Application of K-Means Clustering and Calendar View Visualisation for Air Pollution Index Analysis

Application of K-Means Clustering and Calendar View Visualisation for Air Pollution Index Analysis

Z Ali Omar[1], Siti Rahayu Mohd Hashim[2,4], Justin Sentian[3] and Su Na Chin[2]

[1]Mathematics with Computer Graphics Programme, Faculty of Science and Natural Recourses, Universiti Malaysia Sabah, Malaysia
[2]Mathematics with Economy Programme, Faculty of Science and Natural Recourses, Universiti Malaysia Sabah, Malaysia
[3]Environmental Science Programme, Faculty of Science and Natural Recourses, Universiti Malaysia Sabah, Malaysia

rahayu@ums.edu.my

Abstract. Two years of diurnal concentration of particulate matter (PM_{10}) and nitrogen dioxide with the addition of relative humidity measurement, collected from Putrajaya, Malaysia's ground-based measurement station from January 2014 to December 2015, were analysed. *K*-means clustering was employed and optimal clusters of four were identified for each year based on the most suggested number of clusters from internal cluster validation measures of the total within sum of square, silhouette index and gap statistics. Each cluster was then profiled where each mean pollutant sub-indices were calculated and the contributing pollutant to the air pollution index (API) was determined by looking at the maximum value from all sub-indices. This mechanism closely follows the Recommended Malaysian Air Quality Guidelines (RMG) for determining API. Particulate matter was found to be the dominant sub-index in all clusters and then paired with the mean relative humidity for visualisation. A calendar view was selected to show the temporal patterns and we observed a consistent cluster profile with the actual mean values of the selected parameters for most months. The calendar view also suggested that overall, the API (based on particulate matter) in 2014 was much better as compared to 2015.

1. Introduction

The application of clustering in air pollution analysis has already been done extensively since the 1980s [1]. The ability of clustering to find irregularities and similarities among the data points [2], makes clustering a process that can further classify the data points that can easily be summarised [1]. Among the clustering methods available, *k*-means is one of the most commonly used when it comes to ground-based air pollution measurements as well as clustering air mass trajectories [1]. Although the clustering methods were exploration methods in general [1, 2], it has been shown that clustering of air pollution measurement is useful for efficient pollution monitoring, identification of sources and operative and mitigation control strategies [1]. It explains why there are still recent studies done on using clustering to analyse air pollution data, such as in [3, 4, 5]. We also noticed that previous related studies were focusing on implementing or determining the best clustering methods to employ, as in [4] and [5].

In this paper, we focused more on the implementation of the selected clustering method, *k*-means, as it is computationally fast, easy, and simple, and able to handle huge datasets [1], and reported to be

more efficient [6]. We take it further by visualising the identified clusters, in a calendar view for easy temporal trend (monthly, yearly) assessment. The calendar visualisation has already been employed by [7] and [8] to show the pattern of univariate time series data. Our study used a multivariate time series dataset that included hourly concentration measurement of particulate matter (PM_{10}, which will be referred to *PM*), nitrogen dioxide (NO_2) and relative humidity (*RH*) for 365 days a year. This data was collected from the Putrajaya, Malaysia ground-based station from January 2014 to December 2015. Identifying the relationship between meteorological factors and pollutants has always been an interest as it affects the atmospheric pollutant concentrations and aerosol composition [9, 10, 11]

The main concern in conducting clustering is the determination of the number of *k* (cluster) [12, 13, 6, 2]. Therefore, in our study, we have considered this clustering issue as explained in further detail in the methodology sections along with data pre-processing. The identified clusters were then profiled based on the Air Pollution Index (API) following the mechanism in the Recommended Malaysian Air Quality Guideline (RMG) [14] governed by the Department of Environment (DOE) [15].

2. Methodology

2.1. Data preparation

The hourly averaged air quality and meteorological data were used in the clustering process as well as in the imputation of any missing observation data. In the case of this study, the missing values for pollutants such as NO_2 and meteorological parameters such as *RH* were observed at 27.4% and 14.8% for 2014 respectively. The missing values of pollutants in the following year were comparatively lower (i.e only 4.7% for NO_2). The observed missing values were then treated using the median of nearby points [16]. Due to the differences in the range of readings, the values were normalised by scaling them from 0 to 1 [2, 17]. This technique can also reduce the sensitivity of *k*-means to extreme values, which were retained. We also assumed that during the process, the technique would group the values to be in their cluster.

2.2. The clustering processes

The clustering was implemented yearly using *k*-means with Euclidean as the distance measurement [6, 13]. Since *k*-means is a partitional algorithm based on centroid, the data was separated by years to differentiate the centroid found in both years. The determination of the optimal number of clusters was based on three internal compactness measures using total within sum of square [13] (looking at the elbow), silhouette index [12, 13, 18] (taking the highest index as the number of clusters) and gap statistics [12, 19] (the first peak would suggest the number of the cluster). The most suggested number of clusters will be the optimal number and used to cluster the dataset.

2.2.1. API sub-index for PM and NO_2. The means for all the variables in a cluster were used to identify the description of the cluster. The means for *PM* and NO_2 were mapped to correspond to the API values as calculated based on the formulae in table 1 [14]. The API calculation is based on the 24-hour time-weighted average for *PM* concentration (table 1 a), and on a 1-hour time-weighted average concentration for NO_2 (table 1 b). The API status and pollution level are given in table 2 [14]. Once the main differentiating factors have been identified for the clusters, then the cluster classification will be plotted in a calendar view for visualisation.

Table 1. API sub-index functions calculation.

conc. < 50 μg/m^3	API = conc.
50 < conc. < 150	API = 50 + {[conc. – 50] x 0.5}
150 < conc. < 350	API = 100 + {[conc. – 150] x 0.5}
350 < conc. < 420	API = 200 + {[conc. – 350] x 1.4286}
420 < conc. < 500	API = 300 + {[conc. – 420] x 1.25}
conc. > 500 μg/m^3	API = 400 + [conc. – 500]

(a). PM concentration.

*conc. < 0.17 ppm	API = conc. x 588.23529
*0.17 < conc. 0.6	API = 100 + {[conc. – 0.17] x 232.56}
0.6 < conc. 1.2	API = 200 + {[conc. – 0.6] x 166.667}
Conc. > 1.2 ppm	API = 300 + {[conc. – 1.2] x 250}

(b). NO_2 concentration.

Table 2. API index status and how it may affect general health.

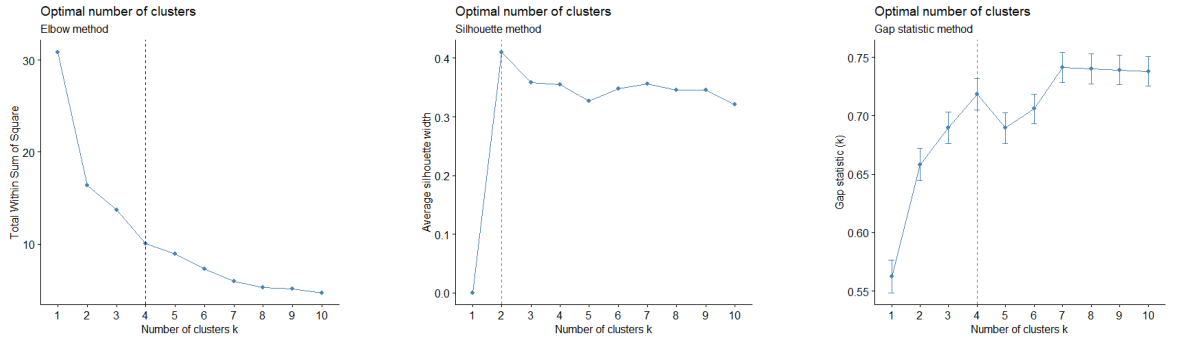
API	Status	Pollutant Level
0 – 50	Good	Pollutant low and has no ill effects on health
51 – 100	Moderate	Moderate pollution and has no ill effects on health
101 – 200	Unhealthy	Mild aggravation of symptoms among high-risk persons, i.e., those with heart or lung disease
201 – 300	Very Unhealthy	Significant aggravation of symptoms and decreases exercise tolerance in person with heart or lung disease
301 – 500	Hazardous	Severe aggravation of symptoms and endangers the health
Above 500	Emergency	Severe aggravation of symptoms and endangers the health

2.2.2. *Relative humidity (RH).* In general, Malaysia's monthly RH values varied greatly between 10% and 90%. Low relative humidity over the Malaysian Peninsular region, particularly in January and February, which coincides with the southwest monsoon and comparatively higher in November, which coincides with northeast monsoon, with some exceptions in some states on the east coast [20].

3. Results and discussion

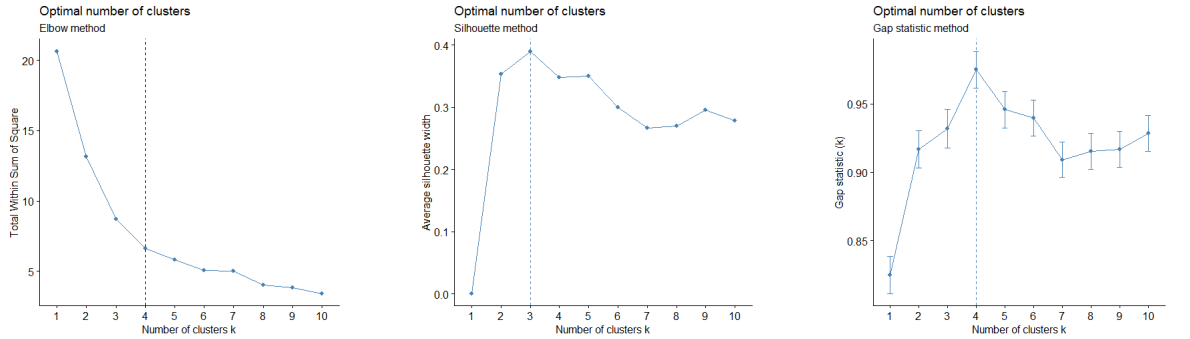
3.1. Optimal cluster

Based on the three methods to determine the optimal number of clusters (figure 1), clusters of four were selected as optimal (based on the total within sum of squares and the gap statistic method) for both the 2014 and 2015 datasets.



2014 dataset

(a). Elbow method k = 4 **(b).** Silhouette method k = 2 **(c).** Gap statistic method k = 4



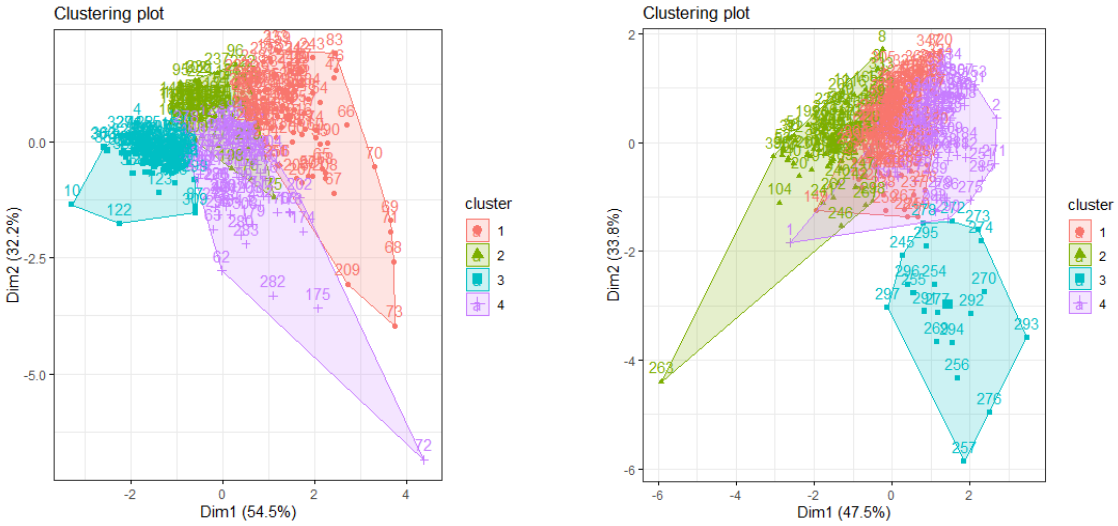
(d). Elbow method k = 4 **(e).** Silhouette method k = 3 **(f).** Gap statistic method k = 4

2015 dataset

Figure 1. Determining the optimal number of clusters for 2014 (a-c) and 2015 (d-f) based on the Elbow, Silhouette and Gap Statistic methods.

3.2. Cluster plot

The *k*-means clustering produced 4 clusters of sizes 104, 71, 104 and 86 for the 2014 dataset and 157, 91, 20, 97 for the 2015 dataset respectively. A significant reduction of samples in cluster 3 from 104 (2014) to 20 (2015) was observed. The grouping patterns of the clusters' objects were also found to be distinctive between these two years' datasets. For instance, the grouping for cluster 3 objects was at the top left in 2014 but had moved to the bottom right in 2015 (figure 2). This could indicate that the two years' grouping patterns are different, justifying the yearly separation of the datasets. For both years, the total variances of the dataset explained by the clusters were within the acceptable range of 67.6% and 67.9%, respectively.



2014 dataset 2015 dataset

Figure 2. Cluster grouping plots based on four clusters for 2014 and 2015 datasets.

3.2.1. *Summary of each variable by cluster.* The summary for each of the variables by cluster and year is given in tables 3 (*PM*), 4 (NO_2) and 5 (*RH*). The corresponding boxplot based on the summary for each variable by cluster and year is shown in figures 3, 4 and 5.

Table 3. Number summary for Particulate Matter by cluster and year.

	Particulate Matter (ug/m^3)							
	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	2014	2015	2014	2015	2014	2015	2014	2015
Min	13.42	23.23	15.42	18.58	15.23	141.4	15.00	28.58
1st Quartile	29.76	39.21	21.33	34.06	29.88	167.9	42.48	42.79
Median	44.19	47.62	26.52	41.50	34.85	209.3	57.31	52.00
Mean	52.64	53.27	31.16	46.06	36.02	211.0	64.76	60.17
3rd Quartile	59.56	60.33	33.90	52.78	42.46	242.6	75.52	67.50
Max	227.62	128.38	120.64	110.38	83.17	328.8	313.79	138.92

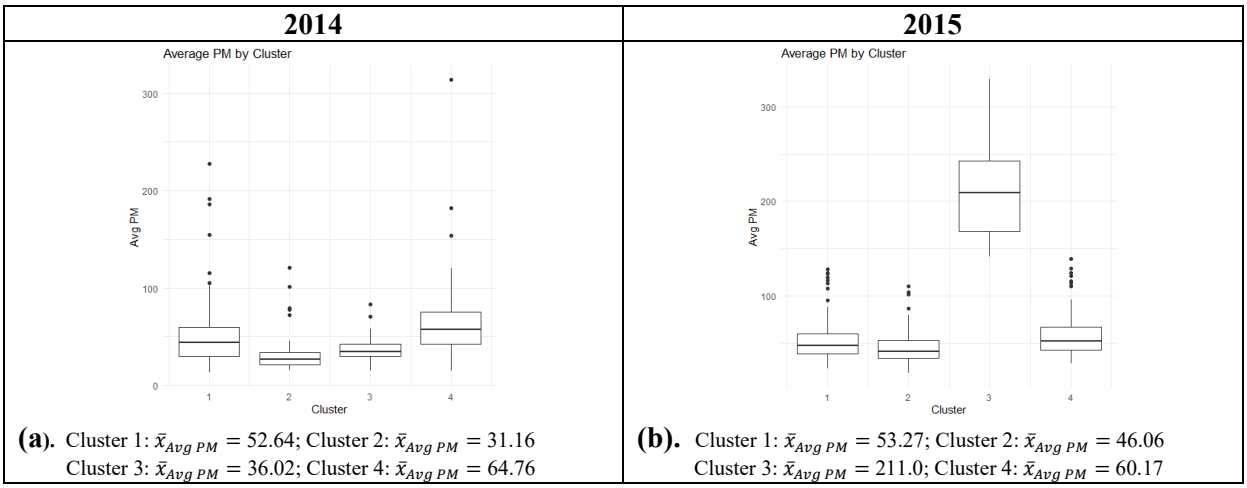
Table 4. Number summary for Nitrogen Dioxide by cluster and year.

	NO_2 (ppm)							
	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	2014	2015	2014	2015	2014	2015	2014	2015
Min	0.00363	0.01170	0.00774	0.00487	0.01730	0.01056	0.01371	0.01970
1st Quartile	0.00839	0.01552	0.01185	0.00902	0.02164	0.01661	0.01680	0.02139
Median	0.01011	0.01661	0.01396	0.01104	0.02164	0.01886	0.01898	0.02233
Mean	0.00999	0.01673	0.01381	0.01043	0.02180	0.01910	0.01917	0.02292
3rd Quartile	0.01187	0.01813	0.01570	0.01233	0.02164	0.02140	0.02164	0.02405
Max	0.01448	0.02000	0.01861	0.01543	0.03461	0.02648	0.02782	0.03440

Table 5. Number summary for Relative Humidity by cluster and year.

	Relative Humidity (%)							
	Cluster 1		Cluster 2		Cluster 3		Cluster 4	
	2014	2015	2014	2015	2014	2015	2014	2015
Min	59.12	54.83	72.79	15.50	75.21	68.71	63.08	43.17
1st Quartile	65.38	73.83	76.54	67.42	78.71	72.50	68.94	74.62
Median	68.67	76.67	78.71	71.17	81.17	74.96	70.88	78.67
Mean	68.17	77.08	78.68	70.81	81.55	75.73	70.53	77.81
3rd Quartile	72.04	79.62	80.85	73.96	84.36	78.91	72.52	81.83
Max	75.17	89.21	86.29	86.79	91.46	82.04	75.30	88.04

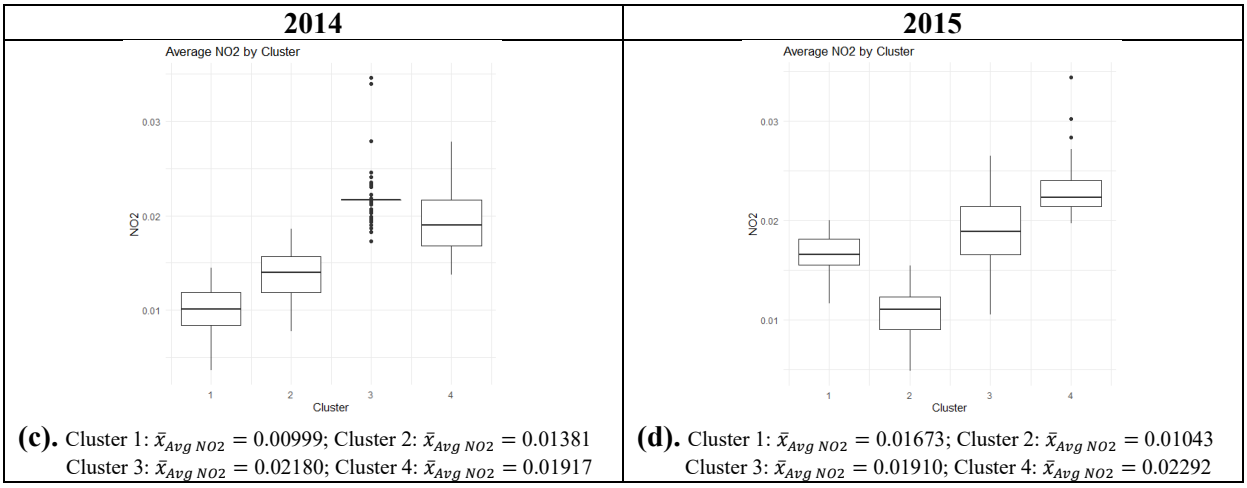
3.2.2. *The box plot for each variable by clusters.*



(a). Cluster 1: $\bar{x}_{Avg\ PM} = 52.64$; Cluster 2: $\bar{x}_{Avg\ PM} = 31.16$
Cluster 3: $\bar{x}_{Avg\ PM} = 36.02$; Cluster 4: $\bar{x}_{Avg\ PM} = 64.76$

(b). Cluster 1: $\bar{x}_{Avg\ PM} = 53.27$; Cluster 2: $\bar{x}_{Avg\ PM} = 46.06$
Cluster 3: $\bar{x}_{Avg\ PM} = 211.0$; Cluster 4: $\bar{x}_{Avg\ PM} = 60.17$

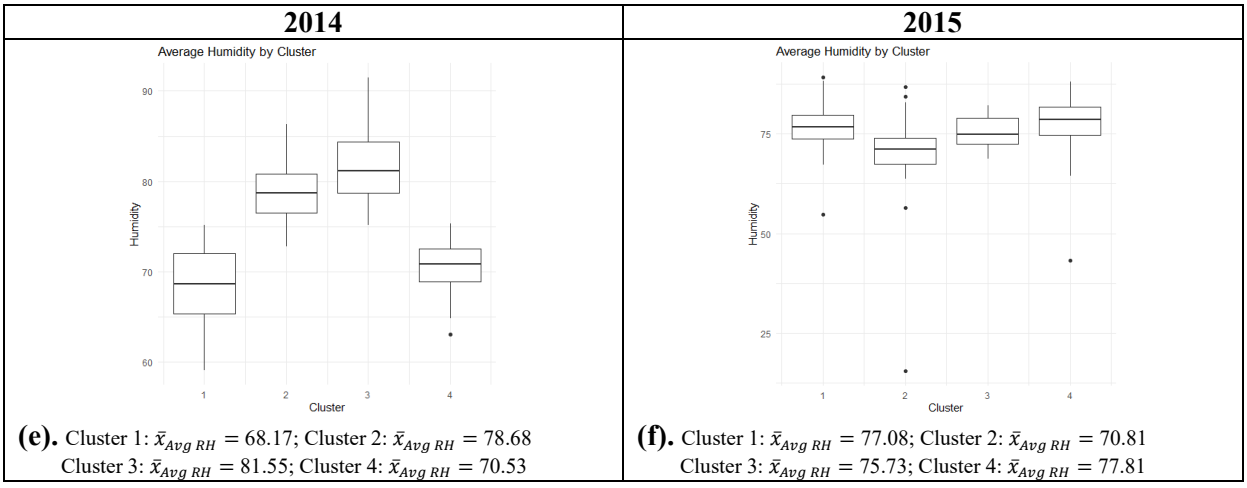
Figure 3. Box plots of Particulate Matter by cluster and year.



(c). Cluster 1: $\bar{x}_{Avg\ NO2} = 0.00999$; Cluster 2: $\bar{x}_{Avg\ NO2} = 0.01381$
Cluster 3: $\bar{x}_{Avg\ NO2} = 0.02180$; Cluster 4: $\bar{x}_{Avg\ NO2} = 0.01917$

(d). Cluster 1: $\bar{x}_{Avg\ NO2} = 0.01673$; Cluster 2: $\bar{x}_{Avg\ NO2} = 0.01043$
Cluster 3: $\bar{x}_{Avg\ NO2} = 0.01910$; Cluster 4: $\bar{x}_{Avg\ NO2} = 0.02292$

Figure 4. Box plots of Nitrogen Dioxide by cluster and year.



(e). Cluster 1: $\bar{x}_{Avg\ RH} = 68.17$; Cluster 2: $\bar{x}_{Avg\ RH} = 78.68$
Cluster 3: $\bar{x}_{Avg\ RH} = 81.55$; Cluster 4: $\bar{x}_{Avg\ RH} = 70.53$

(f). Cluster 1: $\bar{x}_{Avg\ RH} = 77.08$; Cluster 2: $\bar{x}_{Avg\ RH} = 70.81$
Cluster 3: $\bar{x}_{Avg\ RH} = 75.73$; Cluster 4: $\bar{x}_{Avg\ RH} = 77.81$

Figure 5. Box plots of Relative Humidity by cluster and year.

3.3. Cluster description based on API subindex

Based on the mean values for each sub-index in each cluster, the API was calculated for the *PM* and NO_2. The profiles for each cluster according to their API sub-index status and the mean *RH* are shown in table 6. The mean API for NO_2 was good for all clusters in both years, while the mean API for *PM* varied, and the values were larger than the NO_2. Therefore, the *PM* sub-index was chosen as the indicator for the overall API. This was also consistent with the findings of [9].

Table 6. Cluster description based on API sub-indices and meteorology status.

	Cluster 1		**Cluster 2**		**Cluster 3**		**Cluster 4**	
	2014	2015	2014	2015	2014	2015	2014	2015
API_{PM}	Moderate (51.32)	Moderate (51.63)	Good (31.16)	Good (46.00)	Good (36.02)	Unhealthy (130.5)	Moderate (57.38)	Moderate (55.09)
API_{NO2}	Good (5.88)	Good (9.84)	Good (11.06)	Good (6.14)	Good (12.83)	Good (11.24)	Good (11.28)	Good (13.49)
RH	68.17%	77.08%	78.68%	70.81%	81.55%	75.73%	70.53%	77.81%

3.4. The calendar view based on cluster

To view the distribution of the clusters over the calendar year of 2014 and 2015, the clusters were coloured based on the mean API_{PM} concentration. The colouring system also follows the RMG [14],

where a good API is represented by the blue colour, moderate by green and unhealthy by yellow and orange. We further break out the colour intensity, with darker shades meaning relatively higher concentration. Since the range of the unhealthy condition is large, we used two colours to represent the variation in concentration levels.

In 2014, clusters 1 and 4 were classified as moderately unhealthy with their mean API_{PM} being the same range of 51 and 60. Meanwhile, clusters 2 and 3 were classified as good with their mean API_{PM} of between 31 and 40 (figure 6). In 2015, cluster 2 was still classified as good, though the API_{PM} value was slightly higher than in 2014, at between 41 and 50. Cluster 3 was classified as unhealthy with an API_{PM} value of 121 and 130. The monthly occurrences of the clusters throughout the calendar year of 2014 and 2015 were observed at a rather similar API_{PM} level for almost all months except for September, October, and November in 2015. The occurrence of unhealthy level for cluster 3 in September and October 2015, corresponded to the intense haze episodes due to transboundary pollution originating from large-scale biomass burning in Sumatra, Indonesia. [9]. During this period, there were days (Sept. 14 and Oct. 3) where the API_{PM} sub-index value were much higher in cluster 3 of between 181 and 190. The distinctive feature was observed in November 2015, whereby the actual monthly mean API_{PM} (37.15) was classified as good but was clustered as moderate. During this month, the *RH* has recorded the highest monthly average value of 81.59%. By comparing the *RH* for each cluster between 2014 and 2015, the *RH* values have shown almost similar temporal variations with some exceptions in June, July, August, and October. The actual lowest mean *RH* was in March 2014 and July 2015. Based on the calendar view, we can quickly deduce that the API_{PM} for 2014 was much better than in 2015. This was confirmed by the actual average annual mean concentration of PM, which was good (46.58) and moderate (55.98) in those respective years.

4. Conclusion

We investigated the application of *k*-means clustering in analysing multivariate time series data of *PM*, NO_2 and *RH* collected from the Department of Environment ground-based measurement station in Putrajaya, Malaysia, starting from January 2014 to December 2015. We have also added a calendar view for easy interpretation of the classified diurnal patterns based on the clusters profiled by mean API_{PM} paired with the mean *RH* measurements of the clusters. From the calendar view standpoint, we discovered that the *k*-means could be a viable method for analysing air pollutants in conjunction with other meteorological features. Since implementing the standard *k*-means algorithm has already shown promising results, further improvement on the clustering methods can be considered for accurate visualisation of the multivariate time series data. This includes identifying the best imputation methods; using the maximum value for the NO_2 observation instead of the average; varying the distance measurement for the clusters and considering other cluster validation indices to determine the number of clusters. The calendar view also seems to be a beneficial graphical aid for the cluster's patterns. This study has clearly shown the monthly status of air quality based on API values and we were also able to see the yearly differences by yearly comparison. It would also be interesting to see if the calendar could show any relation to the air quality status and monsoon seasons in Malaysia.

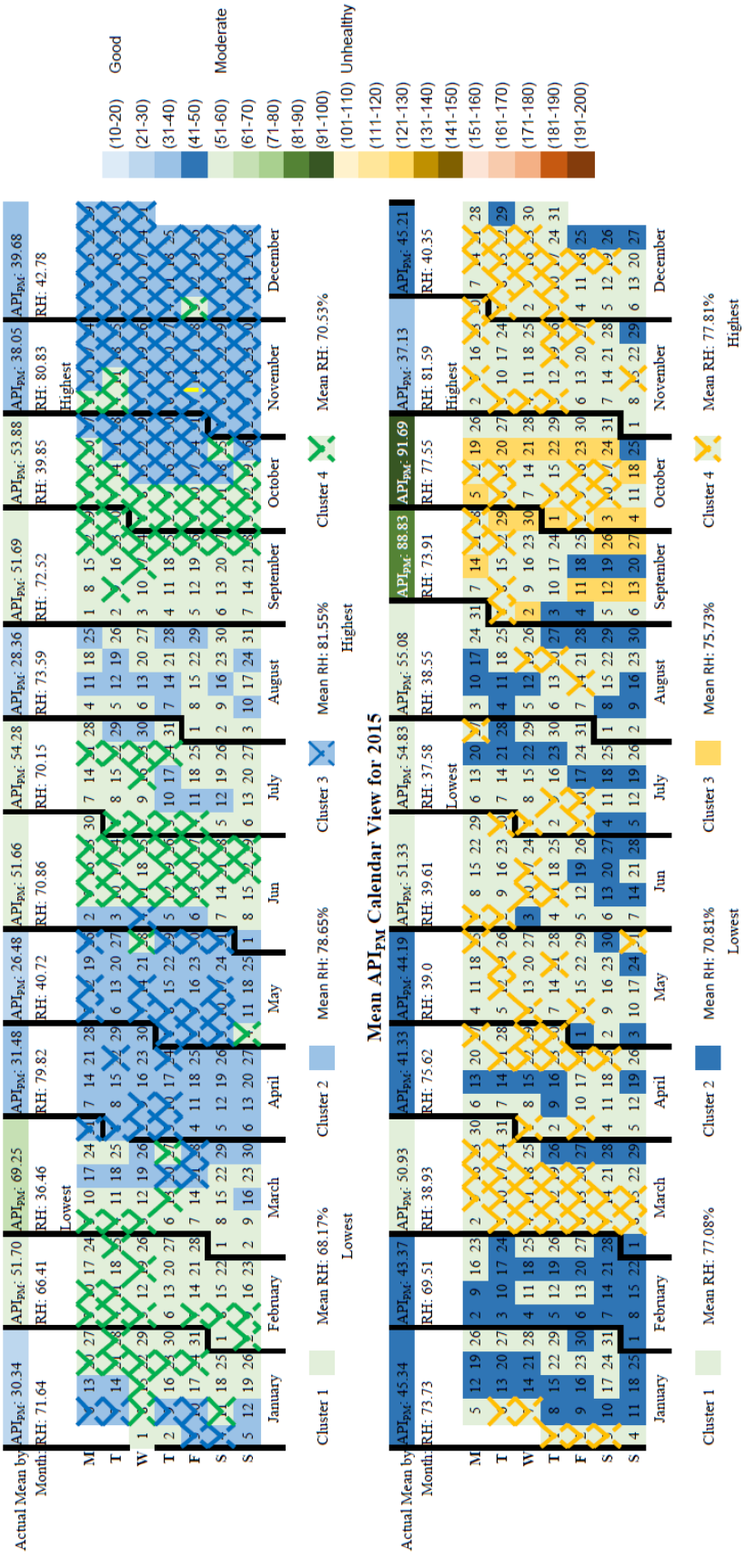


Figure 6. Mean API_{PM} calendar view for all clusters in 2014 and 2015.

5. References

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

[2] Pham D T, Dimov S S and Nguyen C D 2005 Selection of k in k-means clustering *Journal of Mechanical Engineering Science* **219** pp 103-19

[3] Vandhana S and Anuradha J 2020 Environmental air pollution clustering using enhanced ensemble clustering methodology *Environmental Science and Pollution Research* **28** pp 40746-55

[4] Alahamade W, Lake I, Reeves C E and De La Iglesia B 2021 Evaluation of Multi-variate Time Series Clustering for Imputation of Air Pollution Data *Geoscientific Instrumentation Methods and Data Systems* pp 1-23

[5] Zhang L and Yang G 2021 Cluster Analysis of PM2.5 pollution in China using frequent itemset clustering approach *Environmental Research* **204** pp 1-10

[6] Jain A K 2010 Data clustering: 50 years beyond k-means *Pattern Recognition Letter* **31** pp 651-66

[7] Van Wijk J J and Van Selow E R 1999 Cluster and Calendar Based Visualization of Time Series Data *IEEE Symposium on Information Visualization (INFOVIS'99)* (San Francisso) pp 1-6

[8] Liu J, Li J and Li W 2016 Temporal patterns in fine paticulate matter time series in Beijing: A calendar view *Nature Scientific Reports* **6** pp 1-6

[9] Ahmad Mohtar A A, Latif M T, Baharudin N H, Ahamad F, Jing X C, Othman M and Liew J 2018 Variation of major air pollutants in different seasonal conditions in an urban environment in Malaysia *Geoscience Letter* **5** pp 1-13

[10] Wan Mahiyuddin W R, Sahani M, Aripin R, Latif M T, Thuan Q T and Chit M W 2013 Short-term effects of daily air pollution on mortality *Atmospheric Environment* **65** pp 69-79

[11] Hernandez G, Berry T A, Wallis S L and Poyner D 2017 Temperature and humidity effects on particulate matter concentrations in a sub-tropical climate during winter *International Proceeding of Chemical, Biological and Environmental Engineering* **102** pp 41-9

[12] Baidari I and Patil C 2020 A criterion for deciding the number of clusters in a dataset based on data depth *Vietnam Journal of Computer Science* **7** pp 417-31

[13] Muca M, Kutrolli G and Kutrolli M 2015 A proposed algorithm for determining the optimal number of clusters *European Scientific Journal* **11** pp 112-20

[14] Department of Environment Malaysia 2000 A Guide to Air Pollutant Index (API) in Malaysia (Kuala Lumpur: Department of Environment Ministry of Science, Technology and Environment)

[15] Official Portal of Department of Environment 2021 Air Pollutant Index *www.doe.gov.my* Online: https://www.doe.gov.my/portalv1/en/info-umum/english-air-pollutant-index-api/100

[16] Cokluk O and Kayri M 2011 The effects of methods on imputation for missing values on the validity and reliability of scales *Educational Science: Theory and Practice* **11** pp 303-9

[17] Mohamad I and Usman D 2013 Standardization and its effects on K-means clustering algorithm *Research Journal of Applied Sciences, Engineering and Technology* **6** pp 3299-303

[18] Rousseeuw P J 1987 Silhouettes: a graphical aid to the interpretation and validation of cluster analysis *Journal of Computational and Applied Mathematics* **20** pp 53-65

[19] Tibshirani R, Walther G and Trevor H 2001 Estimating the number of clusters in a data set via the gap statistic *Journal of Royal Statistical Society* **63** pp 411-23
[20] Official Website of Malaysian Meteorological Department 2021 *www.met.gov.my* Online: https://www.met.gov.my/?lang=en

Acknowledgement
We would like to thank the Department of Environment, Ministry of Environment and Water for providing the dataset in this study.