MJFAS
Malaysian Journal of Fundamental and Applied Sciences

Azizan *et al.* | Malaysian Journal of Fundamental and Applied Sciences, Vol. 19 (2023) 804-816

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.

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

Received: 14 April 2022
Accepted: 21 Sept. 2023

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

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

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.