An Agglomerative Hierarchical Clustering with Association Rules for Discovering Climate Change Patterns

Mahmoud Sammour[1], Zulaiha Ali Othman[2], Zurina Muda[3], Roliana Ibrahim[4]
Center for Artificial Intelligence Technology, Faculty of Information Science and Technology
Universiti Kebangsaan Malaysia, Bangi, Selangor, Malaysia[1, 2, 3]
Information Systems Department, Faculty of Computing[4]
Universiti Teknologi Malaysia, Skudai, Johor, Malaysia

***Abstract*—Ozone analysis is the process of identifying meaningful patterns that would facilitate the prediction of future trends. One of the common techniques that have been used for ozone analysis is the clustering technique. Clustering is one of the popular methods which contribute a significant knowledge for time series data mining by aggregating similar data in specific groups. However, identifying significant patterns regarding the ground-level ozone is quite a challenging task especially after applying the clustering task. This paper presents a pattern discovery for ground-level ozone using a proposed method known as an Agglomerative Hierarchical Clustering with Dynamic Time Warping (DTW) as a distance measure on which the patterns have been extracted using the Apriori Association Rules (AAR) algorithm. The experiment is conducted on a Malaysian Ozone dataset collected from Putrajaya for year 2006. The experiment result shows 20 pattern influences on high ozone with a high confident (1.00). However, it can be classified into four meaningful patterns; more high temperature with low nitrogen oxide, nitrogen oxide and nitrogen dioxide high, nitrogen oxide with carbon oxide high, and carbon oxide high. These patterns help in decision making to plan the amount of carbon oxide and nitrogen oxide to be reduced in order to avoid the high ozone surface.**

Keywords—Hierarchical clustering; dynamic time warping; ground-level ozone; Apriori Association Rules

I. Introduction

Ozone, scientifically called trioxygen, is an inorganic molecule with the chemical formula O_3. It is a pale blue gas with a distinctively pungent smell [1]. Reports suggest that ground level ozone can be rather harmful to the human respiratory system. In addition, there are also research reports that this can also result in several other detrimental diseases such as severe exposure to the ozone can negatively affect and upset lung function, and can potentially increase inflammation [2]. For instance, studies find that the mortality rate in urban areas is related to the effects of the ozone, and there is a correlation between them [3]. Other studies such as the one by [4] concluded that the effects of ozone can be non-linear, and specifically, extreme exposure to ground level ozone can be dangerous for the health. Therefore, in view of the dangers the ozone layer entails, it could be critical to research and determine the factors which cause the ozone layer to spread the most.

Several studies have been proposed for the task of predicting ozone levels [5], [6]. Such methods utilized the clustering techniques in order to group the spots that have a high level of ozone. However, the results of clustering sometimes would lead to inferior indications regarding the ozone. This is due to multiple reasons. First, the results of clustering significantly change based on the clustering technique and the distance measure used. There are several clustering techniques such as partitioning (e.g., k-means, k-medoids, etc.) and hierarchical (e.g., agglomerative and divisive). Besides that, there are different distance measures that can be utilized with the clustering technique such as Euclidean, Minkowski and Dynamic Time Warping (DTW). These choices lead to different results of clustering. On the other hand, evaluating the results of clustering is a challenging task in which different evaluation methods have been proposed for this purpose. All these mentioned reasons make the process of identifying significant patterns from ozone clustering results a difficult task.

This paper aims to propose the Apriori Association Rules algorithm in order to extract patterns from the clustering results and considered to be an extension of our study in [7]. Therefore, the next section of this paper discussed the existing techniques in the literature. In Section III introduces the proposed algorithm. While Section IV presented the performance and evaluation. A result and discussion are presented in Section V. Finally, Section VI concludes the finding of this study.

II. Related Work

According to [8] who accommodated a review for the trends of ground-level ozone using data from the last century have concluded that the ground-level ozone has dramatically increased in the last three decades. As a response, the research community has attempted to propose statistical models that have the ability to predict the increasing ozone rates. For instance, [9] proposed an agglomerative hierarchical clustering to identify the most polluted area in Houston, Texas, in terms of ground-level ozone. In their study, the authors have declared multiple factors that have a significant impact on the ozone increment, such as wind speed, wind direction, and solar radiation.

In addition, [10] proposed a k-means clustering approach with Euclidean distance measure in order to identify the peaks of ozone rates in an industrial area in Central-Southern Spain. The authors have successfully identified several polluted plots. Another approach was proposed by [11] in which a statistical method of passive sampling was used to investigate the air pollution in Pakistan. Furthermore, [12] proposed a combination of statistical means of quantile regression and agglomerative hierarchical clustering in order to measure the pollution of air in terms of ground-level ozone.

Other researchers have attempted to identify characteristics of ground-level ozone such as [13] who proposed a Hybrid Single Particle Lagrangian Integrated Trajectory (HySPLIT) Model in order to characterize the ground ozone concentration in the gulf of Texas. In their study they figured out that the lowest ozone concentrations are associated with trajectories that remained over the central Gulf for at least 48 hours. On the other hand, higher concentrations are associated with trajectories that pass close to the Northern and Western Gulf Coast. Wang et al. addressed the problem of detecting ground-level ozone from a spatio-temporal aspect [14]. The authors proposed a nearest neighbor clustering approach in order to identify spatio-temporal patterns of the air pollution.

Another observational study was conducted by [15], which concentrated on the pollution in Tangshan, North China. This study mainly relied on statistical analysis. The study implied the dramatic expansion rates of ozone and nitrogen dioxide (NO_X) from 2008 to 2011. The study concluded the reason behind the increment rates as being due to the extent of industries that are located in the city. In addition, [16] accommodated a comparative study of three regression approaches including Neural Network (NN), Support Vector Machine (SVM) and Fuzzy Logic (FL) in terms of predicting ground-level ozone. Based on the Root Mean Square Error (RMSE), SVM has shown superior performance in predicting the ozone levels. Similarly, [17] have examined two NN models including Feed-forward NN and Back-propagation NN in terms of ozone prediction. Basically, multiple features have been encoded and fed into the network including temperature, humidity, wind speed, incoming solar radiation, sulfur dioxide and nitrogen dioxide. Feed-forward NN has outperformed the other model.

In addition, [5] accommodated a comparison among two linear regression methods including SVM and multi-layer perceptron NN to identify ozone levels in the Houston–Galveston–Brazoria area, Texas. The results showed superior performance for SVM. Tamas et al. used three clustering approaches in order to detect pollution in the air including Artificial Neural Network (ANN), Self-Organized Mapping (SOM) and K-means clustering. Using hourly data, the results showed two main sources of pollution including ozone (O_3) and nitrogen dioxide (NO_2) [6].

On the other hand, [18] did a long-term statistical study for ground-level ozone in Japan from 1990 to 2010. The study focused on identifying correlation for the increment rates of ozone. The authors identified three main causes, stated as: (i) the decrease of NO titration effect, (ii) the increase of transboundary transport, and (iii) the decrease of situated photochemical production. Similarly, on an observational study of ozone level causes by [19], the authors indicated that the Asia continent is one of the main sources that affects the ground-level ozone in Western United States.

III. Materials and Method

The proposed method consists of Agglomerative Hierarchical Clustering with Dynamic Time Warping (DTW) as a distance measure. The reason behind selecting the clustering technique and distance measure lie in their superior performance according to the state of the art of ozone clustering. In addition, the Apriori Association Rules will be applied on the clustering results in order to discover knowledge. The following sub-sections will tackle the proposed method components.

A. Agglomerative Hierarchical Clustering

This phase aims to apply the hierarchical clustering technique. In general, hierarchical clustering algorithms work by aggregating the objects into a tree of clusters [20]. Hierarchical clustering can be categorized into two types, agglomerative and divisive. Such categorization is inspired from the mechanism of grouping the objects whether bottom-up or top-down approach. AHC is considered as a bottom-up hierarchical approach where each object is set in a separated cluster [21], then AHC will merge such clusters into larger clusters. The process continues until a specific termination has been reached. A complete linkage algorithm aims to identify the similarity between two clusters by measuring two nearest data points that are located in different clusters. Hence, the merge will be done between the clusters that have a minimum distance (most similar) between each other. In this paper, AHC has been applied as a maximum linkage.

B. Dynamic Time Warping (DTW)

DTW has been widely used to compare discrete sequences and sequences of continuous values [22]. Let $S = \{s_1, s_2, \dots, s_i\}$ and $T = \{t_1, t_2, \dots, t_j\}$ be a two time series sequences. DTW will minimize the differences among these series by representing a matrix of $n \times m$. In such a matrix, the distance/similarity between s_i and t_j will be calculated using Euclidean distance.

However, a warping path $P = \{p_1, p_2, \dots, p_k, \dots, p_K\}$ where $\max(m, n) \leq K \leq m + n - 1$ will be elements from the matrix that meet three constraints including boundary condition, continuity and monotonicity. The boundary condition constraint requires the warping path to start and finish in diagonally opposite corner cells of the matrix. That is $p_1 = (1,1)$ and $p_K = (m, n)$. The continuity constraint restricts the allowable steps to adjacent cells. The monotonicity constraint forces the points in the warping path to be monotonically spaced in time. The warping path that has the minimum distance/similarity between the two series is of interest. Hence, the DTW can be computed as follows:

$$d_{DTW} = min \frac{\sum_{k=1}^{K} p_K}{K} \tag{1}$$

C. Apriori Association Rules (AAR)

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases [23]. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database. In this manner, applying the Apriori algorithm on our dataset would reveal the interesting patterns that occur. In order to distinguish these interesting patterns or rules, it is necessary to consider the value of confidence which is being illustrated as follows:

Confidence: The confidence of a rule is defined as Conf (X implies Y) = supp(X ∪Y)/supp(X) in which supp(X∪Y) means "support for occurrences of transactions where X and Y both appear". Confidence ranges from 0 to 1, where the closeness to 1 indicates an interesting relation. Confidence is an estimate of Pr(Y | X), the probability of observing Y given X. The support supp(X) of an itemset X is defined as the proportion of transactions in the data set which contain the itemset.

IV. Experiment

First, the data was collected from LESTARI, which is the Institution for Environment and Development in Malaysia and the Asia Pacific. The institution has been established since 1994 with the structure of Universiti Kebangsaan Malaysia (UKM) in order to deal with environment and development issues. The data contain ozone levels for one year (i.e. 2006), particularly for the city of Putrajaya. The data are represented hourly as time intervals, which contain 8760 instances and consist of the following attributes: date, hours, O_3, NOx, nitrogen dioxide (NO_2), temperature (Temp), non-methane hydrocarbons (NMHC), and carbon oxide (CO). Hence, the proposed AHC with DTW was applied on the dataset.

Two main approaches were used to validate the clustering process; external and internal validation of clusters [24]. External validation aims to validate the clusters based on the distribution in which the common information retrieval metrics are such as precision, recall and f-measure. However, the mechanism of validation relies on labeled data. Since the real-life data are usually unlabeled, applying external validation tends to be insufficient.

On other hand, internal validation aims to measure the correctness among objects within a cluster (i.e. intra-cluster) and the correctness among objects within multiple clusters (i.e. inter-cluster). Basically, the main aim of the clustering task is to make sure that the objects within a single cluster are mostly similar, while the objects within multiple clusters are mostly dissimilar. Hence, computing the Root Mean Square Error Standard Deviation (RMSE-SD) would measure the homogenous of the objects within a single cluster and multiple clusters, which can be computed as:

$$internal\ RMSE - SD = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x_{i+1})^2 \quad (2)$$

Where n is the number of objects inside a cluster and $x_i - x_{i+1}$ is the distance between two objects in the same cluster. Similarly, RMSE-SD can be computed for the external clusters as:

$$external\ RMSE - SD = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \quad (3)$$

Where n is the number of objects of two inter clusters and $x_i - \bar{x}$ is the distance between an object in one cluster and the other object in other cluster. Similarly, RMSE-SD can be computed for the external clusters as (3).

Note that, the smaller value of RMSE-SD between the objects within a single cluster leads to better performance in which the objects are very similar. In contrast, the bigger value of RMSE-SD between the objects within a single cluster leads to lower performance in which the homogenous among the objects is being maximized. Therefore, the best results associated with a smaller value of RMSE-SD among intra-cluster, and with a greater value of RMSE-SD among inter-clusters. Based on the latter mentioned explanation, the results of applying AHC with DTW can be depicted as in Table I. As shown in Table I, the best results of intra and inter cluster has been achieved at the number of cluster 9.

The US Office of Air and Radiation have discussed the factors that lead to air pollution. In their investigation, the ozone was one of the main factors that could harm the human health. For this matter, [25] provided five categories of air pollution which are shown in Table II.

In order to provide a more critical analysis of the acquired clusters, the best number of cluster based on the RMSE-SD, which is 9, will be considered. In addition, the categorization proposed by [25] also will be considered. Therefore, two number of clusters will be considered in the analysis which are 5 and 9; the next sections will tackle this analysis.

TABLE I. Results of Intra and Inter Cluster of AHC

# Clusters	Intra-Cluster	Inter-Cluster
15	0.0042	0.3869
14	0.0041	0.3825
13	0.0041	0.3814
12	0.0042	0.3813
11	0.0045	0.3901
10	0.0045	0.4031
9	0.0039	0.4077
8	0.0039	0.3401
7	0.0041	0.3252
6	0.0066	0.3221
5	0.0068	0.3153
4	0.0073	0.3149
3	0.0054	0.3608

TABLE II. Categories of Air Pollution

# index	Unhealthy Level
1	Very Unhealthy
2	Unhealthy
3	Unhealthy for Sensitive Groups
4	Moderate
5	Good

V. Result and Discussion

A. *Analysis when K=5*

This section aims to provide a critical analysis of clustering when k=5, by identifying new patterns. This can be conducted by detecting anonymous or abnormal trends for the ground-level ozone rates. In this manner, each cluster included within the five clusters will be discussed separately.

The analysis tackles the days included in this cluster and is conducted based on three 8-hour intervals, according to [26]. Fig. 1 depicts the results of this experiment. Note that the values of the ozone have been measured using the particle per million recorded from the stations. For cluster 1, the first 8-hour interval began with 0.004 ppb and ended with 0.005 ppb, whereas the second interval showed a rise of the ozone values reaching to the peak of 0.061 ppb at 2 p.m. and ended with 0.050 ppb at 5 p.m. In the third interval, the ozone values gradually decreased reaching 0.008 ppb. This pattern is considered to be standard in accordance to the literature [27].

For cluster 2, the first interval began with 0.014 ppb and ended with 0.005 ppb. The second interval showed a rise of ozone values reaching the peak of 0.113 ppb at 2 p.m. and ended with 0.089 ppb. In the third interval, the values decreased to reach 0.014 ppb. A remarkable pattern could be noticed from this cluster, whereby this pattern represented the sharp increase and decrease of the ozone values.

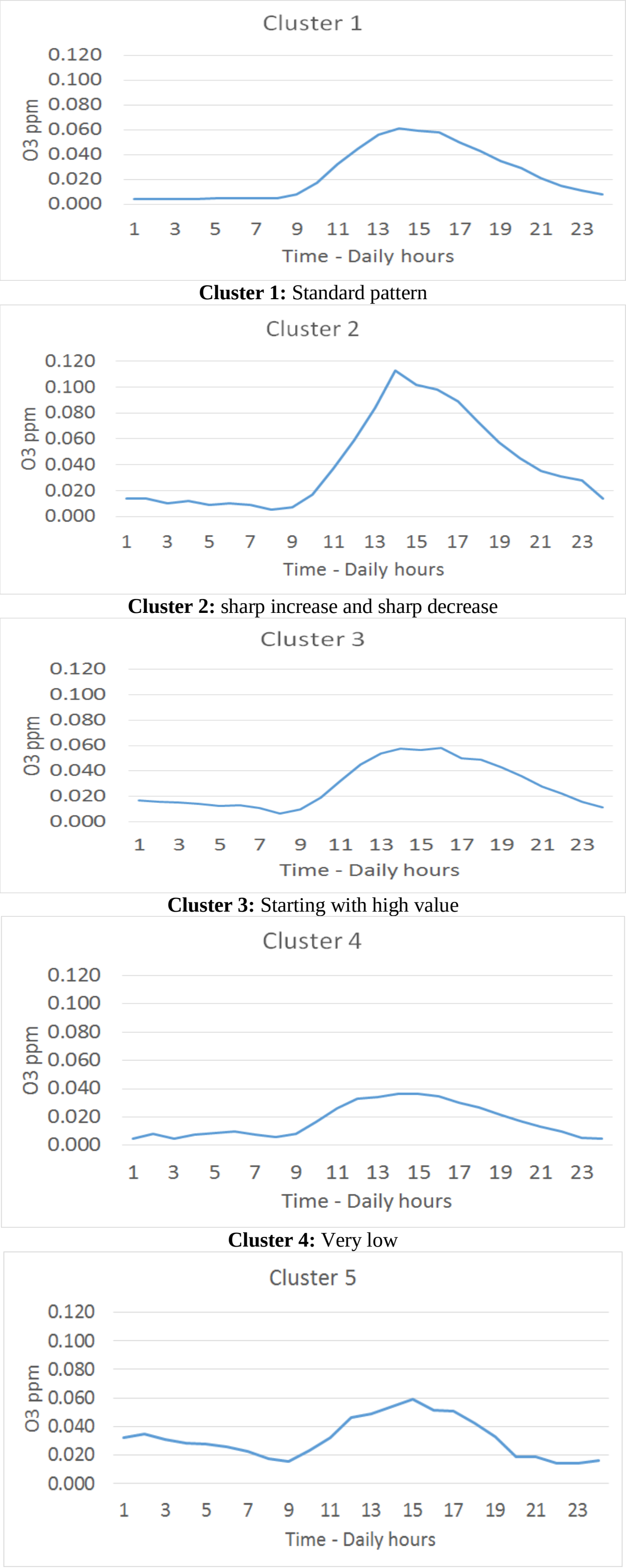
For cluster 3, the first 8-hour interval began with 0.017 ppb and ended with 0.007 ppb. The second interval showed an increase of values reaching the peak of 0.058 ppb at 2 p.m., and this peak did not change until 4 p.m. In the third interval, the values gradually decreased reaching 0.012 ppb. A pattern can be shown as starting with a high value.

For cluster 4, the first 8-hour interval began with 0.005 ppb and ended with 0.006 ppb, whereas the second interval showed an increase of values reaching the peak of 0.037 ppb at 2 p.m., and this peak did not change until 3 p.m. In the third interval, the values gradually decreased reaching 0.005 ppb. A remarkable pattern could be noticed from this cluster, whereby this pattern represented the lowest values of the ozone for the whole day.

For cluster 5, the first 8-hour interval began with 0.033 ppb and ended with 0.016 ppb, whereas the second interval showed an increase of values reaching the peak of 0.059 ppb at 3 p.m. and ended with 0.051 ppb. In the third interval, the values sharply decreased reaching 0.019 ppb at 8 p.m. and ended with 0.017 ppb. A remarkable pattern could be noticed from this cluster, whereby this pattern represented a high value of starting and unusual decline of the ozone values.

B. *Analysis when K=9*

This section aims to provide a critical analysis of clustering when k=9, by identifying new patterns. This can be conducted by detecting anonymous or abnormal trends for the ground-level ozone rates. In this manner, each cluster included within the nine clusters will be discussed separately. The analysis tackles the days included in this cluster and is conducted based on three 8-hour intervals, according to [26]. Fig. 2 depicts the results of this experiment. Note that, the values of the ozone have been measured using the particle per million recorded from the stations.



Cluster 1: Standard pattern

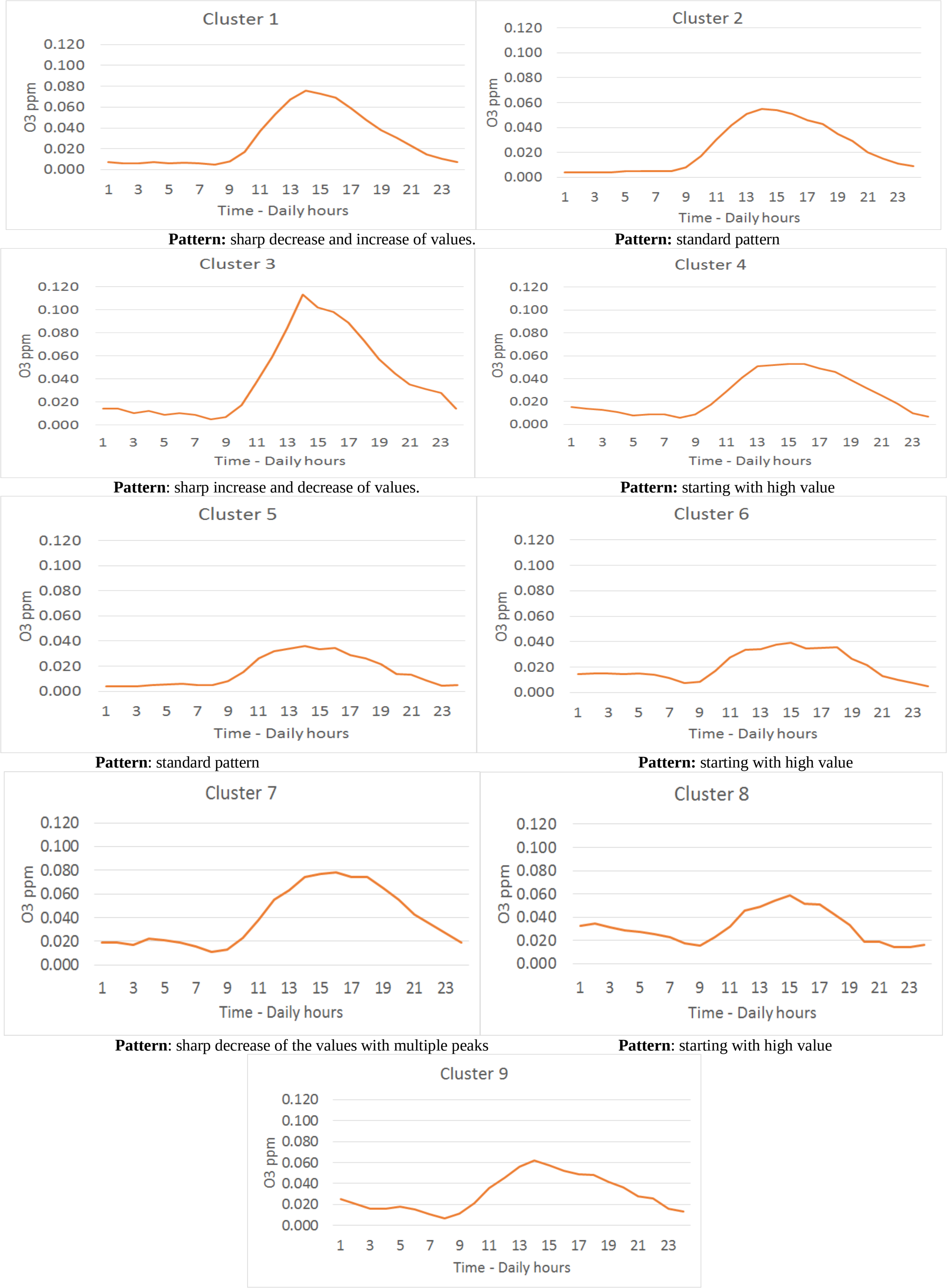
Cluster 2: sharp increase and sharp decrease

Cluster 3: Starting with high value

Cluster 4: Very low

Cluster 5: Starting with high value and unusual low

Fig. 1. Results of clustering when K=5.



Cluster 1

Pattern: sharp decrease and increase of values.

Cluster 2

Pattern: standard pattern

Cluster 3

Pattern: sharp increase and decrease of values.

Cluster 4

Pattern: starting with high value

Cluster 5

Pattern: standard pattern

Cluster 6

Pattern: starting with high value

Cluster 7

Pattern: sharp decrease of the values with multiple peaks

Cluster 8

Pattern: starting with high value

Cluster 9

Pattern: starting with high value

Fig. 2. Results of Clustering when K=9.

For cluster 1, the first 8-hour interval began with 0.008 ppb and ended with 0.005 ppb. Whereas, second interval showed a rise of ozone values reaching the peak of 0.076 ppb at 2pm, then ended up with 0.059 ppb at 5pm. In the third interval, the values gradually decreased reaching 0.007 ppb. A remarkable pattern could be noticed from this cluster, whereby this pattern represented the sharp decrease and increase of ozone values.

For cluster 2, the first interval began with 0.004 ppb and ended with 0.005 ppb. The second interval began with 0.008 ppb and sharply increased to the maximum peak of 0.055 ppb at 2 p.m., then decreased to 0.046 ppb at 5 p.m. The third interval decreased and reached 0.009 ppb. This cluster has a standard pattern

For cluster 3, the first interval began with 0.014 ppb and ended with 0.005 ppb. The second interval showed an increase of values reaching the peak of 0.113 ppb at 2 p.m. and ended with 0.089 ppb. In the third interval, the values decreased to 0.014 ppb. A remarkable pattern could be noticed from this cluster, whereby this pattern represented the sharp increase followed by a sharp decrease of values.

For cluster 4, the first interval began with 0.015 ppb and ended with 0.006 ppb. The second interval showed multiple peaks in which the values at 4 p.m. reached 0.053, and 0.049 ppb at 5 p.m. The third interval showed a decrease of values that reached 0.007 ppb.

For cluster 5, the first interval began with 0.004 ppb and ended with 0.005 ppb. The second interval showed three peak stated as; 0.032 ppb at 12 p.m., 0.036 at 2 p.m., and 0.035 ppb at 4 p.m., and ended with 0.029 ppb. The third interval showed an unstable decrease that reached 0.005 ppb. This cluster has a standard pattern.

For cluster 6, the first interval began with 0.015 ppb then showed a stable decrease until 8 a.m. reaching 0.008 ppb. The second interval showed two peaks of 0.034 at 12 p.m. and 0.039 at 3 p.m. The third interval showed a gradual decrease of the values reaching 0.005 ppb. The pattern embedded in the cluster has a high value of the starting point.

For cluster 7, the first interval began with 0.019 ppb and ended with 0.011 ppb. The second interval showed two peaks of 0.078 ppb at 4 p.m. and 0.074 ppb at 2 p.m. The third interval showed a gradual decrease reaching 0.019 ppb. A remarkable pattern could be noticed from this cluster, whereby this pattern represented a sharp decrease of the values with multiple peaks.

For cluster 8, the first interval began with 0.033 ppb and ended with 0.016 ppb. The second interval showed a maximum peak of 0.059 ppb at 3 p.m. and then ended with 0.051 ppb. The third interval showed a sharp decrease of values reaching 0.017 ppb. The pattern embedded in this cluster has a high value of the starting point.

For cluster 9, the first interval began with 0.025 ppb and ended with 0.007 ppb. The second interval showed a sharp increase of values reaching a maximum of 0.062 ppb at 2 p.m. and ended with 0.049 ppb. The third interval showed a gradual decrease of the values reaching 0.014 ppb. The pattern embedded in this cluster has a high value of the starting point.

C. Comparison between K=5 and K=9

This section aims to accommodate a comparison between the two numbers of cluster 9 and 5 which were analyzed in the previous sections. The comparison will be based on multiple variables including the starting values of ozone, maximum peak, maximum peak of median, and ending values. Table III shows the values of 5 number of clusters.

As shown in Table III, the number of days included in the 'unhealthy' category represents nearly half of the year which seems to be the overestimated categorization. This means that this category should be divided into more categories, whereas the 'moderate' category contains only eight days which seems to be the underestimated categorization. Generally, this category is supposed to contain more days.

However, Table IV shows the values of 9 number of clusters. As shown in Table IV, unlike the standard 5 categorization, the 9 categorization has the ability to provide a better description of the year's days. This can be represented by giving more categories. For instance, the 'unhealthy' category has been split into two categories, namely 'unhealthy' and 'very unhealthy for sensitive group'. These categories have shown a reasonable contained number of days. In addition, the category 'moderate' has been split into three categories, namely 'high moderate', 'moderate' and 'low moderate'. Similarly, these categories contained a reasonable number of days. Finally, the category 'good' has been also divided into two categories as 'very good' and 'good'.

D. Extracting Pattern using AAR

Basically, determining the factors that affect the ozone is a difficult task. Akimoto et al. conducted a study to analyze the causes of ground-level ozone in Japan using 20 years of data [18].

As a conclusion in their study, they have surprisingly found that even with the decrease of NOx and NMHC (i.e. considered as the main causes of increasing the ozone rates), there is still an ongoing increment of ground-level ozone rates. Based on their judgment, they have referred the reason to transportation. Hence, it is a challenging task to identify the factors that would affect the ground-level ozone. However, this study attempts to present an analysis for specific cases of extreme growth of ozone rates. Therefore, the association rules approach has been used in order to clarify the factors that would increase rates of ozone in Putrajaya, Malaysia. Table V depicts the results of applying association rules by showing the most significant patterns with highest confidences.

As we can see in Table V, the first 12 rules are associated with two factors whereas the other rules are associated with a single factor. In particular, the first five rules (i.e. 1-5) are associated with NOx and the temperature, where the increase of temperature with NOx = 0.003 leads to an increase in the ozone rates. In addition, the following two rules (i.e. 6 and 7) are associated with NOx and NO_2, where the decrease of NO_2 with an NOx = 0.003 would lead to an increment in the ozone rates. The five following rules (i.e. 8-12) are associated with NOx and CO, where the decrease of CO with NOx = 0.003 (especially for rule 10 and 11) would lead to an increment in the ozone rates.

On the other hand, the remaining eight rules (i.e. 13-20) are associated with a single factor which is CO. In fact, these rules are related to the peak or highest rates of ozone. Although there is no direct relation between the CO values and the ozone rates, as a general view, CO is related to transportation. This can be evidenced in the findings of [18] study which implies that transportation is one of the main reasons behind the growth of ground-level ozone rates.

TABLE III. VALUES USING CLUSTER = 5

Days	**K=5**	**Morning**		**Afternoon**		**Evening**	**Standard Category**
	Class	***Start***	***end***	***Max***	***Men Max***	***end***	
60	4	0.005	0.006	0.09	0.037	0.005	Good
8	5	0.033	0.016	0.093	0.059	0.017	Moderate
104	3	0.017	0.007	0.105	0.058	0.012	Unhealthy for Sensitive Groups
173	1	0.004	0.005	0.115	0.061	0.008	Unhealthy
19	2	0.014	0.005	0.148	0.113	0.014	Very Unhealthy

TABLE IV. VALUES USING CLUSTER = 9

Days	**K=9**	**Morning**		**Afternoon**		**Evening**	**Proposed Category**
	Class	***Start***	***end***	***Max***	***Men Max***	***end***	
38	5	0.004	0.005	0.06	0.036	0.005	Very Good
22	6	0.015	0.008	0.09	0.039	0.005	Good
63	4	0.015	0.006	0.093	0.053	0.007	High Moderate
121	2	0.004	0.005	0.096	0.055	0.009	Moderate
8	8	0.033	0.016	0.093	0.059	0.017	low Moderate
20	9	0.025	0.007	0.077	0.062	0.014	Unhealthy for Sensitive Groups
52	1	0.008	0.005	0.115	0.076	0.007	Very Unhealthy for Sensitive Groups
21	7	0.019	0.011	0.105	0.078	0.019	Unhealthy
19	3	0.014	0.005	0.148	0.113	0.014	Very Unhealthy

TABLE V. SIGNIFICANT PATTERN EXTRACTED USING AAR

#	Factor 1	Factor 2	=>	Ozone	Confidence
1	NOx=0.003	Temp=33.7	=>	0.089	1.00
2	NOx=0.003	Temp=32.1	=>	0.088	1.00
3	NOx=0.003	Temp=30.4	=>	0.070	1.00
4	NOx=0.003	Temp=29.9	=>	0.061	1.00
5	NOx=0.003	Temp=29.2	=>	0.059	1.00
6	NOx=0.003	NO2=0.013	=>	0.059	1.00
7	NOx=0.003	NO2=0.002	=>	0.070	1.00
8	NOx=0.003	CO=1.7	=>	0.070	1.00
9	NOx=0.003	CO=1.61	=>	0.061	1.00
10	NOx=0.003	CO=0.31	=>	0.088	1.00
11	NOx=0.003	CO=0.27	=>	0.086	1.00
12	NOx=0.003	CO=0.16	=>	0.078	1.00
13	CO=0.75	-	=>	0.148	1.00
14	CO=0.91	-	=>	0.147	1.00
15	CO=0.7	-	=>	0.143	1.00
16	CO=0.44	-	=>	0.140	1.00
17	CO=0.63	-	=>	0.140	1.00
18	CO=0.60	-	=>	0.139	1.00
19	CO=0.59	-	=>	0.139	1.00
20	CO=0.78	-	=>	0.131	1.00

VI. CONCLUSION

This study has proposed a pattern extraction method for ground-level ozone using the Apriori Association Rules method. The data used in the experiment were collected from LESTARI, which is the Institution for Environment and Development in Malaysia and the Asia Pacific. In the beginning, AHC with DTW were applied on the dataset in order to cluster the days based on the ozone levels. Consequentially, the proposed AAR was applied to extract significant patterns. The experiment shows that the extracted patterns are related to CO which is an interesting relation in accordance to the literature. In fact, this study utilized a ground-level ozone data with a single year. Hence, using a dataset with multiple years in future researches has the ability to identify frequent patterns which may facilitate the determination of the important factors of the ozone. However, there is two limitations in this study which is the resources of hardware and time consuming. This is because, AHC requires a high computational cost and time. Therefore, only one-year ozone data were selected in this study to avoid this limitation.

ACKNOWLEDGMENT

This study is supported by the Universiti Kebangsaan Malaysia (UKM) and funded by research grant (FRGS/1/2016/ICT02/UKM/02/8).

REFERENCES

[1] H. Liu et al., “Ground-level ozone pollution and its health impacts in China,” Atmos. Environ., vol. 173, pp. 223–230, 2018.

[2] L. Shen, L. J. Mickley, and E. Gilleland, “Impact of increasing heat waves on U.S. ozone episodes in the 2050s: Results from a multimodel analysis using extreme value theory,” Geophys. Res. Lett., 2016.

[3] M. Ahmadi, Y. Huang, and K. John, “Application of spatio-temporal clustering for predicting ground-level ozone pollution,” in Advances in Geographic Information Science, 2017, pp. 153–167.

[4] W. Zhao, S. Fan, H. Guo, B. Gao, J. Sun, and L. Chen, “Assessing the impact of local meteorological variables on surface ozone in Hong Kong during 2000–2015 using quantile and multiple line regression models,” Atmos. Environ., vol. 144, pp. 182–193, 2016.

[5] W. Sun et al., “Prediction of surface ozone episodes using clusters based generalized linear mixed effects models in Houston–Galveston–Brazoria area, Texas,” Atmos. Pollut. Res., 2015.

[6] W. Tamas, G. Notton, C. Paoli, M. L. Nivet, and C. Voyant, “Hybridization of air quality forecasting models using machine learning and clustering: An original approach to detect pollutant peaks,” Aerosol Air Qual. Res., vol. 16, no. 2, pp. 405–416, 2016.

[7] M. Sammour and Z. Othman, “An Agglomerative Hierarchical Clustering with Various Distance Measurements for Ground Level Ozone Clustering in Putrajaya , Malaysia,” Int. J. Adv. Sci. Eng. Inf. Technol., 2016.

[8] M. Weber et al., “Total ozone trends from 1979 to 2016 derived from five merged observational datasets-the emergence into ozone recovery,” Atmospheric Chemistry and Physics. 2018.

[9] K. B. Ensor, B. K. Ray, and S. J. Charlton, “Point source influence on observed extreme pollution levels in a monitoring network,” Atmos. Environ., 2014.

[10] J. A. Adame, A. Notario, F. Villanueva, and J. Albaladejo, “Application of cluster analysis to surface ozone, NO2and SO2daily patterns in an industrial area in Central-Southern Spain measured with a DOAS system,” Sci. Total Environ., 2012.

[11] S. S. Ahmad and N. Aziz, “Spatial and temporal analysis of ground level ozone and nitrogen dioxide concentration across the twin cities of Pakistan,” Environ. Monit. Assess., vol. 185, no. 4, pp. 3133–3147, 2013.

[12] A. Monteiro et al., “Trends in ozone concentrations in the Iberian Peninsula by quantile regression and clustering,” Atmos. Environ., 2012.

[13] O. Connan, K. Smith, C. Organo, L. Solier, D. Maro, and D. Hébert, “Comparison of RIMPUFF, HYSPLIT, ADMS atmospheric dispersion model outputs, using emergency response procedures, with 85Kr measurements made in the vicinity of nuclear reprocessing plant,” J. Environ. Radioact., 2013.

[14] M. Deng, Q. L. Liu, J. Q. Wang, and Y. Shi, “A general method of spatio-temporal clustering analysis,” Sci. China Inf. Sci., vol. 56, no. 10, pp. 1–14, 2013.

[15] D. Ji et al., “The heaviest particulate air-pollution episodes occurred in northern China in January, 2013: Insights gained from observation,” Atmos. Environ., 2014.

[16] P. Hájek and V. Olej, “Ozone prediction on the basis of neural networks, support vector regression and methods with uncertainty,” Ecol. Inform., 2012.

[17] M. Kandil, A. Gadallah, F. Tawfik, and N. Kandil, “Prediction of Maximum Ground Ozone Levels using Neural Network,” Int. J. Comput. Digit. Syst., vol. 218, no. 1224, pp. 1–8, 2014.

[18] H. Akimoto, Y. Mori, K. Sasaki, H. Nakanishi, T. Ohizumi, and Y. Itano, “Analysis of monitoring data of ground-level ozone in Japan for long-term trend during 1990-2010: Causes of temporal and spatial variation,” Atmos. Environ., 2015.

[19] E. Manzini et al., “Northern winter climate change: Assessment of uncertainty in CMIP5 projections related to stratosphere-troposphere coupling,” J. Geophys. Res. Atmos., vol. 119, no. 13, pp. 7979–7998, 2014.

[20] M. Aghagolzadeh, H. Soltanian-Zadeh, B. Araabi, and A. Aghagolzadeh, “A hierarchical clustering based on mutual information maximization,” in Proceedings - International Conference on Image Processing, ICIP, 2006.

[21] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, “Efficient agglomerative hierarchical clustering,” Expert Syst. Appl., 2015.

[22] S. Salvador and P. Chan, “Toward accurate dynamic time warping in linear time and space,” Intell. Data Anal., 2018.

[23] A. J. Doshi and B. Joshi, “Comparative analysis of Apriori and Apriori with hashing algorithm,” Int. Res. J. Eng. Technol., 2018.

[24] Y. Zhang et al., “Multi-kernel extreme learning machine for EEG classification in brain-computer interfaces,” Expert Syst. Appl., vol. 96, pp. 302–310, 2018.

[25] K. Kuklinska, L. Wolska, and J. Namiesnik, “Air quality policy in the U.S. and the EU – a review,” Atmos. Pollut. Res., 2015.

[26] A. Zare, J. H. Christensen, A. Gross, P. Irannejad, M. Glasius, and J. Brandt, “Quantifying the contributions of natural emissions to ozone and total fine PM concentrations in the northern Hemisphere,” Atmos. Chem. Phys., 2014.

[27] Y. Y. Toh, S. F. Lim, and R. von Glasow, “The influence of meteorological factors and biomass burning on surface ozone concentrations at Tanah Rata, Malaysia,” Atmos. Environ., 2013.