International Journal on Advanced Science Engineering Information Technology

Vol.6 (2016) No. 6
ISSN: 2088-5334

# An Agglomerative Hierarchical Clustering with Various Distance Measurements for Ground Level Ozone Clustering in Putrajaya, Malaysia

Mahmoud Sammour[#1], Zulaiha Othman[#2]

[#]*Center for Artificial Intelligence Technology, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia 43600 Bangi, Selangor Darul Ehsan, Malaysia University, Address, City, ZIP Code, Country*
*E-mail: [1]mahmoud.samour@gmail.com, [2]zao@ukm.edu.my*

***Abstract*— Ground level ozone is one of the common pollution issues that has a negative influence on human health. The key characteristic behind ozone level analysis lies on the complex representation of such data which can be shown by time series. Clustering is one of the common techniques that have been used for time series metrological and environmental data. The way that clustering technique groups the similar sequences relies on a distance or similarity criteria. Several distance measures have been integrated with various types of clustering techniques. However, identifying an appropriate distance measure for a particular field is a challenging task. Since the hierarchical clustering has been considered as the state of the art for metrological and climate change data, this paper proposes an agglomerative hierarchical clustering for ozone level analysis in Putrajaya, Malaysia using three distance measures i.e. Euclidean, Minkowski and Dynamic Time Warping. Results shows that Dynamic Time Warping has outperformed the other two distance measures.**

***Keywords*— agglomerative hierarchical clustering; Dynamic Time Warping; ozone analysis; time series**

## I. INTRODUCTION

Putrajaya is one of the developed cities located in Malaysia. With the dramatic economic development and population expansion, several environmental pollution issues have arisen. One of these issues is the increasing of Ozone pollution. Apparently, such increment has a significant impact on the human health [1]. Several stations have been employed nowadays to observe the ozone trends. In order to analyze such trends, machine learning techniques especially clustering technique can be considered as a great opportunity in terms of detecting significant patterns. Clustering aims to aggregate similar data points within clusters [2]. In this manner, similar trends could be aggregated in a single group which facilitates the cause analysis. However, the key challenge behind clustering ozone levels lies on the representation in which the data is being represented in time manner [3].

Time series has emerged as a response to the data evolution of chronological representation where the data been made in time intervals [4]. There are many kinds of time series data such as financial, weather forecasting, pattern recognition, etc. [5]. The common task of time series data mining is the process of identifying similar sequences. Such process is performed using clustering techniques.

There are many clustering techniques could be used in this task. One of the common clustering technique is hierarchical clustering which has been considered as the state of the art for various environmental and metrological data in the literature [6]-[8]. Hierarchical clustering aims to build a hierarchy of clusters in which the data points are being initialized as one cluster and then split into multiple clusters (Divisive Hierarchical Clustering), or each data point could be initialized as a cluster and then merged into a smaller number of clusters (Agglomerative Hierarchical Clustering) [9].

On the other hand, the similarity or distance function used by the clustering technique plays an essential role in terms of the performance of the clustering task [10]. Several distance measures have been proposed including Euclidean, Minkowski and Dynamic Time Warping distance measures. In fact, integrating an appropriate distance measure with an appropriate clustering technique is a challenging task [10]. Therefore, identifying suitable distance measure for ozone level clustering represents a vital demand process. This paper aims to conduct a comparative analysis between three distance measures including Euclidean distance (ED), Minkowski distance (MD) and Dynamic Time Warping (DTW) using Agglomerative Hierarchical Clustering (AHC).

## II. Material and Method

Various studies have tackled the problem of detecting the ozone trend, for instance, Solazzo et al. [11] have conducted a comprehensive analysis for surface-level ozone based on air quality in Europe and North America in which an ensemble clustering approach have been used to group the similar data.

On the other hand, Saithanu & Mekparyup [8] have proposed an agglomerative hierarchical clustering with Euclidean distance measure for clustering ozone level at the east of Thailand. In their study, the authors have concentrated on the significant factors that lead to increasing the ozone level such as temperature, wind direction, humidity and wind speed.

Similarly, Austin et al. [12] have concentrated on factors associated with ozone levels such as temperature, pressure and sea level for identifying ozone detection using k-means clustering. The data used in such study is a daily data collected from Boston Logan airport. In their studies, the authors have attempted to identify the most appropriate number of clusters. Results showed that five number of clusters has obtained the superior performance.

In addition, Malley et al. [7] have proposed a Hierarchical Clustering Analysis (HCA) with non-negative matrix factorization for classifying ozone level in Europe. Multiple datasets have been used in such study related to ozone variation measurements for the period of 1991-2010. The grouping clustering has been used to identify relationships influence the ozone levels.

Finally, Ahmadi et al. [6] have applied two kinds of clustering including k-means clustering and agglomerative hierarchical clustering for ozone level analysis. Basically, k-means has been used firstly in order to detect significant patterns of ozone. Then, agglomerative hierarchical clustering has been used to identify hourly ozone patterns. Finally, multiple regression tasks have taken a place in order to predict ozone based on seasons and zones.

### A. Proposed Method AHC with DTW

The proposed method of this study is an agglomerative hierarchical clustering that has been carried out as a complete/maximum linkage with Dynamic Time Warping as a distance measure. The application of the proposed method has been performed to classify the ground level ozone in Putrajaya, Malaysia for the year 2006.

### B. Research Method

The research method of this study consists of five main phases as shown in Fig. 1. The first phase is data which discusses the collection, details and characteristics of the data used in the experiments. The second phase is preprocessing which discusses the cleaning tasks that have been performed to turn the data into an appropriate form. The third phase is clustering which discusses the application of agglomerative hierarchical clustering. The fourth phase discusses the distance measures including ED, MD and DTW. Finally, the fifth phase is evaluation in which the clustering results are being validated using certain evaluation method.

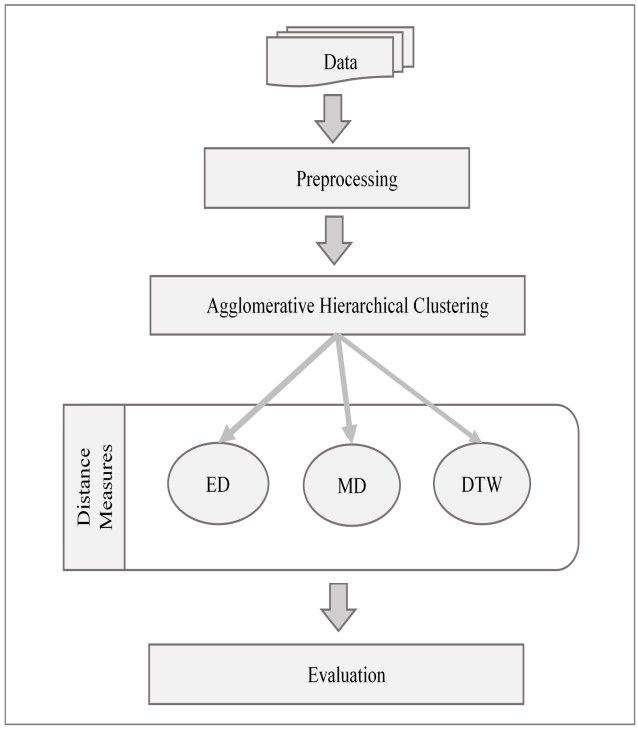


Fig. 1 Research method of the study

### C. Data

Data has been collected from LESTARI [13] which is the Institution for Environment and Development in Malaysia and the Asia Pacific. Such 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 contains ozone levels for one year (i.e. 2006) particularly for Putrajaya city. The data has been represented hourly as time intervals, which contained 8544 instances.

### D. Preprocessing

This phase aims to prepare the data in order to be more suitable for processing. Basically, each data includes irrelevant data, noisy and uncompleted instances. Handling such data plays an essential role in terms of improving the performance of clustering process [14]. Hence, two tasks have been proposed for this purpose; cleaning and discretization. Cleaning aims to handle the missing values and the calibration errors where such values has the ability to cause incorrect matches in the process of clustering [15]. In this manner, Microsoft Excel has been used to detect such values in which the 158 missing values and 431 calibration errors have been identified and dealt with by Matlab ANN prediction algorithm. Whereas, discretization task aims to limit the class values within a specific interval. Such interval will facilitate the process of clustering where the values will be reduced into a particular range. Such process of discretization is essential for specific algorithms such as hierarchical clustering [16].

### E. Hierarchical Clustering

This phase aims to apply a hierarchical clustering technique. In general, hierarchical clustering algorithms work by aggregating the objects into a tree of clusters [17].

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 set in a separated cluster then AHC will merge such clusters into larger clusters [2]. Such process is continuing until a specific termination has been reached. Complete linkage algorithm aims to identify the similarity between two clusters by measure two nearest data points that are located in different clusters. Hence, the merge will be done between the clusters that have minimum distance -most similar- between each other.

In this paper, AHC has been applied as a maximum linkage with three distance measures including Euclidean Distance (ED) [18], Minkowski Distance (MD) [19] and Dynamic Time Warping (DTW) [20], these measures are illustrated in the next sub-section.

### *F. Distance Measures*

The key characteristic behind clustering process lies on the function that will be used to identify the similarity between two data. Such data varies where it could be formed as raw values of equal or non-equal length, or it could be formed as vector space of feature-pairs [21].

For Euclidean distance, let $x_i$ and $v_j$ be a P-dimensional vector, then the Euclidean distance can be measured as [21]:

$$d_E = \sqrt{\sum_{k=1}^{P} \left(x_{ik} - v_{jk}\right)^2} \quad (1)$$

For Minkowski distance, Let $x_i$ and $v_j$ be a P-dimensional vector, Minkowski distance is a generalization of Euclidean distance, which is computed as follows [21]:

$$d_M = \sqrt[q]{\sum_{k=1}^{P} \left(x_{ik} - v_{jk}\right)^q} \quad (2)$$

where q is a positive integer.

On the other hand, DTW has been widely used to compare between discrete sequences and sequences of continuous values [21]. Let $S = \{s_1, s_2, \dots, s_i, \dots, s_n\}$ and $T = \{t_1, t_2, \dots, t_j, \dots, t_n\}$ be a two time series sequences. DTW will minimize the differences among these series by representing a matrix of $n \times m$ [22]. In such 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 [22]. 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 [23]. The warping path that has the minimum distance/similarity between the two series is of interest. Hence, the DTW can be computed as follows:

$$d_{DTW} = min \frac{\sum_{k=1}^{K} p_K}{K} \quad (3)$$

## III. Results and Discussion

### *A. Evaluation*

One of the challenging tasks behind clustering is evaluating its results in which the question 'what is the best way to group the data' should be clarified [24]. Two main approaches have been proposed for validating clustering process; external and internal validation of clusters [25]. External validation aims to validate the clusters based on the distribution in which the common information retrieval metrics such as precision, recall, and f-measure. However, such mechanism of validation relies on a labeled data. Since, the real-life data is usually unlabeled thus, applying external validation tend to be insufficient. On the 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, as well as, 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 within multiple clusters. 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, best results associated with a smaller value of RMSE-SD among intra-cluster, and with a greater value of RMSE-SD among inter-clusters.

### *B. Experiments*

The experiments have been conducted using C# programming language in which the data has been transformed into columns and eliminating the noisy data. In addition, the agglomerative hierarchical clustering has been performed with max-linkage using the three distance measures including Euclidean, Minkowski and DTW. The clustering was performed using a multiple number of clusters as parameters with a range of 3-15 number of clusters. Such ranged has been set as a result of analyzing the data and identifying the appropriate classes.

In this section, the results of the proposed AHC using ED, MD and DTW are being declared. Basically, the results have been obtained based on a multiple number of clusters. Based on the observation of data, the number of clusters should be ranged from 3-15. Table 1 shows the results for intra-cluster and Table 2 shows the results of inter-clusters.

TABLE I
RESULTS OF RMSE-SD FOR INTRA-CLUSTER

| # Clusters | ED | MD | DTW |
|---|---|---|---|
| 15 | 0.03703 | 0.027 | 0.0118 |
| 14 | 0.03573 | 0.02436 | 0.0121 |
| 13 | 0.0366 | 0.02434 | 0.0125 |
| 12 | 0.03607 | 0.02476 | 0.0133 |
| 11 | 0.03562 | 0.02487 | 0.0118 |
| 10 | 0.03144 | 0.02375 | 0.0121 |
| 9 | 0.03125 | 0.02266 | 0.0124 |
| 8 | 0.02907 | 0.02125 | 0.013 |
| 7 | 0.03001 | 0.02043 | 0.0137 |
| 6 | 0.02848 | 0.02302 | 0.0145 |
| 5 | 0.02294 | 0.01471 | 0.0102 |
| 4 | 0.02111 | 0.01766 | 0.0103 |
| 3 | **0.00954** | **0.0127** | **0.0039** |

As shown in Table 1, the minimum results of RMSE-SD have been obtained at 3 number of clusters for ED, MD and DTW by achieving 0.00954, 0.0127 and 0.0039 respectively. As mentioned earlier, the smaller value of RMSE-SD for intra-cluster leads to better performance. Therefore, 3 number of cluster is the most accurate one. However, DTW has shown the smallest value of RMSE-SD which compared to the other distance measures. This means that DTW has outperformed both ED and MD for the intra-cluster.

TABLE II
RESULTS OF RMSE-SD FOR INTER-CLUSTER

| # Clusters | ED | MD | DTW |
|---|---|---|---|
| 15 | 0.098 | 0.222 | 0.29 |
| 14 | 0.101 | 0.2304 | 0.26 |
| 13 | 0.104 | 0.2399 | 0.27 |
| 12 | 0.104 | **0.253** | 0.27 |
| 11 | 0.107 | 0.2494 | 0.28 |
| 10 | **0.118** | 0.252 | 0.3 |
| 9 | 0.103 | 0.2509 | 0.29 |
| 8 | 0.104 | 0.2343 | 0.25 |
| 7 | 0.103 | 0.2338 | 0.24 |
| 6 | 0.086 | 0.2513 | 0.27 |
| 5 | 0.091 | 0.2015 | 0.29 |
| 4 | 0.092 | 0.139 | 0.29 |
| 3 | 0.026 | 0.0786 | **0.34** |

As shown in Table 2, the maximum value of RMSE-SD for ED was at 10 number of clusters by achieving 0.118, for MD at 12 number of clusters by achieving 0.253, and for DTW at 3 number of clusters by achieving 0.34. As mentioned earlier, the maximum value of RMSE-SD for inter-clusters leads to better performance. By comparing the three values of RMSE-SD for the three distance measure, it is obvious that DTW has the greatest value. This means that DTW has outperformed the other distance measures for the inter-clusters. Fig. 2 shows the performances of the three distance measures for both intra-cluster and inter-clusters.

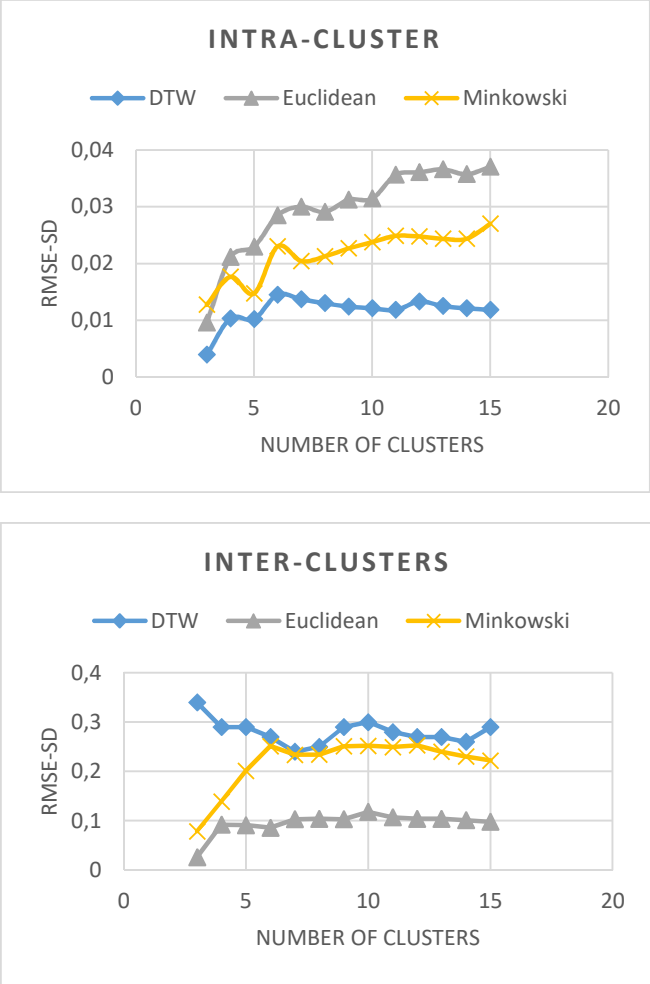


Fig. 2 Performances of the three distance measures for both intra-cluster and inter-clusters

As shown in Fig. 2, DTW has outperformed the other distance measures by obtaining greatest values of RMSE-SD in terms of intra-cluster evaluation for all number of clusters. Similarly, DTW has outperformed the other distance measures in terms of inter-clusters by obtaining the lowest values of RMSE-SD for all number of clusters.

Basically, comparing the results of the proposed AHC with DTW against the related work seems to be a challenging task due to multiple reasons. First, datasets used in the related work are different. Second, the evaluation of clustering is varying among the studies. Third, the aims of applying the clustering are also different in which some studies were addressing relationships that influence the

variation of ozone level. Finally, the number of years and the covered regions are differing among studies.

However, since Euclidean and Minkowski distance measures have been used with AHC in the related work, it can be concluded that the proposed DTW with AHC has shown competitive performance.

### *C. Discussion*

The US Office of Air and Radiation [26] 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 manner, AirNow (2009) has provided 5 categories of air pollution which are shown in Table 3.

TABLE III
CATAGORIES OF AIR POLLUTION

| |
|---|
| Very Unhealthy |
| Unhealthy |
| Unhealthy for Sensitive Groups |
| Moderate |
| Good |

In order to provide 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 AirNow (2009) categorization will be considered. Therefore, a comparison is being held between the two number of clusters 5 and 9. The comparison will be based on multiple variables including starting values of ozone, maximum peak, maximum peak of median and ending values. Table 4 shows the values of 5 number of clusters.

As shown in Table 4, the number of days included in the 'unhealthy' category is nearly representing the half of the year which seems to be overestimated categorization. This means that this category should be divided into more categories. Whereas, the 'moderate' category contains only eight days which also seems to be underestimated categorization. Generally, this category is supposed to contain more days. However, Table 5 shows the values of 9 number of cluster.

TABLE IV
VALUES OF 5 NUMBER OF CLUSTER

| **#Days** | **K=5** | **Morning** | | **Afternoon** | | **Even ing** | |
|---|---|---|---|---|---|---|---|
| | Class | Start | end | Max | Men Max | end | |
| 173 | 1 | 0.004 | 0.005 | 0.115 | 0.061 | 0.008 | Un healthy |
| 19 | 2 | 0.014 | 0.005 | 0.148 | 0.113 | 0.014 | Very Un healthy |
| 104 | 3 | 0.017 | 0.007 | 0.105 | 0.58 | 0.012 | Un healthy for |
| 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 |

TABLE V
VALUE OF 9 NUMBER OF CLUSTER

| **#Days** | **K=5** | **Morning** | | **Afternoon** | | **Even ing** | |
|---|---|---|---|---|---|---|---|
| | Class | Start | end | Max | Men Max | end | |
| 52 | 1 | 0.008 | 0.005 | 0.115 | 0.076 | 0.007 | Un healthy |
| 121 | 2 | 0.004 | 0.005 | 0.096 | 0.055 | 0.009 | low Moderate |
| 19 | 3 | 0.014 | 0.005 | 0.148 | 0.113 | 0.014 | Very Un healthy |
| 63 | 4 | 0.015 | 0.006 | 0.093 | 0.053 | 0.007 | Moderate |
| 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 |
| 21 | 7 | 0.019 | 0.011 | 0.105 | 0.078 | 0.019 | Very Un healthy for Sensitive Groups |
| 8 | 8 | 0.033 | 0.016 | 0.093 | 0.059 | 0.017 | Un healthy for Sensitive Groups |
| 20 | 9 | 0.025 | 0.007 | 0.077 | 0.062 | 0.014 | High Moderate |

As shown in Table 5, unlike the standard 5 categorizations, the 9-categorizaiton 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 as 'unhealthy' and 'very unhealthy for sensitive group'. These categories have shown reasonable contained number of days. In addition, the category 'moderate' has been split into three categories as 'high moderate', 'moderate' and 'low moderate'. Similarly, these categories have contained a reasonable number of days. Finally, the category 'good' has been also divided into two categories as 'very good' and 'good'. However, Fig. 3 and Fig. 4 show the distribution of categories over the number of days.

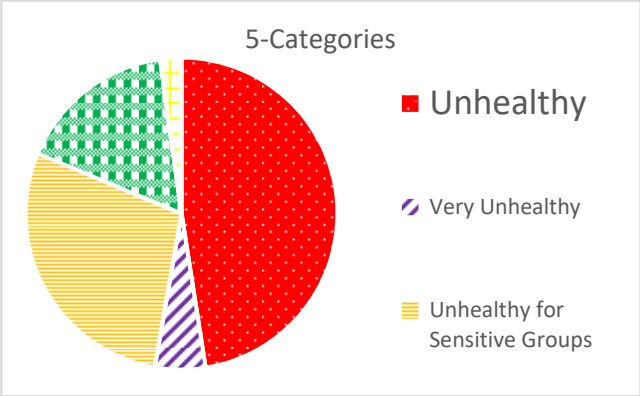


Fig 3 Distribution of days over the five categories

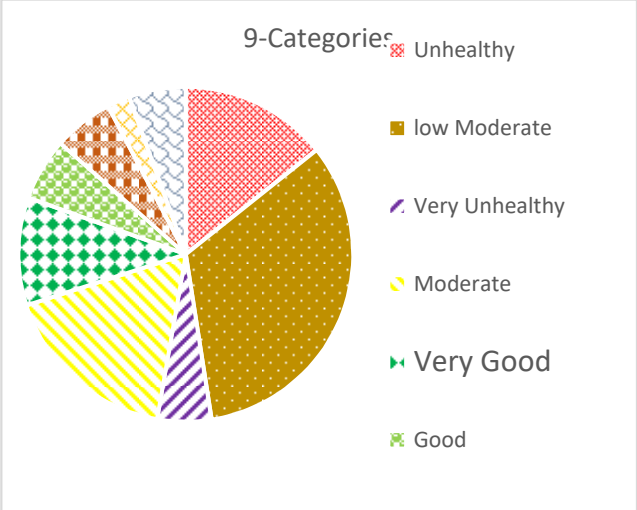


Fig 4 Distribution of days over the nine categories

## IV. Conclusions

This paper has conducted a comparative study between three distance measures including Euclidean Distance (ED), Minkowski Distance (MD) and Dynamic Time Warping (DTW) for clustering ozone level in Putrajaya, Malaysia using Agglomerative Hierarchical clustering (AHC). Data used in this paper is an hourly observation of ozone level for one year (i.e. 2006). Results showed that DTW has superior performance compared to the other two distance measures. In future direction, conducting a comparative analysis of different clustering techniques such as k-means, k-medoids, density-based and others, would contribute toward improving the effectiveness of clustering results.

## Acknowledgment

This study is supported by the University Kebangsaan Malaysia (UKM) and funded by a research grant (AP2013-007).

## References

[1] A Monteiro, A Carvalho, I Ribeiro, M Scotto, S Barbosa, A Alonso, JM Baldasano, MT Pay, AI Miranda, and C Borrego, "Trends in ozone concentrations in the Iberian Peninsula by quantile regression and clustering," Atmospheric environment, vol. 56, pp. 184-193, 2012.

[2] Sangeeta Rani and Geeta Sikka, "Recent techniques of clustering of time series data: A Survey," *Int. J. Comput. Appl*, vol. 52, pp. 1-9, 2012.

[3] Michael Steinbach, Pang-Ning Tan, Vipin Kumar, Steven Klooster, and Christopher Potter, "Discovery of climate indices using clustering," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003, pp. 446-455.doi.

[4] Tak-chung Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, pp. 164-181, 2011.

[5] Axel Wismüller, Oliver Lange, Dominik R Dersch, Gerda L Leinsinger, Klaus Hahn, Benno Pütz, and Dorothee Auer, "Cluster analysis of biomedical image time-series," *International Journal of Computer Vision*, vol. 46, pp. 103-128, 2002.

[6] Mahdi Ahmadi, Yan Huang, and Kuruvilla John, "Predicting Hourly Ozone Pollution in Dallas-Fort Worth Area Using Spatio-Temporal Clustering," 2015.

[7] *Christopher S Malley, Christine F Braban, and Mathew R Heal, "The application of hierarchical cluster analysis and non-negative matrix factorization to European atmospheric monitoring site classification," Atmospheric Research, vol. 138, pp. 30-40, 2014.*

[8] K Saithanu and Jatupat Mekparyup, "Clustering of Air Quality and Meteorological Variables Associated with High Ground Ozone Concentration in the Industrial Areas, at the East of Thailand," *International Journal of Pure and Applied Mathematics*, vol. 81, pp. 505-515, 2012.

[9] P. Berkhin, "A Survey of Clustering Data Mining Techniques," in *Grouping Multidimensional Data*, J. Kogan, C. Nicholas, and M. Teboulle, Eds., ed: Springer Berlin Heidelberg, 2006, pp. 25-71.

[10] Konstantinos Kalpakis, Dhiral Gada, and Vasundhara Puttagunta, "Distance measures for effective clustering of ARIMA time-series," in *Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on*, 2001, pp. 273-280.doi

[11] Efisio Solazzo, Roberto Bianconi, Robert Vautard, K Wyat Appel, Michael D Moran, Christian Hogrefe, Bertrand Bessagnet, Jørgen Brandt, Jesper H Christensen, and Charles Chemel, "Model evaluation and ensemble modelling of surface-level ozone in Europe and North America in the context of AQMEII," *Atmospheric environment*, vol. 53, pp. 60-74, 2012.

[12] Elena Austin, Antonella Zanobetti, Brent Coull, Joel Schwartz, Diane R Gold, and Petros Koutrakis, "Ozone trends and their relationship to characteristic weather patterns," *Journal of Exposure Science and Environmental Epidemiology*, 2014.

[13] LESTARI. (2016). *The Institute for Environment and Development in Malaysia and the Asia Pacific*. Available: http://www.ukm.my/lestari

[14] Dino Isa, Lam Hong Lee, VP Kallimani, and Rajprasad Rajkumar, "Text document preprocessing with the Bayes formula for classification using the support vector machine," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 20, pp. 1264-1272, 2008.

[15] Ramesh SV Teegavarapu and V Chandramouli, "Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records," *Journal of Hydrology*, vol. 312, pp. 191-206, 2005.

[16] Sumi S Monira, Zaman M Faisal, and H Hirose, "Comparison of artificially intelligent methods in short term rainfall forecast," in *Computer and Information Technology (ICCIT), 2010 13th International Conference on, 2010*, pp. 39-44.doi.

[17] M. Aghagolzadeh, H. Soltanian-Zadeh, B. Araabi, and A. Aghagolzadeh, "A Hierarchical Clustering Based on Mutual Information Maximization," in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, 2007, pp. I - 277-I - 280.doi:10.1109/ICIP.2007.4378945.

[18] Per-Erik Danielsson, "Euclidean distance mapping," *Computer Graphics and image processing*, vol. 14, pp. 227-248, 1980.

[19] Patrick JF Groenen and Krzysztof Jajuga, "Fuzzy clustering with squared Minkowski distances," *Fuzzy Sets and Systems*, vol. 120, pp. 227-237, 2001.

[20] Donald J Berndt and James Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," in *KDD workshop*, 1994, pp. 359-370.doi.

[21] T Warren Liao, "Clustering of time series data—a survey," *Pattern recognition*, vol. 38, pp. 1857-1874, 2005

[22] François Petitjean, Alain Ketterlin, and Pierre Gançarski, "A global averaging method for dynamic time warping, with applications to clustering," *Pattern Recognition*, vol. 44, pp. 678-693, 2011.

[23] Vit Niennattrakul and Chotirat Ann Ratanamahatana, "On clustering multimedia time series data using k-means and dynamic time warping," in *Multimedia and Ubiquitous Engineering, 2007. MUE'07. International Conference on*, 2007, pp. 733-738.doi.
[24] Andreas Rauber¹, Jan Paralic, and Elias Pampalk¹, "Empirical Evaluation of Clustering Algorithms*.
[25] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu, "Understanding of internal clustering validation measures," in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, 2010, pp. 911-916.doi.
[26] AirNow. (2009). *Ozone and Your Health*. Available: https://cfpub.epa.gov/airnow/index.cfm?action=ozone_health.index#request.PDFPath#ozone-c.pdf