MJFAS
Malaysian Journal of Fundamental and Applied Sciences

RESEARCH ARTICLE

A Framework to Spatially Cluster Air Quality Monitoring Stations in Peninsular Malaysia using the Hybrid Clustering Method

Nurul Alia Azizan[a], Ahmad Syibli Othman[a], Asheila AK Meramat[b] , Siti Noor Syuhada Muhammad Amin[c], Azman Azid[d*]

[a]School of Biomedical Science, Faculty of Health Sciences, Universiti Sultan Zainal Abidin, Gong Badak Campus, 21300 Kuala Nerus, Terengganu, Malaysia; [b]Department of Basic Medical Science, Faculty of Medicine & Health Science, Universiti Malaysia Sarawak, 94300 Kota Samarahan, Sarawak, Malaysia; [c]Universiti Sultan Zainal Abidin Science and Medicine Foundation Centre, Gong Badak Campus, 21300 Kuala Nerus, Terengganu, Malaysia; [d]School of Animal Science, Aquatic Science and Environment, Faculty of Bioresources and Food Industry, Universiti Sultan Zainal Abidin, Kampus Besut, 22200 Besut, Terengganu, Malaysia

Abstract Multiple variables must be analyzed in order to assess air quality trends. It turns into a multidimensional issue that calls for dynamic methods. In order to provide an improved spatial cluster distribution with distinct validation, this study set out to illustrate the hybrid cluster method in air quality monitoring stations in Peninsular Malaysia. The Department of Environment, Malaysia (DOE), provided the data set, which covered the two-year period from 2018 to 2019. This study included six air quality pollutants: PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3, and CO. Principal component analysis (PCA), a multivariate technique, was used to condense the information found in enormous data tables in order to better comprehend the variables (to reduce dimensionality) prior to grouping the data. The PCA factor scores were then used to produce the AHC. The clusters were validated using discriminant analysis (DA). 36 of 47 stations required additional analysis using AHC, according to the PCA factor scores. Low Polluted Region (LPR = seven stations), Moderate Polluted Region (MPR = 20 stations), and High Polluted Region (HPR = nine stations) were created from AHC and share the same characteristics. The DA results showed 84 % correct classification rate for the clusters. With regard to identifying and categorizing stations according to air quality characteristics, the framework presented here offers an improved method. This illustrates that the hybrid cluster method utilized in this work can produce a new method of pollutant distributions that is helpful in air pollution investigations.

Keywords: Hybrid clustering method, air quality, multivariate technique, Peninsular Malaysia.

***For correspondence:**
azmanazid@unisza.edu.my

Received: 14 April 2022
Accepted: 21 Sept. 2023

Introduction

With the economic and technological development of cities, environmental pollution problems are arising, such as water, noise and air pollution. Particularly, air pollution is growing in importance as a global environmental problem since poor air quality can have a negative impact on people's health, the environment, and national economies. Any material of any sources within the atmosphere could be exists in particulate matter (PM_{10}, $PM_{2.5}$ and Ultra Fine Particulate (UFP)), sulphur dioxide (SO_2), nitrogen dioxide (NO_2), ozone (O_3), carbon monoxide (CO), heavy metals and volatile organic compounds (VOCs). Both industrialised and developing nations are concerned about the quality of the air they breathe in today's world. The primary reasons of the reduction in air quality may include both anthropogenic (such as emissions from industry and cars) and natural (such as volcanic emissions and

10.11113/mjfas.v19n5.2620

forest fires as transboundary pollution) sources [1]. Studies showed evidence that air pollution could be a potential role in neurological diseases [2, 3] as well as in cognitive impairment [4-6]. Moreover, air pollution showed an association in diabetes mellitus prevalence in Malaysia [7]. Recent studies in school-aged children have also demonstrated an elevated risk for conditions including autism [8], respiratory problems [9, 10], and asthma [11, 12] that are associated to genotoxicity brought by air pollution [13]. With the significance health impacts from the polluted air, therefore air quality index becoming the main indicator towards human health status. Air Pollution Index (API) is providing easily comprehensible information about the air pollution indicator. Since 1989, the Malaysian Department of Environment (DOE) has embraced API as a crucial tool to educate the public about air quality, potential health consequences, and other environmental concerns [14]. The bigger the API number, the more hazardous the air is to human health. According to API values, the air quality is divided into five categories: good, moderate, unhealthy, very unhealthy, and hazardous, with values corresponding to 0 to 50, 51 to 100, 101 to 200, 201 to 300, and greater than 300. For instance, an API score of 50 indicates good air quality, whereas an API value over 300 indicates hazardous air quality [15]. In the middle of year 2017, DOE has upgraded the index calculation by using six pollutants parameter instead of five parameters. $PM_{2.5}$ has been introduced as one of the pollutants to be incorporated in the calculation index [16]. The Recommended Malaysia Ambient Air Quality Guideline (RMAQG) has been replaced by a New Ambient Air Quality Standard (NAAQS). This new standard incorporates six criteria for air pollutants, including five existing ones: PM_{10}, SO_2, CO, NO_2, and O_3. Additionally, it includes an additional parameter, $PM_{2.5}$. The implementation of the new standard includes interim targets such as IT-1 in 2015 and IT-2 in 2018, with full implementation scheduled for 2020.

Due to DOE's programmes on air pollution in Malaysia, it is essential for Malaysia, a growing nation, to have an effective system for monitoring air quality. Chemometric techniques, sometimes referred to as multivariate analysis, are one of the most up-to-date and trustworthy statistical methods used by researchers to examine massive volumes of data. It is founded on the statistical principle, which calls for monitoring and assessing several variables at once while keeping the workload manageable. Because they can prevent incorrect interpretation of results, these strategies are the best ones to employ when applying to a significant amount of complicated environmental monitoring data [17]. It has been demonstrated to be a more effective tool for analysing air quality than conventional statistical methods, such as for spatial variations, which offers an understanding of the key trends and underlying relationships in data, for contamination sources identification, data reduction, and interpretation [18,19]. These strategies offer better processing and interpretation of air quality data as well as effective management of air quality monitoring programmes by minimising database complexity. Numerous scientific investigations [14,19-23] have utilised principal component analysis (PCA), agglomerative hierarchical cluster analysis (AHC), and discriminant analysis (DA), particularly in the monitoring of air quality. The formulation of suitable plans for the efficient administration of air quality monitoring programmes is made possible by the application of these techniques for decoding challenging databases, which improves our understanding of the air quality in the research area [24].

Clustering is an exploratory data analysis technique that examines the data's underlying structure. K-means and AHC, two well-known and commonly utilised techniques, have been used in air pollution research since the 1980s and have attracted a lot of attention [25]. AHC analysis is a technique for categorising items into clusters in which the objects (monitoring stations) inside a cluster are similar to one another while objects in other clusters are distinct [26]. Characterization of the spatial variation of air quality parameters can deliver an enhanced understanding of the ecological circumstance and aid strategy producers to plan needs for practical air quality administration. The level of air quality is dictated by measured air pollutants. Numerous studies have been done on these techniques, such as an evaluation of $PM_{2.5}$ in Malaysia based on spatial cluster analysis [14], a study on spatial $PM_{2.5}$ using k-means cluster analysis [27], a study on the classification of significant pollutants using AHC [28, 29], and a study using cluster analysis to determine the pattern of air quality in Klang Valley [1]. However, several multivariate statistical approaches, including agglomerative hierarchical cluster analysis (AHC), discriminant analysis (DA), principal component analysis (PCA), and factor analysis (FA), were used to analyse and reveal significant information from huge, complex data about air quality studies.

Using data gathered between 2018 and 2019, the hybrid clustering method (PCA-AHC) was used in this study to classify sites throughout Peninsular Malaysia according to air quality pollutants. The study's goals are to determine whether this newly developed approach will enable a better comprehension of the heterogeneity in air quality pollutants. This shows that the hybrid methodology used in this work can produce better pollutant distributions that are helpful in investigations of air pollution. We hope that the identified clusters can be used to further investigate the heterogeneity in the relationship between air pollutants concentration of the sampling sites and morbidity across the Peninsular Malaysia.

Materials and Methods

Study Area

Malaysia has a total land area of 329,960.22 km^2. Peninsular Malaysia (West Malaysia) and Borneo, which included the states of Sabah and Sarawak, were the two primary landmasses known as East Malaysia. Peninsular Malaysia covers a total area of 131,798 km^2 [30] and is divided into 12 states, with a total of 47 air monitoring stations in the National Continuous Air Quality Monitoring Network (NCQMN), out of a total of 65 in Malaysia (Figure 1). The northeast and southwest monsoons, which occur from November to March and May to September respectively, are the two monsoon seasons in Malaysia. The northeast monsoon occasionally provides significant rain, whereas the southwest monsoon resulting in less rainfall at this time.

The Malaysian Department of Environment (DOE) categorised air monitoring stations according to land use characteristics. One issue was to avoid being too close to residential areas in order to protect human health. Industrial estates and large traffic volumes could be the sources of pollution. All of Peninsular Malaysia's air quality monitoring sites (47 stations) were selected to provide a comprehensive picture of the region's air quality. These 47 monitoring locations are supervised and managed by a commercial company (Pakar Scieno TW Sdn. Bhd.) on behalf of the DOE. The majority of air monitoring stations are found in suburban, urban and industrial settings. Based on their latitude and longitude, Figure 1 depicts the location of the study area.

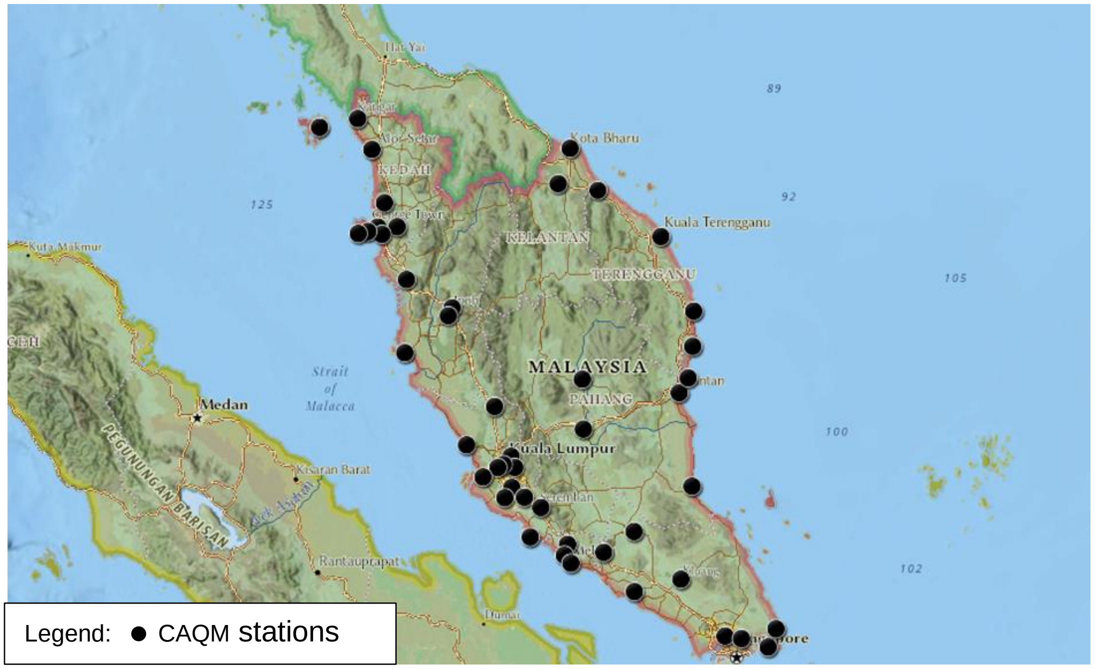

Figure 1. Location of Continuous Air Quality Monitoring (CAQM) stations in Peninsular Malaysia

Frame of Data

Peninsular Malaysia's air quality was measured at 47 stations located throughout the country (Figure 1). Data for the air pollutants PM$_{10}$, PM$_{2.5}$, SO$_2$, NO$_2$, O$_3$, and CO as well as meteorological variables wind direction, wind speed, temperature, humidity, and solar radiation were obtained from the Air Quality Division of the Department of the Environment (DOE) and monitored and collected by the DOE-authorized agency named Pakar Scieno TW Sdn. Bhd. The source of the data was a monthly average calculated from hourly monitoring locations.

Eight characteristics, including factors related to air pollutants, were initially gathered for this investigation. Due to the significant amount of missing data for two parameters, NO and NOx, only six out of eight parameters were chosen for further study. Before being provided to the stakeholders, the data generated by Pakar Scieno TW Sdn. Bhd. shall be checked and verified by the DOE in its capacity as an authority. Before being submitted to the DOE, all air quality data from the Continuous Air Quality

Monitoring Network (CAQM) goes through QA/QC protocols. Every two weeks, gas detection devices are manually inspected, and PM_{10} and $PM_{2.5}$ instruments are calibrated once a month in accordance with standard operating procedures. Data deletion (which results in negative values) is triggered by insufficient data, while outliers activate second-level QC tests. Some of the findings were supported by observable evidence, while others were ruled out due to instrument failure. Data from air monitoring stations is sent to the DOE Environmental Data Centre (EDC) in Putrajaya, where it is subjected to quality assurance and quality control (QA/QC). The data was collected hourly at each station and subjected to QA/QC procedures to confirm its accuracy. PM_{10} and $PM_{2.5}$ were measured using a Thermo Scientific tapered element oscillating microbalance (TEOM) 1405-DF (USA) analyzer; CO and O_3 were measured using a Thermo Scientific Model 48i (USA) CO analyzer and a Thermo Scientific Model 49i (USA) O_3 analyzer; SO_2 was measured using a Thermo Scientific Model 43i (USA) SO_2 analyzer; and NO_2 was measured using a Thermo Scientific Model 42i (USA) NO_2 analyzer.

Data Pretreatment

Pretreatment data were applied to the raw data in order to assure the accuracy of the data being studied. As several parameters had a high percentage of missing values, this included resolving the issue of missing data. Furthermore, data transformation was used to standardise the range of each parameter and normalisation was carried out to remove magnitude discrepancies. Due to the observation of different ranges for dependent and independent values, the normalisation procedure was required. These pretreatments ensured that no dominant parameters were present in the dataset. As a result, before being exposed to additional analysis, each parameter underwent pre-processing through normalisation to equalise their magnitudes.

Kaiser Meyer Olkin (KMO) and Bartlett's Tests

To determine whether the data are appropriate for the subsequent analysis using the chosen parameters, the Kaiser Meyer Olkin (KMO) and Bartlett's tests are conducted at the beginning of a multivariate study. The Bartlett's test looks for associated parameters in the study, while the KMO test assesses how well the data have been factored. Table 1 provides the guidelines for evaluating the KMO test results. Prior to extracting factors in the procedure known as principal component analysis (PCA), the appropriateness of the samples is confirmed by computing the Measure of Sampling Appropriateness (MSA) using the KMO value, which should range between 0.60 and 1.00 [32].

Table 1 Rules of guidance for interpreting results of KMO test

Value of KMO	Interpretation
0.80-1.00	Adequate
0.70-0.79	Middling
0.60-0.69	Mediocre
0.50-0.59	Not adequate
0.00-0.49	Unacceptable

Multivariate Analysis

Multivariate techniques are outstanding tools that are frequently used in the field of environment to recognise the spatial variation. Three strategies for reaching the goal were determined by this study. Using the XLSTAT software (XLSTAT, 2019, Addinsoft, New York, NY, USA), principal component analysis (PCA), agglomerative hierarchical cluster analysis (AHC), and discriminant analysis (DA) were performed. We were able to identify a pattern of air quality in Peninsular Malaysia via new approach using this hybrid clustering method.

Principal Component Analysis (PCA)

PCA, one of the most used and useful statistical approaches for identifying the possible structure of a group of variables [33], can be used to reduce the dimensions of a large data set. With the aim to summarise the content of big data, PCA was applied [34]. The links between observations and variables, as well as among the variables, may be revealed by this overview [35]. The most common application of PCA is to represent a multivariate data table as a smaller number of variables (summary indices). Therefore, trends, clusters and outliers may be observed. PCA was calculated based on equation 1 below:

$PC_i = l_{1i} X_1 + l_{2i} X_2 + \ldots + l_{ni} X_n$ [36]
where PC*i* is define as *i*th principal component, *lji* is define as variable loading and *Xj* is define as observed variable.

Prior to grouping the data, PCA was utilised in this study to obtain a clearer image of the variables. It is believed to enhance clustering outcomes (noise reduction). By condensing a large number of connected variables into a smaller set of uncorrelated (independent) variables known as principle components (PCs), this method was used to explain the variance of a large number of connected variables [22]. This may aid in determining which explanatory variables have the most impact on the dependent variable. By doing so, only the most important PCA factors were used as input parameters for the specific type of region. The consistent variables were construed by PCA to recognize the source of air pollution and the most significant parameters in this study. Result from factor score PCA then used to clustering the data.

Hierarchical Agglomerative Cluster Analysis (AHC)

An unsupervised statistical technique called AHC is used to group or cluster observations according to how similar or dissimilar they are. This spatial classification of air quality monitoring stations can be shown using a dendrogram that assesses the degree of risk homogeneity using Ward's method and Euclidean distance measurement [36]. The Euclidean distance is calculated using the ratio of the linkage distance divided by the maximal distance (D_{link}/D_{max}), which is multiplied by 100 to standardise the linkage distance represented by the y-axis [37]. In this study, the PCA and AHC were combined with the intention of generating a better clustering method.

Discriminant Analysis (DA)

DA is typically used to find the factors that best discriminate between the AHC groups and to assist in the construction of new discriminant functions (DFs) for each group in order to assess the regional variation in atmospheric air quality. Equation 2 is used to determine DFs:
$F(Gi) = Ki + \sum^{n}{}_{j}=1\; w_{ij}\;\; Pij$ [1]
where, *i* is the number of group *G*; *kj* is constant inherent to each group; *n* is the number of parameters used to classify a set of data into a given group; w_j is the weight coefficient assigned by discriminant function analysis (DFA) to a given parameter *Pj*.

The three clusters created by AHC using standard mode were employed in this study using DA [1]. Based on another study, DA also had been used to predict group membership [21]. Air pollutants (PM_{10}, $PM_{2.5}$, SO_2, NO_2, O_3, and CO) were chosen as the dependent variables and Clusters 1, 2, and 3 as the independent variables.

Results and Discussion

Normality Test

The normality of the data was assessed using the Jarque-Bera test [38]. This statistical test determines whether the series follows a normal distribution. The analysis revealed that the data did not exhibit normal distribution, as evidenced by a p-value (<0.0001) lower than the significance level alpha (0.05). As a result, the alternative hypothesis (H_a) was accepted rather than the null hypothesis (H_0), showing that the extracted variable did not follow a normal distribution. Data transformation was used to solve this problem and stop any variable from controlling performance [39].

Kaiser-Meyer-Olkin (KMO) and Bartlett’s Tests

KMO and Bartlett's tests were run before the analysis to determine whether the data were appropriate for PCA and to assess its sufficiency [18]. All metrics in this study were used for PCA, with the exception of NO and NOx, which had a significant amount of missing data. Although a small percentage of missing data can be handled using PCA [40], a high number of missing data can cause the analysis to produce incorrect conclusions. Furthermore, PCA is susceptible to missing data brought on by a lack of data for particular parameters [41].

The KMO test was used to evaluate the samples' suitability, which may have been affected by underlying causes [42]. The KMO sample adequacy measure is shown in Table 2. For the KMO test, a value larger than 0.5 was chosen as the reference point [43]. As a result of the KMO values exceeding 0.5, which indicate adequate data for PCA extraction, the findings obtained demonstrated that the measure of sampling adequacy (MSA) was acceptable.

Table 2. Kaiser-Meyer-Olkin measure of sampling adequacy

PM_{10}	0.537
$PM_{2.5}$	0.541
SO_2	0.829
NO_2	0.625
O_3	0.720
CO	0.679
KMO	**0.612**

The Bartlett's sphericity test was performed in addition to KMO to assess the parameters' association with the suitability of the data for PCA construction. With a significance level of 0.05, the test result revealed a p-value less than 0.0001. As a result, the data were eligible for multivariate analysis because there was enough information to perform PCA and a significant correlation between the parameters.

As the p-value was lower than the significance level of 0.05, the null hypothesis (H_0) was rejected in favour of the alternative hypothesis (H_a). This means that at least one of the correlations between parameters was significantly different from zero. Consequently, the air quality parameters were found to be correlated and not orthogonal, allowing for a diverse interpretation of the data's variability.

Air Quality Pattern using the Hybrid Clustering Method (PCA-AHC)

There are 47 air monitoring stations in Peninsular Malaysia and six variables measured to be calculated as API. DOE categorized according to urban, sub-urban, industrial and rural settings in order to easily monitor any implications towards human health and environmental. In this study, six air pollutants had been used in PCA as independent variables and 47 air monitoring stations as dependent variables. Results showed Table 3 and the corresponding chart are both related to a mathematical entity known as eigenvalues, which reflect the quality of the projection from the N-dimensional original table (N=6) without any changes in the number of dimensions. As stated below, the first eigenvalue equals 2.6 and represents 43.3 % of the total variability. Each eigenvalue corresponds to a factor and each factor to a one dimension. Each eigenvalue is corresponding to a factor and each factor is associated with a single dimension. A factor is a linear combination of the initial variables with no correlation between them (r=0). The eigenvalues and corresponding factors are arranged by how much of the initial variability they reflect in descending order.

Table 3 Eigenvalues

	F1	F2	F3	F4	F5	F6
Eigenvalue	2.600	1.684	0.723	0.656	0.267	0.069
Variability (%)	43.337	28.071	12.047	10.940	4.457	1.147
Cumulative %	43.337	71.408	83.455	94.396	98.853	100.000

The first two or three eigenvalues should ideally correspond to a large percentage of the variance, ensuring that the maps based on the first two or three components are a high-quality projection of the original multi-dimensional table. As seen in Figure 2, the first two factors account for 71.4 % of the initial variability of the data, even though there are six factors, indicating that the variables are significant in determining air quality. This is because all six variables have been automatically determined as having useful dimensions.

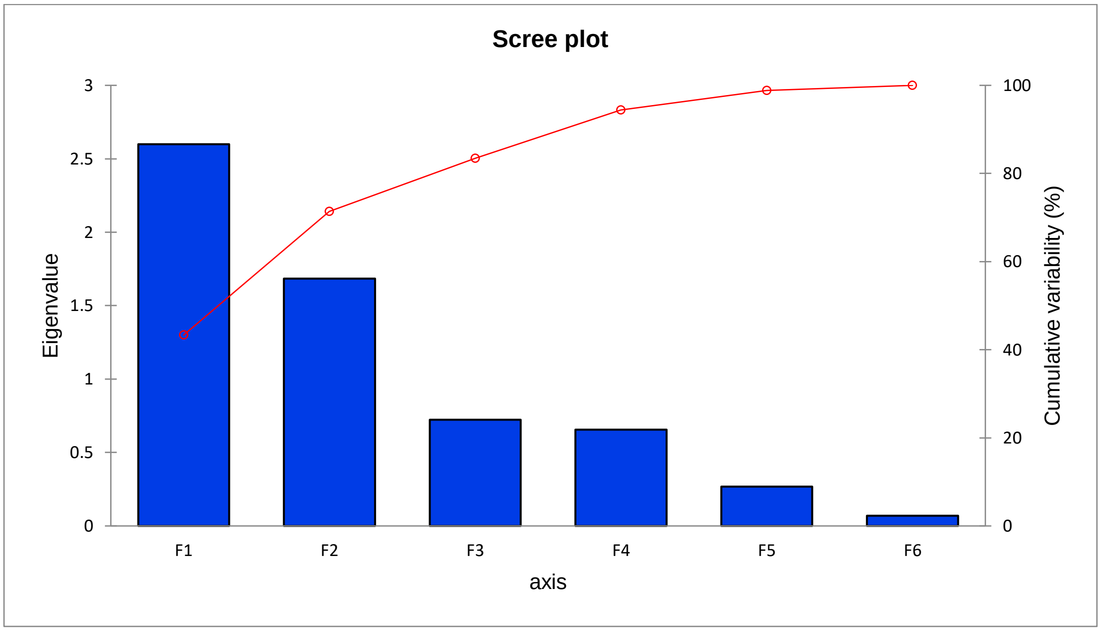

Figure 2 Scree plot of eigenvalue vs cumulative variability (%)

The factor score PCA was able to summarise all 47 stations with six variables into 36 stations as the significant observation outcome since all six variables were found to be significant in connection to air quality. PCA was typically employed [1, 23, 33] to pinpoint potential sources of variation. In order to pinpoint emission sources, the extra factors known as principal components (PCs) were later examined. In contrast, PCA was employed in this study to evaluate data where observations were described by a number of connected quantitative dependent variables. The most significant data is extracted by PCA analysis and displayed in a new space as a set of orthogonal and linear variables known as principle components (sometimes referred to as factors) (F1+F2....+Fn), where n is the total number of variables. Each variable or observation on each primary component is represented geometrically by a factor score or factor loading. PCA relationships can be seen using a number of methods, such as bi-plots, scree plots, and correlation plots. The squared cosine is used to describe the component's contribution to the squared distance of a variable or observation from the origin [46]. A component's variability as well as overall variability are significantly influenced by variables or observations with larger squared cosines [47]. We are still searching for more relevant data that shows how these two multivariate approaches are used in the classification of air quality. In our research, PCA was used to convert air quality observations into factorial axes and then examine the relationships between the observations and the variables. To clarify the most crucial contributing observations to air quality variability, we used PCA in particular. Finally, all of the stations in Peninsular Malaysia's stations were sorted by our analysis into the most significant observations with the most significant variables.

The trends in Peninsular Malaysia's air quality were then investigated using AHC and the summary indices using factor score PCA. It is based on each variable identified in the PCA findings for the years 2018 and 2019, as well as the monthly variation for the period of January through December, which pertains to the chosen data for that year. Classes were the main subject of the analysis, with emphasis on the newly formed cluster. AHC was used to study the air quality pattern, which is shown in Figure 3 as a dendrogram. It illustrates how the algorithm groups the observations and then breaks them down into smaller groups. The algorithm successfully grouped each observation. The monthly average air quality data from 36 monitoring sites was analysed using AHC analysis based on factor scores from prior PCA analysis in this study. This section uses AHC to classify the most significant air quality stations according to the homogeneity level of each station by looking at the historical values of each of the six air pollutants separately. The dendrogram's AHC results show how the clusters identified as High Pollution Regions (HPR), Medium Pollution Regions (MPR), and Low Pollution Regions (LPR) are distinct from one another. The profile clusters were grouped into categories based on the average value (mean) of the pollutant variables. The clusters with the highest average values were labelled as HPR, clusters with medium values were categorized as MPR, and clusters with lower values were identified as LPR, as shown in Table 4.

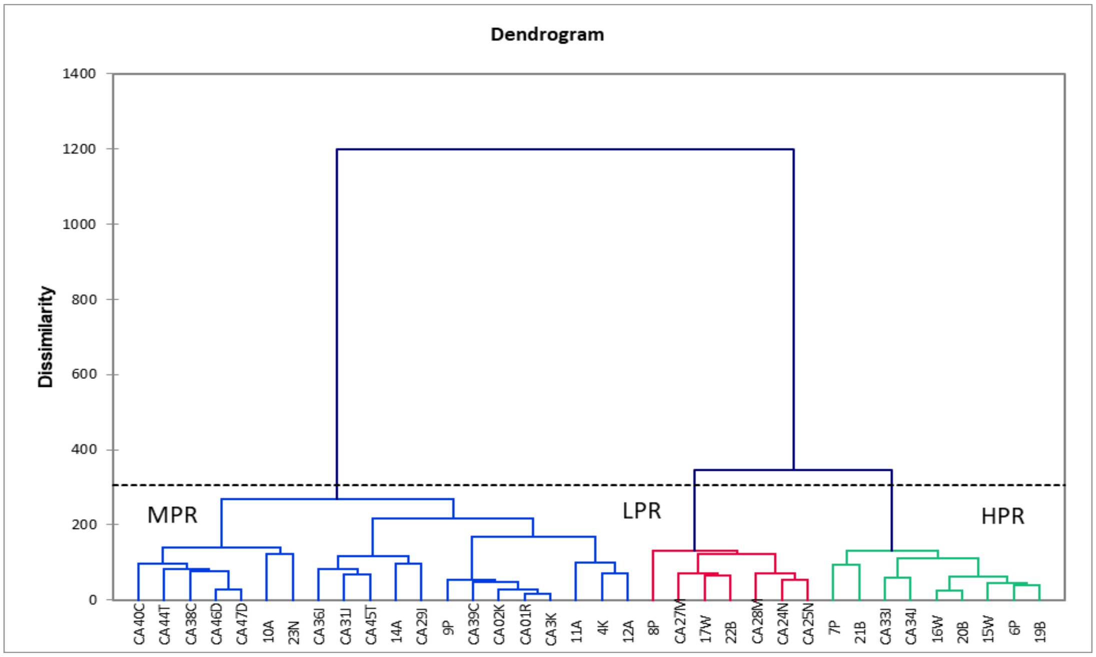

Figure 3 Dendrogram representing various groups of air quality monitoring stations

Figure 4 shows the classification of stations in Malaysia using AHC (HPR, MPR, and LPR) based on six air pollution concentrations. There are nine stations in total for HPR, twenty for MPR, and seven for LPR. The central region of Peninsular Malaysia, as well as portions of the northern and southern regions, are where the majority of HPR stations are situated.

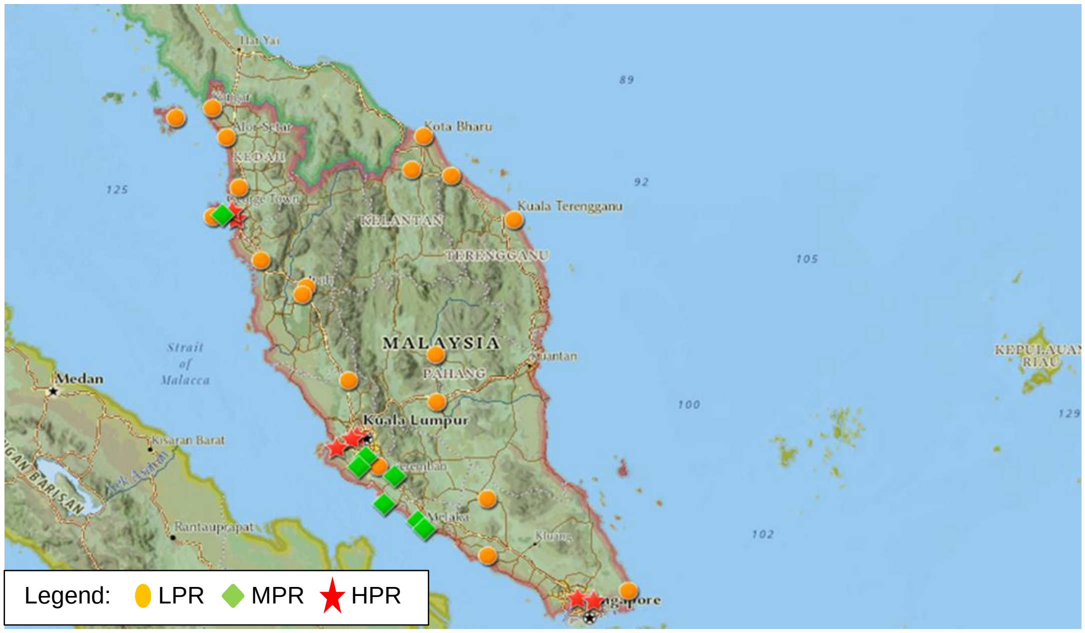

Figure 4 Classification of stations using a hybrid clustering method based on the average monthly concentrations of six air pollutants

Three clusters were created using the clustering algorithm, and the stations in these clusters all have the same homogeneity criteria. This study reached a level where grouping groups are already homogenous but are heterogenous among themselves as shown in Table 4. This study found that average values for both particulate matters (PM) are above the NAAQS for the specific two years period. PM_{10} concentration showed slightly above the guideline as reported in mean concentrations at all three clusters (HPR, MPR, LPR). It is more worrying in $PM_{2.5}$ concentration, where it is reported to contribute higher mean concentration at all clusters and above the limitation set by the NAAQS. There was a significant disparity between the average and maximum values. However, SO_2, NO_2, O_3 and CO average value showed in LPR, MPR and HPR stayed in a good condition compared to PM where under the limit.

Table 4 Descriptive analysis of six parameters based on cluster of air monitoring stations

		Parameters					
Station ID	Statistic items	PM_{10} MAX (µg/m3)	$PM_{2.5}$ MAX (µg/m3)	SO_2 MAX (ppm)	NO_2 MAX (ppm)	O_3 MAX (ppm)	CO MAX (ppm)
Cluster 1 MPR	Minimum	24.633	19.397	0.001	0.007	0.029	0.649
	Maximum	1113.107	1112.491	0.034	0.056	0.143	4.808
	Median	89.718	71.527	0.003	0.021	0.061	1.473
	Mean	**116.997**	**92.643**	**0.004**	**0.023**	**0.062**	**1.550**
	Standard deviation (n-1)	100.696	84.936	0.004	0.008	0.018	0.508
Cluster 2 HPR	Minimum	49.960	41.411	0.001	0.025	0.036	1.027
	Maximum	1453.981	1353.878	0.061	0.097	0.156	5.410
	Median	96.596	79.284	0.007	0.049	0.083	2.641
	Mean	**126.084**	**107.512**	**0.009**	**0.049**	**0.084**	**2.744**
	Standard deviation (n-1)	137.459	130.712	0.007	0.012	0.021	0.710
Cluster 3 LPR	Minimum	33.965	25.154	0.002	0.014	0.048	0.955
	Maximum	405.660	384.394	0.042	0.054	0.136	5.267
	Median	89.712	74.109	0.011	0.032	0.083	1.613
	Mean	**108.859**	**91.283**	**0.012**	**0.032**	**0.084**	**1.698**
	Standard deviation (n-1)	59.164	55.161	0.007	0.007	0.016	0.522
New Ambient Air Quality Standard (NAAQS)		100	35	0.04	0.17	0.10	9.00

Discriminating Classes using Discriminant Analysis (DA)

Further analysis by using DA, which exposed dissimilarities within the study sites was implemented in view of the clustering obtained from the AHC for HPR, MPR and LPR. Here, the class of clusters were considered as the independent parameters, while the air pollutants were considered as the dependent parameters.

Table 5 Discriminant analysis for all classes (standard)

Sampling stations	Cluster 1-MPR	Cluster 2-HPR	Cluster 3-LPR	Total	% Correct
Cluster 1-MPR	451	8	21	480	93.96%
Cluster 2-HPR	25	177	14	216	81.94%
Cluster 3-LPR	62	9	96	167	57.49%
Total	538	194	131	863	83.89%

Table 5 shows the classification matrix of DA, which involved three types of clusters, namely Cluster 1-MPR, Cluster 2-HPR and Cluster 3-LPR. Based on the table, there were 480 numbers of data from the Cluster 1-MPR with the percentage accuracy of 93.96 %, 216 numbers of data from the Cluster 2-HPR with the percentage accuracy of 81.94 % and 167 numbers of data from the Cluster 3-LPR with the percentage accuracy of 57.49 %. In general, the regions discriminated well with an average of 83.89 % correct classification.

When compared to another study [30] that used a standard clustering technique, the current study's percentage of correct classification of each cluster was substantially higher. While accuracy in a prior study was 63.94 % of the average percentage of correct classification, accuracy in the present study was 83.89 %. It is showing that through hybrid clustering method, percentage of correct classification either in each cluster and total clusters could be improved and have a better percentage rather than using a routinely method for clustering analysis. However, there were few studies done previously using normal clustering method were able to perform good percentage of accuracy [19, 21, 33]. Their findings showed more than 90 % of accuracy for spatial variation. However, there was a limitation from previous study because only selected stations in small numbers were chosen to examine the spatial variation rather than the present study. This new approach was able to include higher numbers of observations rather than involved selected stations in order to reduce error or avoid any bias from the data selection. Hence, this approach will improve the observation/station representation in this study. Study by Ab. Rahman *et al*. (2022) also performed AHC in relation to determine the spatial variation on $PM_{2.5.}$ However, there was no validation of the classification of the groups were done to verify the percentage of correct classification.

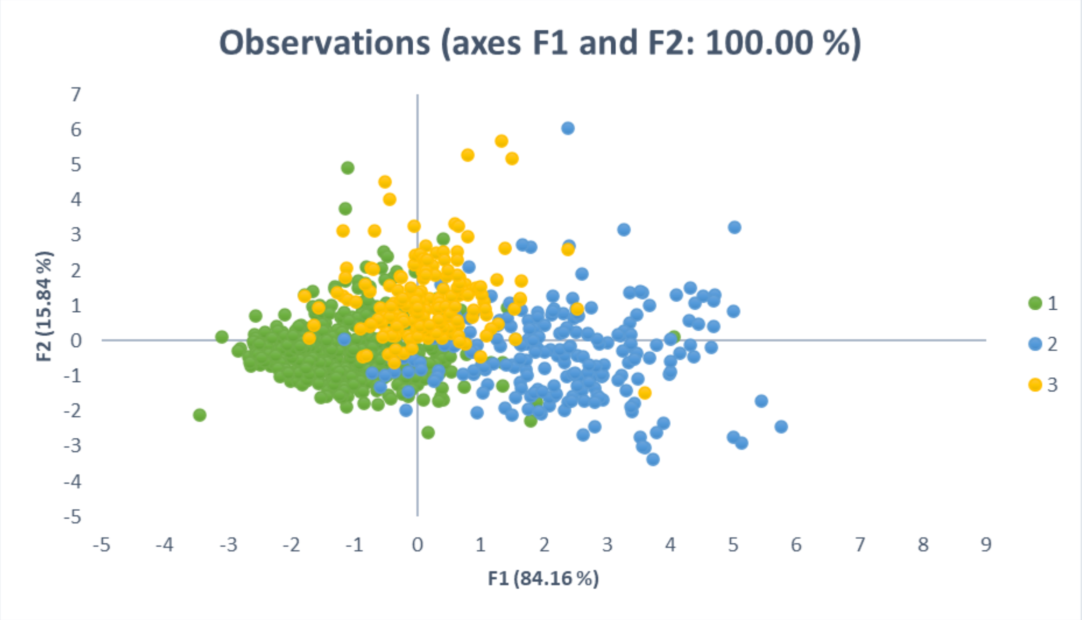

Figure 5 Observation dots based on facts at three classes

Clusters one and two were more dominating in terms of colour in Figure 5. When compared to cluster three, this is due to the large percentage of correct classification. Cluster three had fewer observation dots and was yellow in colour. The study discovered a lambda value of 0.257, refer $p<0.0001$ and $\alpha = 0.05$, using the Wilks lambda test.
The hypothesis is stated as below:
H_0: The mean vectors of three clusters are equal
H_a: At least one of the means vector is different from another

When $p<\alpha$ is calculated, the null hypothesis (H_0) should be rejected and the alternative hypothesis should be accepted (H_a). While H_0 is accurate, the chance of rejecting it is less than 0.01 %. As a result, each cluster was distinct from the others. Following the study, DA was able to discriminate all variables into three clusters: one, two, and three, which represent the total number of air monitoring stations in Peninsular Malaysia.

Conclusions

Based on the findings, the hybrid clustering method was able to evaluate the patterns of air monitoring. The architecture presented here offers an alternative method for locating and categorizing stations according to air quality parameters. It is particularly crucial for integrating PCA with the clustering technique. Three unique air quality clusters were produced by AHC. According to the current study, hybrid clustering (PCA, AHC) is an improved approach for similar iterations when taking into account the connection link that binds an object to one class. The cluster classifications were also validated using discriminant analysis (DA), and the findings indicated that they were reliable variables in Peninsular Malaysia. This suggests that the hybrid clustering strategy used in this work is capable of developing new pollutant distribution approaches that are helpful in air pollution investigations.

Conflicts of Interest

The author(s) declare(s) that this study was carried out without any self-interest, commercial or financial conflicts and they state that they have no competing interests with the funders.

Acknowledgment

This work is part of a research project, UniSZA/2021/DPU1.0/20. N.A. Azizan is supported by Skim Latihan Akademik IPTA-SLAI (Ministry of Higher Education, Malaysia) and Universiti Sultan Zainal Abidin. We also thank Department of Environment of Malaysia for their support.

References

[1] Sahrir, S., *et al.* (2019). *Environmetric study on air quality pattern for assessment in Klang Valley, Malaysia. International Journal of Recent Technology and Engineering (IJRTE), 8*, 17-24.

[2] Jankowska-Kieltyka, M., A. Roman, and I. Nalepa. (2021). The air we breathe: air pollution as a prevalent proinflammatory stimulus contributing to neurodegeneration*. Frontiers in Cellular Neuroscience, 15*, 647643-647643.

[3] Bandyopadhyay, A. (2016). Neurological disorders from ambient (urban) air pollution emphasizing UFPM and PM2.5. *Current Pollution Reports, 2*(3), 203-211.

[4] Chen, M.-C., *et al*. 2021. Air pollution is associated with poor cognitive function in Taiwanese adults*.* International Journal of Environmental Research and Public Health, *18*(1), 316.

[5] Iaccarino, L., *et al*. (2021). *Association between ambient air pollution and amyloid positron emission tomography positivity in older adults with cognitive impairment*. *JAMA Neurology, 78*(2), 197-207.

[6] Shehab, M. A. and F. D. Pope. (2019). Effects of short-term exposure to particulate matter air pollution on cognitive performance*. Sci Rep., 9*(1), 8237.

[7] Wong, S. F., *et al.* (2020). Association between long-term exposure to ambient air pollution and prevalence of diabetes mellitus among Malaysian adults*. Environ Health*, *19*(1), 37.

[8] Volk, H. E., *et al.* (2013). Traffic-related air pollution, particulate matter, and autism*. JAMA Psychiatry, 70*(1), 71-77.

[9] Arifuddin, A. A., Jalaludin, J. and Hisamuddin, N. H. (2019). Air pollutants exposure with respiratory symptoms and lung function among primary school children nearby heavy traffic area in Kajang*. Asian J. Atmos. Environ*., *13*, 21-29.

[10] Tellez-Rojo, M. M., *et al.* (2020). Children's acute respiratory symptoms associated with PM2.5 estimates in two sequential representative surveys from the Mexico City Metropolitan Area. *Environ Res*, *180*, 108868.

[11] Zainal Abidin, E., *et al.* (2014). The relationship between air pollution and asthma in Malaysian schoolchildren. *Air Quality, Atmosphere & Health, 7*(4), 421-432.

[12] Zakaria, J., M.s. Lye, and Z. Hashim. (2012). Asthma severity and environmental health risk factor among asthmatic primary school children in the selected areas. *American Journal of Applied Sciences, 9*, 1553-1560.

[13] Hisamuddin, N. H., *et al.* (2022). The influence of environmental polycyclic aromatic hydrocarbons (PAHs) exposure on DNA damage among school children in urban traffic area, *Malaysia. International Journal of Environmental Research and Public Health*, *19*(4).

[14] Ab. Rahman, E., *et al*. (2022). Assessment of PM2.5 patterns in Malaysia using the clustering method. *Aerosol and Air Quality Research*, *22*(1), 210161.

[15] Abd Rani, N. L., *et al.* (2018). Air pollution index trend analysis in Malaysia, 2010-15. *Polish Journal of Environmental Studies*, *27*(2), 801-807.

[16] Department of Environment (DOE). (2020). New Malaysia ambient air quality standard department of environment*.* Accessed March. http://www.doe.gov.my/portalv1/en/category/info-umum/indeks-pencemaran-udara.

[17] Shafii, N. Z., *et al*. (2019). Application of chemometrics techniques to solve environmental issues in Malaysia*. Heliyon, 5*(10), e02534.

[18] Azid, A., *et al*. (2015). Identification source of variation on regional impact of air quality pattern using chemometric. *Aerosol and Air Quality Research*, *15*(4), 1545-1558.

[19] Azizan, N. A., *et al.* (2022). Air quality pattern in Central of Malaysia: A new approach*. Bioscience Research, 19*(SI-1), 126-134.

[20] Zakaria, U., *et al.* (2017). The assessment of ambient air pollution pattern in Shah Alam, Selangor, Malaysia. *Journal of Fundamental and Applied Sciences, 9*, 772-788.

[21] Azid, A., *et al*. (2017). Air quality modelling using chemometric techniques*. Journal of Fundamental and Applied Sciences, 9*, 443-466.

[22] Dominick, D., *et al*. (2012). Spatial assessment of air quality patterns in Malaysia using multivariate analysis*. Atmospheric Environment, 60*, 172-181.

[23] Liyana Zakri, N., *et al*. (2018). Identification source of variation on regional impact of air quality pattern using chemometric techniques in Kuching, Sarawak*. International Journal of Engineering & Technology, 7*(3): Special Issue 14.

[24] Azid, A., *et al*. (2014). Prediction of the level of air pollution using principal component analysis and artificial neural network techniques: A case study in Malaysia. *Water, Air, & Soil Pollution, 225*(8), 2063.

[25] Govender, P. and V. Sivakumar. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research, 11*(1), 40-56.

[26] Pires, J. C. M., *et al.* (2008). Management of air quality monitoring using principal component and cluster analysis—Part I: SO_2 and PM_{10}*. Atmospheric Environment, 42*(6), 1249-1260.

[27] Austin, E., *et al.* (2013). A framework to spatially cluster air pollution monitoring sites in US based on the $PM_{2.5}$ composition. *Environment International, 59*, 244-254.

[28] Azid, A., *et al.* (2016). Selection of the most significant variables of air pollutants using sensitivity analysis*. Journal of Testing and Evaluation, 44*, 1-9.

[29] Azizan, N. A., *et al.* (2022). Identification of the most significant of air pollutants using sensitivity analysis with spatial assessment using clustering method*. Bioscience Research, 9*(SI-1), 105-114.

[30] Ismail, A., A. Abdullah, and M. Samah. (2017). Environmetric study on air quality pattern for assessment in Northern Region of Peninsular Malaysia*. Journal of Environmental Science and Technology, 10*, 186-196.

[31] Zheng, D., *et al*. (2018). Prediction and sensitivity analysis of long-term skid resistance of epoxy asphalt mixture based on GA-BP neural network*. Construction and Building Materials*, *158,* 614-623.

[32] Shrestha, N. (2021). Factor analysis as a tool for survey analysis*. American Journal of Applied Mathematics and Statistics, 9*(1), 4-11.

[33] Hua, A. (2018). Applied chemometric approach in identification sources of air quality pattern in Selangor, Malaysia. *Sains Malaysiana*, *47*, 471-479.

[34] Azid, A., *et al*. (2018). Assessing indoor air quality using chemometric models*. Polish Journal of Environmental Studies, 27*.

[35] Shafii, N., *et al*. (2017). Spatial assessment on ambient air quality status: A case study in Klang, Selangor*. Journal of Fundamental and Applied Sciences*, *9*, 964-977.

[36] Jamalani, M., *et al*. (2016). Monthly analysis of PM_{10} in ambient air of Klang Valley, Malaysia. *Malaysian Journal of Analytical Sciences.*

[37] Hamza Ahmad, I. and A. Azman. (2015). Air quality pattern assessment in Malaysia using multivariate techniques*. Malaysian Journal of Analytical Sciences*, *19*(5), 966-978.

[38] Kim, N. (2016). A robustified Jarque–Bera test for multivariate normality*. Economics Letters*, *140*, 48-52.

[39] Lee, D. (2020). Data transformation: A focus on the interpretation*. Korean Journal of Anesthesiology*, *73*, 503-508.

[40] van Ginkel, J. R., L. A. van der Ark, W. H. M. Emons, and R. R. Meijer. (2023). Handling missing data in principal component analysis using multiple imputation, *in Essays on Contemporary Psychometrics*. Springer International Publishing. 141-161.

[41] Franceschi, F., M. Cobo, and M. Figueredo. (2018). Discovering relationships and forecasting PM_{10} and $PM_{2.5}$ concentrations in Bogotá, Colombia, using artificial neural networks, principal component analysis, and k-means clustering*. Atmospheric Pollution Research*, *9*(5), 912-922.

[42] Shrestha, S. and F. Kazama. (2007). Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan*. Environmental Modelling & Software, 22*(4), 464-475.

[43] Ul-Saufie, A. Z., *et al*. (2013). Future daily PM_{10} concentrations prediction by combining regression models and feedforward backpropagation models with principle component analysis (PCA). *Atmospheric Environment, 77*, 621-630.

[44] Abdullah, S., *et al*. (2016). Neural network fitting using Levenberg-Marquardt training algorithm for PM_{10} concentration forecasting in Kuala Terengganu*. Journal of Telecommunication,*

Electronic and Computer Engineering, *8*, 27-31.

[45] Li, G. (2007). Measuring the quality of life in City of Indianapolis by integration of remote sensing and census data*. International Journal of Remote Sensing - INT J REMOTE SENS, 28*.

[46] Kruskal, J. B. a. W., M. (1978). Multidimensional scaling. *Sage University Paper Series on Quantitative Applications in the Social Sciences.* Sage Publications. 07-011.

[47] Kamalha, E. and E. Omollo. (2017). Clustering and classification of cotton lint using principle component analysis, agglomerative hierarchical clustering, and K-means clustering*. Journal of Natural Fibers. 15*.