International Journal on Advanced Science Engineering Information Technology

Vol.8 (2018) No. 6
ISSN: 2088-5334

Rules Discovery of High Ozone in Klang Areas using Data Mining Approach

Zulaiha Ali Othman[#1], Noraini Ismail[#2], Azuraliza Abu Bakar[#3], Mohd Talib Latif[*], Sharifah Mastura Syed Abdullah[+]

[#]*Center for Artificial Intelligence and Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*
E-mail: [1]zao@ukm.edu.my, [2]norainismail14@yahoo.com, [3]azuraliza@ukm.edu.my

[*]*School of Environmental and Natural Resource Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*
E-mail: talib@ukm.edu.my

[+]*Social, Environmental and Developmental Sustainability Research Centre (SEEDS), Faculty of Social Sciences and Humanities, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia*
E-mail: [5]pghikp@ukm.edu.my

***Abstract*— Ground level ozone (O_3) is one of the common pollution issues that has a negative influence on human health. However, the increasing trends in O_3 level nowadays which due to rapid development has become a great concern over the world. Thus, developing an accurate O_3 forecasting model is necessary. However, the interesting pattern from the data should be identified beforehand. Association rules is a data mining technique that has an advantage to discover frequent patterns in a dataset, which subsequently will be useful in the research domain. Therefore, this paper presents the discovering knowledge based on association rules and clustering technique towards a climatological O_3 dataset. In this study, the data was analysed to find the behaviour of each precursors. Later K-means clustering technique was used to find the suitable range for each chosen variable independently, then applied Apriori based association rules technique to present the behaviours in a meaningful and understandable format. The climatological O_3 time series data has been collected from Department of Environment for Klang station from year 1997 to 2012. However, the proposed method only applied on high O_3 concentration data during stated years to find the association pattern. The outcome has discovered 17 strong rules. The patterns and behaviours of the selected variables during high O_3 concentration has been discovered. The rules are benefit to the government on how to control the air quality later.**

***Keywords*—data mining; ozone; association rule; apriori; clustering.**

I. INTRODUCTION

In recent years, the concentration of Ozone (O_3) pollution has been reported increases in several countries such as Japan, China, California and many others [1]. This air polluter is very harmful to the human health and the ecosystem balance. In 2013, the World Health Organization reported that the O_3 threshold standard for public health might not be practical for a sensitive group population. This is because these group populations are also affected even at a low level of O_3. This group population is including people with asthma, old citizens, and children. Therefore, the development of strategies for effective risk reduction towards high O_3 is needed. Thus, further data analysis and comprehensive monitoring should be taken.

O_3 pollution is one of the greenhouse gases produced By human activity through pollutant emissions such as Nitrogen Dioxide (NO_2), Carbon Monoxide (CO), Non-Methane Organic Compound (NMHC) and Organic Compounds (VOC) that reacts in the atmosphere with the presence of sunlight [2]. The primary anthropogenic source that produces these pollutants is from the smoke of the vehicle as well as industrial factories. An O_3 chemical is formed by a simple oxidation reaction of CO with the presence of Nitric Oxide (NO). While the Hydroxyl (OH), Hydroperoxyl (HO_2), NO and NO_2 chains and act as a

catalyst in this reaction. Additionally, organic compounds such as NMHC may also join a chain of reactions when the carbonyl or ketone species are formed beside O_3 [3]. This shows that many precursors are involved in O_3 formation. Therefore, the deep understanding is necessary to investigate the complicated relationship between O_3 and its precursors [4].

There is various method has been proposed for analyzing and detecting O_3 trends, especially using a statistical method and data mining techniques [1-12]. Based on past study, the analysis of O_3 has been conducted to investigate the variables that give a significant impact on O_3 formation using a statistical method. According to a research in [2], the result shows that O_3 rapidly increased when NMHC and NO_2 react with the ambient oxygen at the high temperature. Also, [3] found that NMHC has a similar seasonal variation with NO_2 during summer and winter. Meanwhile, O_3 concentration shows the different distribution with NMHC and NO_2. The study in [4] has also done the similar research to evaluate several variables such as NO, NO_2, NMHC, and CO toward O_3 level. As a result, the selected parameter has shown to give a significant impact on O_3 concentration. However, during the analysis, there is zero situation where O_3 and its precursor's concentration exceed the norms level that can threaten human health. Hence, the averaging method used in this studies cause the information lost. On the other hand, these study only focus on pattern distribution of each selected variable without enclosing the exact values for each variable. Therefore, many researchers have used data mining method to solve this problem [5-8].

Data mining is a method used to discover hidden or exciting knowledge from a past data that for the future [5]. Based on previous research, data mining can be categorized into three task including clustering, classification, and rules discovery. Rules discovery is a data mining technique that aims to attract an exciting relationship between variables in an extensive database [6]. On the other hand, one of the data mining technique namely association rules may explain the acquisition of useful knowledge and patterns from data in an understandable format. The generated rules found are helpful in decision making for the related domain. However, there are only several studies that applied association rules in this problem domain based on state of the art. In [9], the researcher applying association rules hybrid with a genetic algorithm to discover knowledge among correlated real-world O_3 data. The attribute used in this research is temperature, wind, and O_3. On the other hand, he also proposed enhanced association rules for mining multi-dimensional time series data. Although this technique is compelling, however, it still has an inadequacy in dealing with quantitative data [6]. In order to overcome this constraint, a dataset should undergo a discretization process [8].

Clustering algorithm can be applied as a discretization process by partitioning the data into intervals with similar features to make the interval more meaningful [7]. K-means can explain the characteristic of data distribution by classifying a given dataset into a *k* number of a cluster with the nearest Euclidean distance between two item set [8]. Therefore, this study proposed a clustering technique into association rules framework to discretize a dataset.

II. Material and Method

This section is divided into three subsections to discuss the dataset background, pre-processing step and algorithm explanation in part *A*, *B*, and *C* below:

A. Data

This study uses the actual climatological O_3 time series data taken from Klang station located in Selangor, Malaysia (Latitude N 03° 00,620' and longitude E 101° 24,484'). Klang is known as a royal town and a former capital of the state of Selangor. Klang has divided into two regions, which is North Klang and South Klang. Both regions are a commercial and residential area. However, South Klang tends to be busier compared to North Klang due to the development of government officials and the healthcare facilities that placed in that region. The quick progress of innovative and contemporary townships in Klang has led to the increasing pollution level. One of this pollution is O_3. Therefore, this study aims to find a set of rules that discover the knowledge of high concentration O_3 pattern surrounding Klang areas.

This station was selected based on the completeness of the data and the only station that recorded NMHC concentration. This dataset contains 140,280 instances and eight attributes (all continuous input attribute). Table 1 below describes each variable that available in a dataset. The table shows a low percentage of missing data with the highest percentage is 8.83%. The parameter readings for all variables is ppm unit, and the data were recorded hourly.

B. Data Pre-processing

This section will explain the data preparation phase including data analysis, dimensionality reduction, and data discretization.

1) Data Analyzing: Data analyzing is an important step that should be taken in the pre-processing phase to ensure the relationship between variables is identified. This is because uncorrelated variables will not give useful information. In this experiment, only three variables were selected which is NO_2, NMHC, and O_3. Figure 1 below shows the correlation between NO_2 towards O_3, NMHC towards O_3 and NO_2 towards NMHC. The correlation graph for NO_2 towards O_3 and NMHC towards O_3 shows the high negative correlation, while a correlation graph for NO_2 towards NMHC shows high positive correlation. The high negative correlation for NO_2 and NMHC towards O_3 is due to the oxidization process, and the strong positive correlation between NO_2 and NMHC is due to the emission process from a motor vehicle.

Figure 2 depicted the diurnal variation for each variable. Based on the graph, there is the apparent peak of NO_2, NMHC, and O_3 concentration. In general, the daily O_3 concentration was found to increase in the morning at 9:00 hours and return to its earlier and lower concentration at 21:00 hours in the late evening. The highest concentration of O_3 also was recorded during mid-afternoon between 13:00 hours and 15:00 hours. This phenomenon is a typical normal for diurnal variation patterns for O_3. On the other hand, the daily concentration of parameters NO_2 was recorded as being highest during morning peak between 8:00 – 10:00 hours and late evening at 21:00 – 23:00 hours. While NMHC

concentration was found to peak at 7:00 – 9:00 hours in the morning and 22:00 – 24:00 hours in the late evening that about one hour before NO_2. This phenomenon related to the peak hour of people going to work and returning from work with the number of a number of the motor vehicle that can increase NO_2 and NMHC formation. However, the concentration was found stable from midday towards the late afternoon. The graph also shows that O_3 concentration is high during low NO_2 and low NMHC, and vice versa. Based on this analysis, we found the correlation between the variables, but the statistical approach cannot be used in extracting important behavior of the data. This is because no high O3 exceeded 0.04 ppm due to the lose the crucial features of the data after the values were averaged in a 24-hour format.

2) Dimensionality Reduction: Real-world time series data are often high in dimensionality and recorded in continuous form. Some of machine learning algorithm cannot cope with this high dimensional data, especially in term of storage, speed, and accuracy of the query. Therefore, this study reduces the dimensionality of the data by reducing random variables, which are not correlated, not relevant or not give a significant impact on a result (or also known as features selection).

TABLE I
PARAMETERS, PERIOD OF DATA COLLECTIONS AND PERCENTAGE OF MISSING VALUES OF KLANG STATION

No	Parameter	Years	Duration	Missing Data %	Total Instances
1	O_3	16	1997-2012	7.32%	139536
2	NO	16	1997-2012	7.9%	139536
3	NO_2	16	1997-2012	8.12%	139536
4	NO_X	16	1997-2012	8.83%	128220
5	CO	16	1997-2012	6.83%	139214
6	NMHC	14	1997-2010	8.25%	116112
7	Ambient Temperature	16	1997-2012	2.69%	139536
8	Wind Speed	16	1997-2012	6.54%	140280

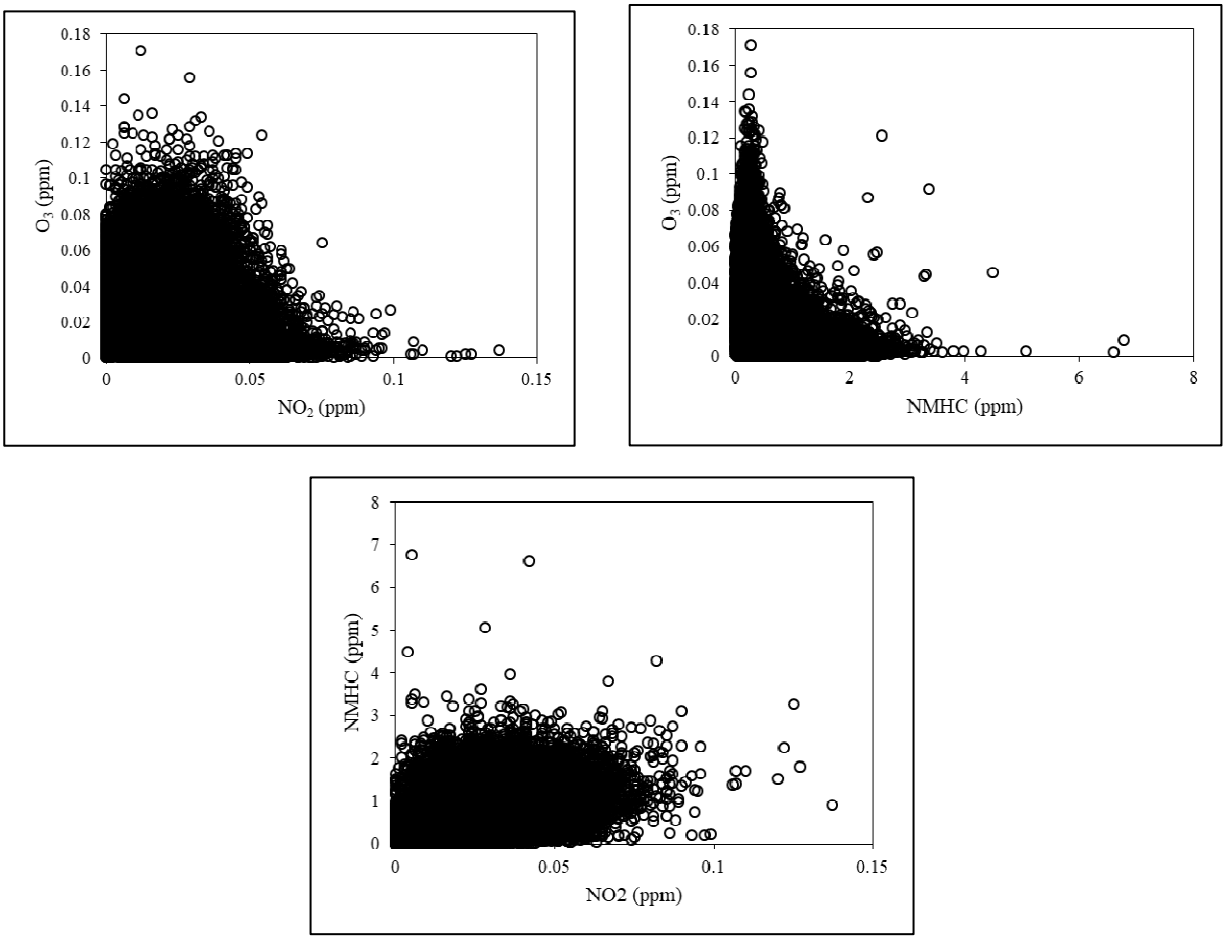

Fig. 1. Dispersion graph between parameter NO_2, NMHC and O_3

3) Data Cleaning: This phase involves identifying and fixing errors in the data. In this study, the collected data has missing values, and the data between the parameter are inconsistent. Several methods exist to replace missing data with reasonable numbers of missing values. One of them is using a k-nearest neighbor. K-nearest neighbor will replace missing values with the corresponding 24 value from the nearest-neighbor column using Euclidean distance function. If the corresponding value from the nearest-neighbor column is also empty, the next nearest 24 columns are used. For a missing data that exceed more than a month continuously, the corresponding row for that month will be removed from dataset to avoid biased result. After the cleaning process, only 113,232 instances was selected from total 140,280

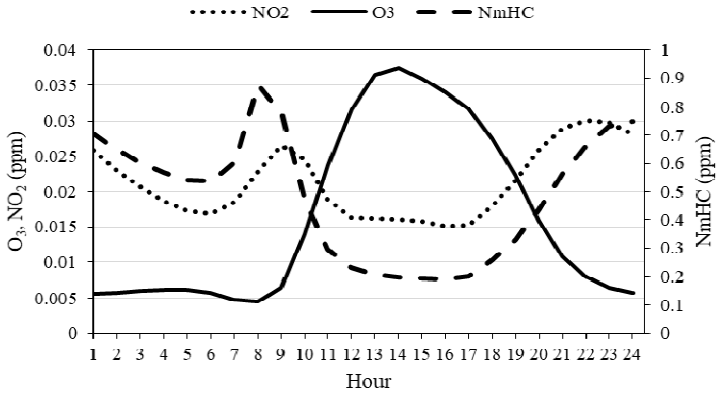

Fig. 2. Correlation graph for average values in a day between NO_2, NMHC, and O_3

TABLE II
THE INTERVAL BETWEEN O_3 GROUP BY CLUSTER AND TOTAL INSTANCES IN THE CLUSTER

Cluster	Range	Total Instances	Cluster	Range	Total Instances	Cluster	Range	Total Instances
1	0-0	2750	17	0.016-0.017	3986	33	0.039-0.039	734
2	0.001-0.001	6533	18	0.018-0.018	1856	34	0.04-0.04	693
3	0.002-0.002	7629	19	0.019-0.019	1813	35	0.041-0.043	1940
4	0.003-0.003	7837	20	0.02-0.02	1594	36	0.044-0.045	1196
5	0.004-0.004	7473	21	0.021-0.022	3405	37	0.046-0.047	1013
6	0.005-0.005	6068	22	0.023-0.023	1569	38	0.048-0.049	920
7	0.006-0.006	4879	23	0.024-0.025	2991	39	0.05-0.05	370
8	0.007-0.007	4220	24	0.026-0.027	2758	40	0.051-0.051	370
9	0.008-0.008	3777	25	0.028-0.029	2602	41	0.052-0.054	969
10	0.009-0.009	3338	26	0.03-0.03	1209	42	0.055-0.059	1280
11	0.01-0.01	2900	27	0.031-0.032	2216	43	0.06-0.065	1049
12	0.011-0.011	2723	28	0.033-0.033	1015	**44**	**0.066-0.072**	**767**
13	0.012-0.012	2490	29	0.034-0.034	994	**45**	**0.073-0.082**	**552**
14	0.013-0.013	2389	30	0.035-0.036	1847	**46**	**0.083-0.099**	**370**
15	0.014-0.014	2208	31	0.037-0.037	892	**47**	**0.1-0.171**	**112**
16	0.015-0.015	2109	32	0.038-0.038	827			

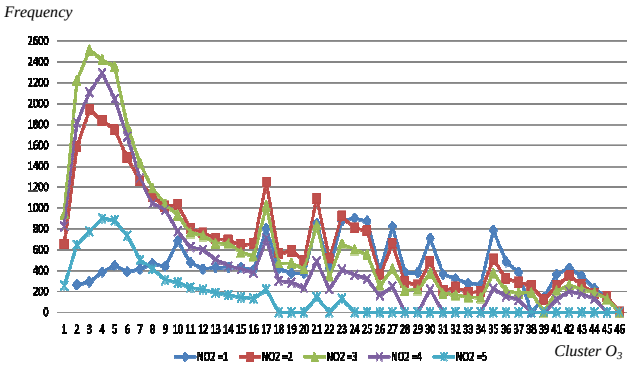

Fig. 3. Graph frequency of NO_2 against cluster O_3

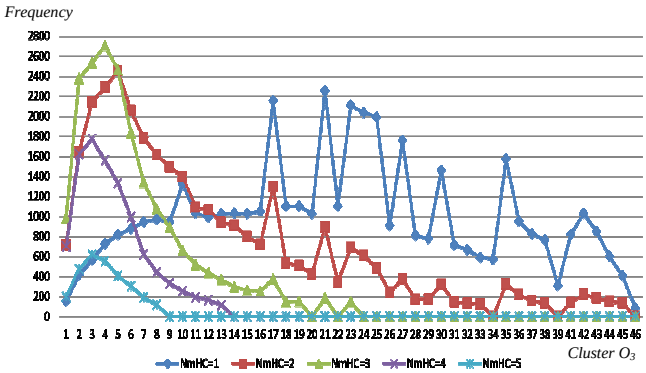

Fig. 4. Graph frequency of NMHC against cluster O_3

instances. The new data is from January 1997 to December 2009.

4) Discretization: Discretization is a method used to convert continuous values into discrete values. Besides the conventional method, a clustering technique can be used as discretization tools. Hence, it can become a more powerful method compared to the conventional method. This is because of the features in a clustering technique that can divide the item into intervals with similar features to make the interval more meaningful [7]. Therefore, this experiment used a k-means clustering technique as a discretization method. K-means works by classifying a dataset into *k* number of a group with the nearest mean. The distance between two points of the item in a dataset will be measured using Euclidean distance [10]. The interval of a cluster explained the characteristics of the data distribution. Other than that, the discretization interval can be more meaningful by clustered all the attributes alone. Table 2 below shows the range of interval for an O_3 variable after the discretization process. Based on table 2, the O_3 variable was clustered into 47 group to find the nearest interval that near to 0.1 ppm. This is because the values of O_3 that above or equal to 0.1 ppm can consider high and dangerous [13].

According to a clustering result in table 2, the total cases of O_3 exceed or equal to 0.1 ppm is only 112 cases, which is only 0.08% occurrences in 16 years from 1997-2012. This shows the occurrence of O_3 exceedance at Klang area is not so frequent. However, the increasing amount of O_3 from time to time give a significant impact on human life. Also, the old citizens, children and people who suffer from asthma disease are affected by O_3 pollution even at a low level. Hence, this study will select a cluster that near to 0.1 ppm which is cluster 44, 45, 46 and 47 to be mined in the next phases. On the other hand, NO_2 and NMHC concentration were clustered into 5 group categories as shown in Table 3.

TABLE III
GROUPING RANGE FOR NO_2 AND NMHC CONCENTRATION

	NO_2 (ppm)		**NMHC (ppm)**	
No	***Range***	***Total***	***Range***	***Total***
1	0.00-0.01	19986	0-0.28	45550
2	0.011-0.018	30586	0.29-0.55	32512
3	0.019-0.026	30374	0.56-0.92	20829
4	0.027-0.039	23815	0.93-1.46	10858
5	0.04-0.137	8471	1.47-6.77	3483

The clustering result was illustrated in the graph to see the frequency pattern of cluster NO_2 and NMHC over O_3. Based on figure 3, the graph shows that the high frequency of NO_2 happens in a small range of O_3. The graph looks consistent for all groups of NO_2. The study in [1] and [3] state that, although NO_2 concentration has a high impact on the formation of O_3, however, it needs some time to oxidize from NO_2 to O_3 and other involvement of precursors also affect the formation of O_3. Additionally, the O_3 formation will be stopped at a certain level of NO_2, but this fact still unexplained yet.

Furthermore, figure 4 depicted the graph for the frequency of group NMHC against a group of O_3. Based on the graph, the high frequency of NO_2 happens in a small range of O_3, and the frequency values depend on the group of NMHC e.g., NMHC group 1 gives a smaller frequency and vice versa. The graph looks consistent for all groups of NMHC. Figure 4 also shows that a high concentration of O_3 happens when the frequency of NMHC is low. This situation also happens for NO_2 against O_3.

This experiment only selects the intervals with a high O_3 (above 0.066 ppm) start with cluster 44 until 47, with total cases of 1801 as a new dataset. The corresponding interval for NO_2 and NMHC also were selected. Then, a new data with high O3 will be mined using Apriori algorithm to extract the meaningful hidden pattern in a data.

C. Apriori Algorithm in Association Rule

Association rule is a data mining technique that can generate a frequent pattern in a rules form, which implies a particular relationship among item set. It is learning by finding typical groups of items that frequently occur in a data set. The most critical parameter component in an association rule is minimum support and minimum confidence [15]. This parameter was used to measure the strength of the rules that were generated by the algorithm. The support measure will evaluate the statistical importance of the dataset. A low support rule may likely to be uninteresting in a transaction [16]. It is guidance on how many the itemset appears in the dataset. Meanwhile, minimum confidence is the guidance of how frequent the rule has been found to be true. For example, let *N* be an item set, $N \Rightarrow M$ is an association rule and *Z* is a set of transactions. The support of *N* concerning *Z* is defined as a part of transaction *z* in the dataset which contains the item set *N*, support (*N*). The confidence value of a rule ($N \Rightarrow M$), concerning a set of transactions *Z*, is the part of the transactions that contains *N* which also contains *M*. Thus, minimum support and confidence can be defined as below:

$$confidence\,(N \Rightarrow M) = support\,(N \cup M)/support\,(N)$$

TABLE IV
PSEUDO CODE OF THE APRIORI ALGORITHM FOR RULE MINING

Apriori Algorithm: Pseudocode

Start
L_1 = {frequent itemset in a database, *D*}
for (*k*=1; L_k !=Ø; *k*++)
do begin C_{k+1} = candidates generated from L_k
for each transaction *t* in the database **do**
increment the count of all candidates in C_{k+1} that are contained in *t*
L_{k-1} = candidates in C_{k+1} with minimum support
End
Return $\cup_k L_k$

The popular algorithms in association rules are Apriori. Apriori were functions by using a "bottom-up" approach by finding a repeated individual item that frequently occurs in a database and this process were extended to a more substantial item set [16]. Similar research had been done in [17]. However, the method used and the type of pollution

were different. The researcher used the Apriori algorithm to analyze extreme Particulate Matter condition in Jing-Jin-Ji, a China region, based on spatial and temporal. Another similar research background in [20] was conducted to find the relations between atmospheric pollution and climatological conditions using association rules with the intention to implement those rules in a forecasting model. This shows that the Apriori-based association rules was so popular and had been done with a different purpose and objective. Table 4 above presents pseudo-code for the Apriori algorithm. Based on table 4, C_k represents a candidate in item set *k*, and L_k represents a frequent item in item set *k*.

III. Result and Discussion

In this study, the proposed method was applied to the climatological O_3 data to find a set of essential rules between the selected attribute using the Apriori technique. The experiment involved a relationship between three main attributes namely NO_2, NMHC, and O_3. The experiment was conducted using the Waikato Environment for Knowledge Analysis (WEKA) software. The selection value for minimum confidence level and minimum support level can affect the rules generation. Therefore, the selection value for this two parameter is very crucial in producing meaningful rules. Commonly, the rules generated with high confidence value was given a top priority in the selection process because they are considered strong, but this method does not provide an opportunity to discover an odd data. Therefore, this study sets the confidence value from 60% to 100% and the minimum support value is set as 0.1. This is to ensure that the frequent and meaningful rules at low confidence and support values can be discovered.

The result was generated by if-then rules which are the scientific statement that connecting variables to form a conclusion. A rules statement has two-part, which is premise (antecedent), and conclusion (consequent) part. The factors (NO_2 and NMHC) have been forced to belong to the antecedent and O_3 were set as a consequent. After rules were generated, the evaluation process needs to be done to ensure that the rules generated are meaningful for decision-making. Some rules may be less meaningful for decision-making. Therefore, these rules should be analyzed to ensure that only meaningful rules are taken. The most significant rules were selected and presented in table 5 below. Only 17 rules were selected among the rules found by the Apriori and it can be noticed that there is no rule of cluster 47 for O_3 (range 0.1-0.171). The rules disclose that NO_2 provides more information about the O_3 formation than NMHC. However, there is a high frequency of low NMHC affecting O_3 concentration.

The graph in figure 5 shows that NO_2 and NMHC concentration were low at the high O_3 level and O_3 concentration are contradicted with NO_2 and NMHC concentration. However, NMHC increased at hour 22. This graph supports the rules finding above which is the frequency for the rule "If NO_2 and NMHC are low, then O_3 is high" is more frequent. Also, there is no NMHC more than 0.55 and NO_2 more than 0.039 in generated rules. The graph also shows that at hour 21, NMHC started to increase but did not affect the O_3 concentration and NO_2 remained the same.

TABLE V
Description of Rules Found by Apriori Algorithm for Group Cluster O_3 44, 45 and 46

Rules O_3 = 44 (0.066-0.072)		**Frequency**
IF	NMHC=1 (0-0.28) **OR**	603
	NMHC=2 (0.29-0.55) **OR**	152
	NO_2=1 (0.00-0.01) **OR**	231
	NO_2=2 (0.011-0.018) **OR**	181
	NO_2=3 (0.019-0.026) **OR**	190
	NO_2=4 (0.027-0.039) **OR**	132
	NO_2=1 (0.00-0.01) **AND** NMHC=1 (0-0.28) **OR**	203
	NO_2=2 (0.011-0.018) **AND** NMHC=1 (0-0.28) **OR**	162
	NO_2=3 (0.019-0.026) **AND** NMHC=1 (0-0.28) **THEN** O_3 = 44 (0.066-0.072)	143
Rules O_3 = 45 (0.073-0.082)		**Frequency**
IF	NMHC=1 (0-0.28) **OR**	408
	NMHC=2 (0.29-0.55) **OR**	137
	NO_2=1 (0.00-0.01) **OR**	137
	NO_2=2 (0.011-0.018) **OR**	149
	NO_2=3 (0.019-0.026) **OR**	129
	NO_2=1 (0.00-0.01) **AND** NMHC=1 (0-0.28)) **OR**	120
	NO_2=2 (0.011-0.018) **AND** NMHC=1 (0-0.28) THEN **O_3** = 45 (0.073-0.082)	127
Rules O_3 = 46 (0.083-0.099)		**Frequency**
IF	NMHC=1 (0-0.28) **THEN** O_3 = 46 (0.083-0.099)	85

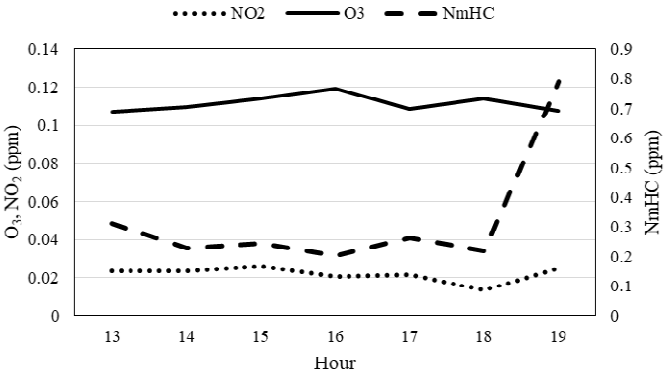

Fig. 5. Graph average in 24 hours of NO_2 and NMHC during high O_3

IV. Conclusions

In this study, a data mining method was proposed in preparing the dataset in order to discover meaningful information among correlated real-world climatological O_3 data around Klang areas. Then, the Apriori-based association rules techniques were applied to a dataset to generate a set of rules. The generated rules were selected based on several criteria and relevance. Only 17 rules were extracted to shows a pattern for high O_3 concentration around Klang station that can answer the research objective. The selected rules stated in Table 5 proved the research findings in [4] and [21] that the concentration of O_3 influenced by the amount of NO_2 and NMHC because of oxidation and photochemical process, with the exact threshold values for each parameter. These

rules provide essential knowledge, especially in controlling the main factor that influences O_3 formation to balance the ecosystem. Also, the generated rules can be implemented in a prediction model to forecast O_3 levels in the future. Besides, this study provides a pattern of NO_2, NMHC, and O_3 that can be used as a reference for the environmental department.

ACKNOWLEDGMENT

This research is funded by Ministry of Higher Education research grant (No: FRGS/1/2016/ICT02/UKM/02/8).

REFERENCES

[1] R. Atkinson et al., "Long-term exposure to ambient ozone and mortality: A quantitative systematic review and meta-analysis of evidence from cohort studies," in *Epidemiology*, 2016, vol. 6.

[2] Rozalina Chuturkova, "Ozone and ozone precursors in urban atmosphere," *Journal scientific and applied research*, 2015, vol. 8, pp. 31-39.

[3] C. Gavrila, A. Coman, I. Gruia, F. Ardelean, and A. Vartires, "Prediction method applied for the evaluation of the tropospheric ozone concentrations in Bucharest," *Rom. Journ. Phys.*, Bucharest, 2016, vol. 61, pp. 1067–1078.

[4] Fatimah Ahmad, Mohd Talib Latif, Rosy Tang, Liew Juneng, Doreena Dominick, and Hafizan Juahir, "Variation of surface ozone exceedance around Klang Valley, Malaysia," *Atmospheric Research*, 2014, vol. 139, 116-127.

[5] Mahmoud Sammour and Zulaiha Ali Othman, "An Agglomerative Hierarchical Clustering with Various Distance Measurements for Ground Level Ozone Clustering in Putrajaya, Malaysia," *International Journal on Advanced Science, Engineering and Information Technology*, 2016, vol. 6, pp. 1127-1133.

[6] D. Adhikary, and S. Roy, "Mining quantitative association rules in real-world databases: A review," *1st International Conference on Computing and Communication Systems*, 2015, vol. 1, pp. 87-92.

[7] J. Caiado, E.A. Maharaj, and P. D'urso, *Time series clustering*. Handbook of cluster analysis. Chapman and Hall/CRC, Boca Raton, Florida Google Scholar. 2015.

[8] S. Aghabozorgi, A.S. Shirkhorshidi and T.Y. Wah, "Time-series clustering: A decade review," *Information Systems*, 2015, vol. 53, pp.16-38.

[9] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, and J. C. Riquelme, *Quantitative association rules applied to climatological time series forecasting*, Springer-Verlag Berlin Heidelberg, 2009 pp. 284–291.

[10] C. Tew, C. Giraud-Carrier, K. Tanner and S. Burton, "Behavior-based clustering and analysis of interestingness measures for association rule mining," *Data Mining and Knowledge Discovery*, 2014, vo. 28, pp. 1004-1045.

[11] K. N. Jallad, and Cynthia Espada-Jallad, "Analysis of ambient ozone and precursor monitoring data in a densely populated residential area of Kuwait," *Journal of Saudi Chemical Society*, 2010, vol. 14, pp. 363–372.

[12] Ghassan Saleh Al-Dharhani, Zulaiha Ali Othman, Azuraliza Abu Bakar, and Sharifah Mastura Syed Abdullah, "Fuzzy-based shapelets for mining climate change time series patterns," *Advances in Visual Informatics*, Lecture Notes in Computer Science, 2015, vol. 9423, pp. 38-50.

[13] World Health Organisation. Review of evidence on health aspects of air pollution - REVIHAAP project: final technical report. 2013 [10 April 2017].

[14] T. Tassa, "Secure mining of association rules in horizontally distributed databases," *IEEE Transactions on Knowledge and Data Engineering*, 2014, vol. 26(4), pp. 970-983.

[15] M. Kantardzic, "Data mining: concepts, models, methods, and algorithms," *John Wiley & Sons*, 2011.

[16] J. Dongre, G. L. Prajapati, and S. V. Tokekar, "The role of Apriori algorithm for finding the association rules in Data mining," *IEEE International conference In Issues and Challenges in Intelligent Computing Techniques (ICICT)*, 2014, pp. 657-660.

[17] S. Qin, F. Liu, C. Wang, Y. Song, and J. Qu. "Spatial-temporal analysis and projection of extreme particulate matter (PM10 and PM2. 5) levels using association rules: A case study of the Jing-Jin-Ji region, China," *Atmospheric Environment*, 2015, vol. 120, 339-350.

[18] J. R. Horne, "Impact of global climate change on ozone, particulate, and secondary organic aerosol concentrations in California: a model perturbation analysis," University of California, Irvine, 2015.

[19] M. Martínez-Ballesteros, A. Troncoso, F. Martínez-Álvarez, and J. C. Riquelme, "Mining quantitative association rules based on evolutionary computation and its application to atmospheric pollution," *Integrated Computer-Aided Engineering*, 2010, vol. 17 (3), 227-242.

[20] M. Martínez-Ballesteros, S. Salcedo-Sanz, J. C. Riquelme, C. Casanova-Mateo, and J. L. Camacho, "Evolutionary association rules for total ozone content modeling from satellite observations," *Chemometrics and Intelligent Laboratory Systems*, 2011, vol. 109 (2), 217-227.

[21] S. N. Matsunaga, S. Chatani, T. Morikawa, S. Nakatsuka, J. Suthawaree, Y. Tajima, and H. Minoura, "Evaluation of non-methane hydrocarbon (NMHC) emissions based on an ambient air measurement in the Tokyo area, Japan," *Atmospheric Environment*, 2010, vol. 44 (38), 4982-4993.