ppm). Cluster 3 (n = 725), the largest cluster, represented the relatively higher NO_2 pattern, particularly in the Centre region (0.0150 ppm), North region (0.0077 ppm), and South region (0.0077 ppm).

The cluster profiles indicate considerable regional variation in daily NO_2 concentrations. The Centre region consistently exhibited the highest concentration levels across all three clusters, ranging from 0.0100 to 0.0150 ppm. This pattern may be associated with higher traffic density, urbanisation, industrial activity, and combustion-related emissions. In contrast, Borneo and the East Coast generally recorded lower NO_2 concentrations, although elevated concentrations were observed in Cluster 2. The corrected North-region values ranged from 0.0005 to 0.0077 ppm across the three clusters.

Overall, the daily clustering analysis confirms that NO_2 concentrations were not spatially uniform across Malaysia. The three clusters represent recurring low, intermediate, and relatively high regional NO_2 concentration conditions, reflecting the combined influence of local emission sources, atmospheric dispersion, and temporal changes in pollutant accumulation.

4.2 Clustering of monthly-averaged concentrations

4.2.1 PM_{10} Concentration

The K-means clustering analysis identified eight distinct clusters, each representing unique patterns of PM_{10} concentrations across Borneo, Centre, East Coast, North, and South (Figure 4a). Cluster 1 exhibited the highest PM_{10} concentrations in the Centre and North regions, with mean values of 45.62 µg/m³ and 50.14 µg/m³, respectively. In contrast, Cluster 3 showed the lowest concentrations, particularly in the Borneo region, with a mean value of 20.52 µg/m³. Additionally, Cluster 8 contributed to the overall clustering structure, further defining the regional patterns, with a CCC value of 1.35121. The clustering results reveal significant regional variations in PM_{10} concentrations, with the Centre and North regions consistently exhibiting higher levels. This suggests that these regions are more prone to higher pollution levels, likely due to their dense urbanization and industrial activities. Further analysis of the clusters indicates that the East Coast and South regions also experienced elevated PM_{10} levels, though not as high as the Centre and North regions. For instance, Cluster 4, which had relatively high PM_{10} concentrations, showed mean values of 40.87 µg/m³ in the Centre and 45.34 µg/m³ in the North. The Borneo region, characterized by lower industrial activity and population density, consistently showed the lowest PM_{10} concentrations across all clusters. This regional disparity highlights the significant impact of urbanization and industrialization on air quality. The lower PM_{10} levels in Borneo can be attributed to its extensive forest cover and lower human activity, which act as natural filters for air pollutants. Given the limited number of monthly observations, the monthly clustering results should be interpreted cautiously. Clusters with very small memberships may represent isolated extreme months rather than stable recurring pollution patterns. Further sensitivity analysis across alternative values of k is recommended to assess the robustness of these monthly clustering solutions.

The seasonal variation in PM_{10} levels was evident, with higher concentrations recorded during the dry season, coinciding with transboundary haze episodes originating from neighboring countries. These haze episodes significantly contribute to elevated PM_{10} levels, particularly in the Centre and North regions. The monsoon phenomenon further influences the monthly variation trend of PM_{10}, with lower concentrations typically observed during the wet season. This seasonal pattern underscores the importance of considering temporal factors in air quality management. The dry season's higher PM_{10} levels are likely due to increased biomass burning and reduced rainfall, which limits the natural dispersion of pollutants (Naidin et al., 2024).

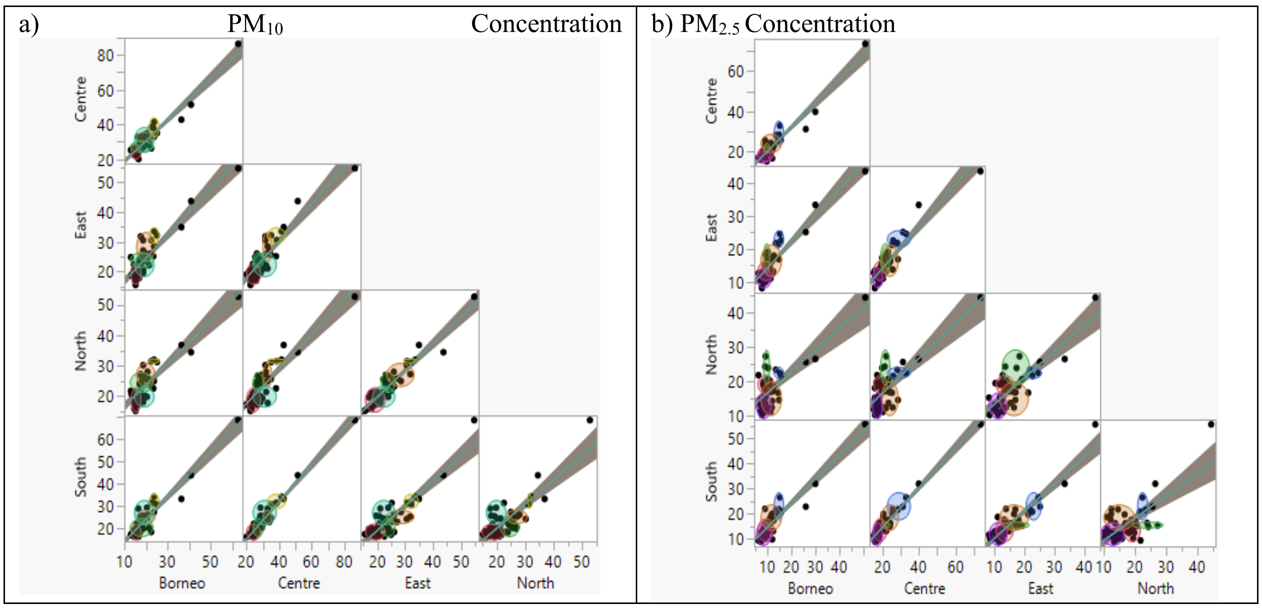

Figure 4: Scatter plot matrix illustrating regional variations in monthly a) PM_{10} concentration and b) $PM_{2.5}$ concentrations across Malaysia's major zones

Additionally, the clustering analysis provides valuable insights into the spatial and temporal patterns of PM_{10} concentrations in Malaysia. Identifying distinct clusters allows for a more nuanced understanding of regional air quality dynamics. These findings underscore the need for targeted air quality management strategies considering regional and seasonal variations. Effective policies and measures, such as stricter emission controls and regional cooperation to address transboundary pollution, are essential to mitigate the impact of PM_{10} on public health and the environment. The results also suggest that localized interventions, tailored to the specific pollution sources and seasonal patterns of each region, could be more effective in improving air quality (Shakor et al., 2020).

The findings from the K-means clustering analysis highlight significant regional variations in PM_{10} concentrations across Malaysia. The Centre and North regions consistently exhibited higher PM_{10} levels, which can be attributed to several factors, including higher population density, industrial activities, and vehicular emissions. These regions are known for their urbanization and industrialization, which contribute to elevated PM_{10} levels. The Borneo region, on the other hand, showed the lowest PM_{10} concentrations. This can be attributed to its lower population density and industrial activity, which results in fewer sources of PM_{10} emissions. The findings align with previous studies that highlight the impact of urbanization and industrialization on air quality.

Seasonal variations also play a crucial role in influencing PM_{10} levels. The highest concentrations were recorded during the dry season, coinciding with transboundary haze episodes originating from neighbouring countries. These haze episodes significantly contribute to elevated PM_{10} levels, particularly in the Centre and North regions. The monsoon phenomenon further influences the monthly variation trend of PM_{10}, with lower concentrations typically observed during the wet season. This seasonal pattern underscores the importance of considering temporal factors in air quality management. The dry season's higher PM_{10} levels are likely due to increased biomass burning and reduced rainfall, which limits the natural dispersion of pollutants.

4.2.2 $PM_{2.5}$ Concentration

The K-means clustering analysis revealed eight distinct groups, each characterized by specific patterns of $PM_{2.5}$ concentrations observed across the Borneo, Centre, East Coast, North, and South regions. Notably, Cluster 1 exhibited the highest $PM_{2.5}$ concentrations in the Centre and North regions, with mean values of 18.62 µg/m³ and 20.14 µg/m³, respectively. In contrast, Cluster 3 showed the lowest concentrations, particularly in the Borneo region, with a mean value of 9.52 µg/m³. Additionally, Cluster 8 was identified as another group showing specific regional patterns, contributing to the overall clustering structure with a CCC value of 1.30657 (Figure 4b). The clustering results reveal significant regional variations in $PM_{2.5}$ concentrations, with the Centre and North regions consistently exhibiting higher levels. This suggests that these regions are more prone to higher pollution levels, likely due to their dense urbanization and industrial activities

Further analysis of the clusters indicates that the East Coast and South regions also experienced elevated $PM_{2.5}$ levels, though not as high as the Centre and North regions. For instance, Cluster 4, which had relatively high $PM_{2.5}$ concentrations, showed mean values of 19.87 µg/m³ in the Centre and 21.34 µg/m³ in the North. The Borneo region, characterized by lower industrial activity and population density, consistently showed the lowest $PM_{2.5}$ concentrations across all clusters. This regional disparity highlights the significant impact of urbanization and industrialization on air quality. The lower $PM_{2.5}$ levels in Borneo can be attributed to its extensive forest cover and lower human activity, which act as natural filters for air pollutants.

The seasonal variation in $PM_{2.5}$ levels was evident, with higher concentrations recorded during the dry season, coinciding with transboundary haze episodes originating from neighboring countries. These haze episodes significantly contribute to elevated $PM_{2.5}$ levels, particularly in the Centre and North regions. The monsoon phenomenon further influences the monthly variation trend of $PM_{2.5}$, with lower concentrations typically observed during the wet season. This seasonal pattern underscores the importance of considering temporal factors in air quality management. The dry season's higher $PM_{2.5}$ levels are likely due to increased biomass burning and reduced rainfall, which limits the natural dispersion of pollutants.

Additionally, the clustering analysis provides valuable insights into the spatial and temporal patterns of $PM_{2.5}$ concentrations in Malaysia. The identification of distinct clusters allows for a more nuanced understanding of regional air quality dynamics. These findings underscore the need for targeted air quality management strategies that consider both regional and seasonal variations. Effective policies and measures, such as stricter emission controls and regional cooperation to address transboundary pollution, are essential to mitigate the impact of $PM_{2.5}$ on public health and the environment. The results also suggest that localized interventions, tailored to the specific pollution sources and seasonal patterns of each region, could be more effective in improving air quality.

The findings from the k-means clustering analysis highlight significant regional variations in $PM_{2.5}$ concentrations across Malaysia. The Centre and North regions consistently exhibited higher $PM_{2.5}$ levels, which can be attributed to several factors, including higher population density, industrial activities, and vehicular emissions. These regions are known for their urbanization and industrialization, which contribute to elevated $PM_{2.5}$. The Borneo region, on the other hand, showed the lowest $PM_{2.5}$ concentrations. This can be attributed to its lower population density and industrial activity, which results in fewer sources of $PM_{2.5}$ emissions. The findings align with previous studies that highlight the impact of urbanization and industrialization on air quality.

Seasonal variations also play a crucial role in influencing $PM_{2.5}$ levels. The highest concentrations were recorded during the dry season, coinciding with transboundary haze episodes originating from neighbouring countries. These haze episodes significantly contribute to elevated $PM_{2.5}$ levels, particularly in the Centre and North regions. The monsoon phenomenon further influences the monthly variation trend of $PM_{2.5}$, with lower concentrations typically observed during the wet season. This seasonal pattern underscores the importance of considering temporal factors in air quality management. The dry season's higher $PM_{2.5}$ levels are likely due to increased biomass burning and reduced rainfall, which limits the natural dispersion of pollutants.

4.2.3 CO Concentration

The K-means clustering algorithm was utilized to discern patterns in the monthly-averaged carbon monoxide (CO) concentrations across five regions in Malaysia, which are Borneo, Centre, East, North, and South. The optimal number of clusters was determined to be five (K = 5), as indicated by the highest CCC = 0.34465, suggesting a moderate clustering structure (Figure 5a). The clustering results revealed significant spatial and temporal variability in CO concentration trends across Malaysia. Cluster 2 emerged as the largest group with 13 data points, while Clusters 1 and 5 contained only a single point each, indicating distinct outliers or unique temporal events. Clusters 3 and 4, with 11 and 15 data points respectively, represented more stable groupings of monthly-averaged values. The cluster means highlighted differentiated regional contributions to CO concentration levels. For instance, Cluster 5 exhibited the highest mean CO concentration across all regions, with particularly elevated values in the Centre (1.1193), North (0.8404), and South (0.8576) regions. This suggests that during the months corresponding to this cluster, Malaysia experienced generally high CO levels likely attributable to regional-scale pollution events or transboundary haze episodes.

Conversely, Cluster 4 displayed the lowest regional mean values, especially in the East (0.4177) and North (0.4726), indicating relatively cleaner air quality during the months represented by this cluster. Cluster 3 reflected intermediate values, suggesting moderate pollution levels. The Borneo region maintained relatively stable CO concentration means across clusters (ranging from

0.4966 to 0.9022), implying less fluctuation in CO pollution, potentially due to its lower population density and limited industrial activities compared to Peninsular Malaysia.

The scatterplot matrix visualizes the relationships among regional CO concentrations, coloured by cluster membership. Strong positive correlations are observable between most regional pairs, notably Centre–North and North–South, confirming the hypothesis of spatial coherence in pollution patterns, possibly due to similar meteorological and emission characteristics during certain months (Dominick et al., 2015). The clustering results provide critical insights into the spatial-temporal dynamics of CO pollution in Malaysia. The formation of distinct clusters highlights that while CO concentrations fluctuate across regions, there are identifiable periods with shared characteristics, likely influenced by seasonal factors such as monsoon transition periods, dry seasons, and anthropogenic activities like biomass burning and increased vehicular emissions (Yin et al., 2020). Notably, the detection of extreme clusters (Clusters 1 and 5) may correspond to episodic events such as haze occurrences caused by forest fires in Kalimantan or Sumatra. These transboundary pollution episodes are well-documented in the region and have been shown to significantly elevate ambient CO and particulate matter concentrations.

The high values in the Centre and North regions, particularly in Cluster 5, support the notion that urbanized and industrial zones contribute disproportionately to CO emissions. These identify Klang Valley and adjacent urban corridors as pollution hotspots due to dense traffic and industrial activities. Moreover, the relatively low variability in Borneo further validates the impact of urbanization on CO emissions. Sabah and Sarawak, while experiencing some degree of pollution due to biomass burning, maintain lower baseline CO concentrations in comparison to Peninsular Malaysia. The observed correlation among regions in certain clusters suggests a need for synchronized air quality management policies across regional boundaries.

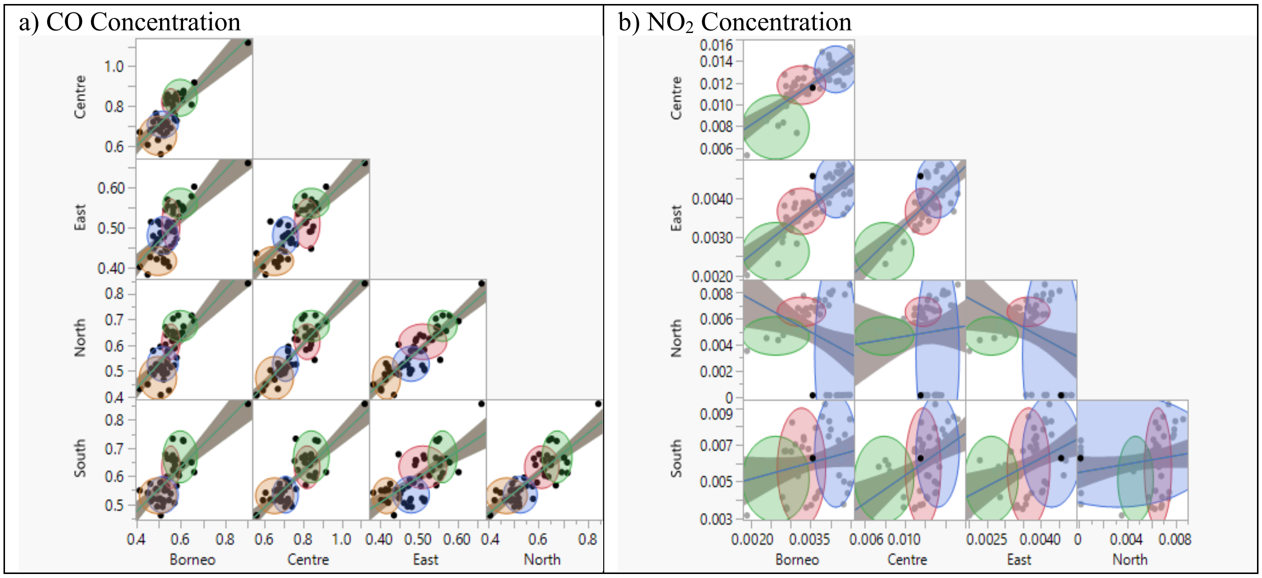

Figure 5: Scatter plot matrix illustrating regional variations in monthly a) CO concentration and b) NO_2 concentrations across Malaysia's major zones

4.2.4 NO_2 Concentration

The monthly NO_2 dataset, consisting of 48 observations from January 2018 to December 2021, was also grouped into three clusters using the K-means method (Fig. 5b). The monthly analysis showed a similar spatial pattern to the daily analysis, confirming that the regional differences in NO_2 concentrations persisted after daily fluctuations were averaged.

Cluster 1 (n = 12) represented a pattern with relatively higher concentrations in Borneo (0.0058 ppm) and the East Coast (0.0058 ppm), while the North region recorded the lowest mean concentration (0.0005 ppm). Cluster 2 (n = 15) was characterised by moderate NO_2 concentrations, including 0.0112 ppm in the Centre region, 0.0063 ppm in the North region, and 0.0053 ppm in the South region. Cluster 3 (n = 21), which was the largest monthly cluster, represented the relatively higher NO_2 condition, with the Centre region recording a mean concentration of 0.0148 ppm. Higher concentrations were also observed in the North (0.0077 ppm) and South (0.0076 ppm) regions within this cluster.

The monthly clustering pattern was more compact than the daily clustering pattern, suggesting that monthly averaging reduced short-term concentration variability. Nevertheless, the Centre region remained the dominant higher-concentration region in all clusters, with monthly mean values ranging from 0.0105 to 0.0148 ppm. This finding indicates that the higher NO_2 pattern in the Centre region was not limited to isolated daily episodes but persisted over the monthly timescale.

The results demonstrate that the three-cluster solution provides a meaningful representation of recurring monthly NO_2 concentration patterns across Malaysia. All concentrations were interpreted using the corrected original NO_2 values in ppm. Therefore, no negative NO_2 concentrations were reported or interpreted in the revised analysis.

4.3 Identification of High-Pollution Regimes and Potential Exposure Risk

The classification of pollutant concentration patterns through clustering analysis provides an important framework for identifying high-pollution regimes that may pose significant exposure risks to urban populations. Across the five study regions, clusters characterised by elevated concentrations of PM_{10}, $PM_{2.5}$, CO, and NO_2 were consistently associated with highly urbanised zones, particularly within the Centre and North regions of Peninsular Malaysia. These findings suggest that periods classified under high-concentration clusters may correspond to intensified anthropogenic emission activities, including vehicular congestion, industrial combustion processes, and transboundary biomass burning events.

In addition to spatial variability, the temporal occurrence of high-pollution clusters indicates the presence of episodic atmospheric conditions that may significantly elevate short-term exposure levels. Such pollution episodes are frequently linked to reduced atmospheric dispersion during dry monsoon periods or meteorological stagnation events, which facilitate the accumulation of airborne pollutants near the surface. The persistence of these clusters within urbanised regions underscores the potential for increased population exposure to harmful air pollutants, particularly in densely populated metropolitan areas such as the Klang Valley.

Furthermore, the identification of moderate and transitional clusters suggests fluctuating atmospheric states that may precede the onset of severe pollution episodes. Monitoring these transitional regimes may offer valuable insights for the development of early-warning systems aimed at mitigating public health impacts associated with short-term air pollution exposure. From a policy perspective, the ability to distinguish between background, moderate, and extreme pollution regimes enables more targeted intervention strategies that account for regional emission characteristics and seasonal variability. Overall, the clustering-based identification of pollution regimes enhances the understanding of spatial–temporal exposure patterns in Malaysia, thereby supporting the implementation of adaptive air quality management measures in rapidly urbanising regions.

5.0 Descriptive Statistical Distribution of Pollutants

This study examined the spatial–temporal variability of key air pollutants, namely PM_{10}, $PM_{2.5}$, CO, and NO_2, across five major regions of Malaysia using a K-means clustering framework. The descriptive statistical analysis revealed substantial regional heterogeneity in pollutant concentrations, with the Centre and North regions consistently recording higher mean values compared to the East and Borneo regions. These disparities are likely influenced by differences in urbanisation intensity, industrial development, population density, and regional emission sources.

The clustering analysis further demonstrated the presence of distinct pollution regimes across both daily- and monthly-averaged concentration datasets. High-concentration clusters were predominantly associated with urbanised zones, indicating the influence of transportation activities, industrial emissions, and seasonal biomass burning events on ambient air quality. In contrast, lower concentration clusters were typically observed in less industrialised regions such as Borneo, reflecting the role of land use and reduced anthropogenic activity in shaping baseline pollutant levels. Transitional clusters, representing moderate pollution conditions, were also identified, suggesting the occurrence of fluctuating atmospheric states influenced by seasonal meteorological variability.

Importantly, the temporal distribution of these clusters indicates that severe pollution episodes are episodic rather than persistent, often occurring under unfavourable atmospheric conditions such as reduced rainfall or weak wind circulation during dry monsoon periods. These findings highlight the potential for short-term exposure risks in densely populated urban environments, particularly in regions experiencing rapid infrastructure development and increased vehicular dependency.

From a methodological perspective, the application of unsupervised machine learning techniques such as K-means clustering provides an effective approach for segmenting large-scale environmental datasets into meaningful pollution regimes. This enables improved interpretation of spatial–temporal pollutant dynamics beyond conventional trend-based analyses and supports the identification of high-exposure episodes that may not be readily observable through aggregated statistical summaries.

In terms of policy implications, the identification of region-specific pollution regimes underscores the necessity for adaptive and localised air quality management strategies. Targeted mitigation measures, including enhanced vehicle emission control, industrial regulation, and transboundary haze monitoring, may be prioritised in regions classified under high-pollution clusters. Furthermore, the detection of moderate and transitional pollution regimes presents an opportunity for the development of early-warning systems aimed at reducing population exposure during episodic pollution events.

Overall, the findings of this study contribute to a data-driven understanding of atmospheric pollution dynamics in Malaysia and provide empirical support for the formulation of regionally tailored environmental policies. Integrating clustering-based analytical frameworks into national air quality monitoring systems may enhance the effectiveness of mitigation strategies and support sustainable urban development initiatives aligned with the objectives of Sustainable Development Goal (SDG) 11.

Acknowledgement

This study is supported by a Universiti Sains Malaysia, Short-Term Grant with Project No: R501-LR-RND002-0000001135-0000 and academic resources that contributed to the completion of this review. The authors gratefully acknowledge the Malaysian Department of Environment for providing the surface air pollutant data used in this study.

Conflicts of Interest

The author declares that there is no conflict of interest regarding the publication of this paper.

Reference

Abdullah, S., Rozali, N. H. S. C., Abu, A., Mansor, A. N. A., Ahmed, A. N., Dom, N. C., ... & Ismail, M. (2023). Inconsistency Of Air Quality Level During Several Phases Of Covid-19 Movement Restrictions in Malaysia. *Journal of Sustainability Science and Management*, 18(11), 147-162.

Ariff, N. M., Bakar, M. A. A., & Lim, H. Y. (2023). Prediction of PM_{10} concentration in Malaysia using k-means clustering and LSTM hybrid model. *Atmosphere, 14*(5), 853. https://doi.org/10.3390/atmos14050853

DOE., (2022). Air Management. Department of Environment Malaysia. Retrieved from https://www.doe.gov.my/en/air-management/

Dominick, D., Juahir, H., Latif, M. T., Zain, S. M., & Aris, A. Z. (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric Environment*, 60, 172–181. https://doi.org/10.1016/j.atmosenv.2012.06.021

Dominick, D., Latif, M. T., Juneng, L., Khan, M. F., Amil, N., Mead, M. I., ... & Pyle, J. A. (2015). Characterisation of particle mass and number concentration on the east coast of the Malaysian Peninsula during the northeast monsoon. *Atmospheric Environment*, 117, 187-199.

Gaveau, D. L. A., et al. (2014). Four decades of forest persistence, clearance and logging on Borneo. PLoS ONE, 9(7), e101654. https://doi.org/10.1371/journal.pone.0101654

Health Effects Institute. (2022). *State of Global Air 2022*. Boston, MA: Health Effects Institute. https://www.stateofglobalair.org

Latif, M. T., Dominick, D., Hawari, N. S. S. L., Mohtar, A. A. A., & Othman, M. (2020). The concentration of major air pollutants during the movement control order due to the COVID-19 pandemic in the Klang Valley, Malaysia. *Sustainable Cities and Society*, 66, 102660.

Mohd Nadzir, M. S., Ooi, M. C. G., Alhasa, K. M., Bakar, M. A. A., Mohtar, A. A. A., Nor, M. F. F. M., Latif, M. T., Hamid, H. H. A., Ali, S. H. M., Ariff, N. M., Anuar, J., Ahamad, F., Azhari, A., Hanif, N. M., Subhi, M. A., Othman, M., & Nor, M. Z. M. (2020). The impact of Movement Control Order (MCO) during pandemic COVID-19 on local air quality in an urban area of Klang Valley, Malaysia. *Aerosol and Air Quality Research, 20*(6), 1237–1248. https://doi.org/10.4209/aaqr.2020.04.0163

Mohd Shafie, S. H., Mahmud, M., Mohamad, S., Rameli, N. L. F., & Abdullah, R. (2022). Influence of urban air pollution on the population in the Klang Valley, Malaysia: A spatial approach. Ecological Processes, 11(42). https://doi.org/10.1186/s13717-021-00342-0

Naidin, S., Sentian, J., Adnan, F.A.F., Herman, F., Mohd Hashim, S.R. (2024). Decade-Long Analysis: Unravelling the Spatio-Temporal Dynamics of PM10 Concentrations in Malaysian Borneo. Aerosol Air Qual. Res. 24, 230176. https://doi.org/10.4209/aaqr.230176

Shakor, A. S. A., Pahrol, M. A., & Mazeli, M. I. (2020). Effects of population weighting on PM_{10} concentration estimation. *Journal of Environmental and Public Health, 2020*, 1561823. https://doi.org/10.1155/2020/1561823

World Health Organization. (2021). *WHO global air quality guidelines: Particulate matter ($PM_{2.5}$ and PM_{10}), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*. Geneva: World Health Organization. https://www.who.int/europe/publications/i/item/9789240034228

Yin, P., Brauer, M., Cohen, A. J., Wang, H., Li, J., Burnett, R. T., Liu, J., Liu, Y., Liang, X., Wang, L., Zhou, M., & Murray, C. J. L. (2020). The effect of air pollution on deaths, disease burden, and life expectancy across China and its provinces, 1990–2017: An analysis for the Global Burden of Disease Study 2017. *The Lancet Planetary Health, 4*(9), e386–e398. https://doi.org/10.1016/S2542-5196(20)30161-3

Zoran, M. A., Savastru, R. S., Savastru, D. M., & Tautan, M. N. (2023). Peculiar weather patterns effects on air pollution and COVID-19 spread in Tokyo metropolis. *Environmental Research, 228*, 115907. https://doi.org/10.1016/j.envres.2023.115907