MONITORING AND ANALYSIS OF PM10 DIURNAL VARIATION AND ITS SPATIAL DISTRIBUTION IN PENINSULAR MALAYSIA USING FUNCTIONAL DATA

N. Shaadan[1,*], S.M. Deni[1,] , A.A. Jemain[2]

[1]*Center for Statistics and Decision Science Studies, Faculty of Computer and Mathematical Sciences, UiTM , 40450, Shah Alam, Selangor, Malaysia*

[2]*School of Mathematical Sciences, Faculty of Science and Technology, UKM, 43600, Bangi, Selangor, Malaysia*

[*]For correspondence; Tel. + (60) 0355435323, E-mail: shahida@tmsk.uitm.edu.my

ABSTRACT: *Particulate matter (PM10) can produce harmful effects on human health and the environment in general. Therefore, information on diurnal variation of PM10 is important as it provides insights on exposure time as well as potential sources of the pollutant. This study aims to identify the major pattern of variations in the diurnal PM10 levels over Peninsular Malaysia during the summer monsoon. Using Functional Principal Component Analysis, it shows that Peninsular Malaysia experiences three major patterns of PM10 diurnal variation. The first mode characterized the day-to-day vertical shift in the level that dominantly occupied the western coastal region which is strongly contributed by PM10 concentration during morning and late evening busy hours. The second mode exhibits the day and night contrast of the level which is likely to be the result of the photochemical reaction during the daylight time and is dominant in the northwest coastal region. The third mode features a shift in the level at around 3.00 pm and achieved the maximum at 8.00 pm as the results of two bimodal peaks of PM10 concentration during the day and night hours. This pattern of variation delineates the northern part of the Malaysian Peninsular. The results have also provided evidence that vehicular emission is the primary source of PM10 pollution in Peninsular Malaysia. Industrial activity and mixing factors are also proven to be the second and third major contributing sources towards PM10 variation.*

Keywords: PM10, diurnal variation, functional data analysis, air pollution

1. INTRODUCTION

The recurrent occurring of PM10 pollution has become a common problem in Southeast Asian Countries, including Malaysia [1]. PM10 pollution is often associated with haze incidences and often reported to occur almost every year. PM10 has become an important pollutant in the country and was proven to be significantly related with respiratory mortality, particularly in the busy areas such as in the Klang Valley [2].

Temporal variation of a pollutant is defined as the variation over time within a certain period, while spatial variation refers to the variation at different locations in a region. The knowledge on the spatial and temporal variation of the pollutant has become an essential input to many fields of research [3]. In the exposure and health related studies for example, both temporal and spatial variation are critical since the information provide better understanding on the representativeness of air quality towards the potential exposure and impact level on human health [4].

In environmental monitoring activities, characterizing the spatial and temporal behaviour of pollutant assists in investigating the formation, transport, and accumulation of the pollutant in the atmosphere [5]. Other than that, spatio-temporal analysis can also provide insights on the sources of PM10, as well as suggesting a better model that can describe PM10 emissions, transport, and atmospheric concentrations [6]. As suggested by [7], the shape of pollutant diurnal curve could also help to ascertain if a site is exposed to the locally generated or long-range transported pollutant.

Several studies have been conducted to investigate the temporal and spatial variation of PM10 worldwide. However, noticeably, the study analyses are often conducted using point average data to explore the variation of PM10. By this approach, consequently, the continuous nature of PM10 level and its dynamic behavior are often ignored in the data analysis. This approach could lead to limited findings (i.e., about the implicit information) due to the insufficiency of input data since the average data are summary values and static in behavior. As for example, in the study by [8], spatio-temporal analysis of PM10 concentration was conducted to investigate seasonal variation for the Malaysian environment. Based on the employment of daily point average data and the conventional principal component analysis (PCA), the findings of the study have led to an understanding of the regionalization nature of PM10 and its seasonal variation over a one year period. However, diurnal variations were not discussed. Therefore, to further increase the knowledge on PM10 diurnal variations for the Malaysian environment, this study was conducted.

In this current study, the analysis is conducted using functional data to incorporate the continuous and dynamic nature of PM10 concentration level by means of Functional Data Analysis (FDA). Functional data are defined as recorded data that arise in a continuum such as time or space. In the context of PM10, the data are defined as the function of time. The other objective of this study is to explore several important modes of diurnal variation that exist in the environment of Peninsular Malaysia and its spatial distribution, as well as their associated contributing sources using Functional Principal Component Analysis (FPCA).

2. METHODOLOGY

In the study analyses, the daily and hourly recorded PM10 data are treated as functional data or curves instead of daily average values. Therefore, the original daily by hourly recorded data must be first converted into functional data or curves before the FPCA is conducted. Figure (1) shows the example of five days PM10 curves after the data conversion process. Data conversion is defined as converting discrete data into functional data or curves.

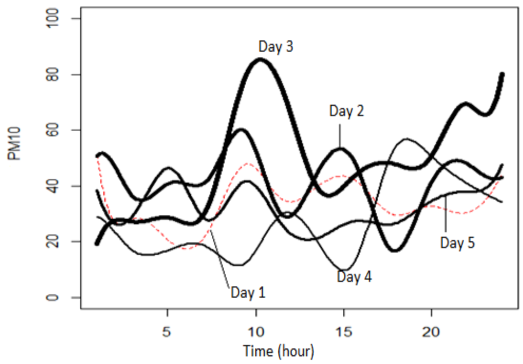


Fig (1) physical forms of five PM10 daily data (curves)

For this study, any day x_i , the PM10 data (y) recorded at time t_j with j=1,…,24 that is $x_i = [y_{t_1}, y_{t_2}, ..., y_{t_{24}}]$, is converted into a continuous function $x_i(t)$ using basis function:

$$x_i(t) = \sum_{k=1}^{K} C_k \phi_k(t) \qquad (1)$$

where the term $\phi_k(t)$ is the basis system consisting of k number of basis functions and C_k is the basis coefficients. The coefficient C_k was determined using the ordinary least squares (OLS) method. After the data were converted into curves (functional data) then analysis to extract the diurnal variations is conducted using FPCA.

FPCA is one of the most important exploratory tools in FDA. The analysis is conducted using mean centered curves because the interest is primarily in characterizing the main deviations of each curve from the average curves [9]. In comparison to FPCA, PCA is known as a classical approach to the exploration of variation in multivariate data, while FPCA is used for functional data or curves. Both methods are generally aimed and used for data reduction, but the key difference is that, PCA uses discrete point data, while FPCA uses functional data. The methods use an eigenvalue decomposition of the covariance matrix to find direction in the observation space along which the data have the highest variability. The direction of variation in PCA is represented by loading vectors, while in the functional context; each principal component is specified by a principal component weight function $\xi(t)$.

In this study, FPCA is conducted using the mean centered curves $z_i(t) = x_i(t) - \bar{x}(t), i = 1,...,30$ for 30 average diurnal curves from 30 air quality monitoring stations in Peninsular Malaysia. A degree three b-spline is chosen as the basis, and k=15 equal number of basis functions were used in the data conversion process. The number of basis k was determined by applying the Bayesian Information Criteria (BIC) to the functional mean curves, $\bar{x}(t)$ of the 30 stations [10]. The main aim of FPCA is to search for several important (principal) components that can describe the major variation in the PM10 curves. Before FPCA is conducted, it is a need to determine how many numbers of the principal components (PC) to be retained so that the components are enough to convey important information in the data. In this case, following the approach by [11], the technique of principal component ranking with respect to eigenvalues using a scree-plot is used.

The first step in FPCA is to find the first eigen-weight function $\xi_1(t)$ that maximizes the mean square of the component score, that is $n^{-1}\sum_i S_{i1}^2$ for which $S_{i1} = \int \xi_1(t) z_i(t) dt, \ i = 1,...,30$ subject to the normality constraint $\int \xi_1(t)^2 dt = 1$. Repeat the subsequent step until the desired number of PC for example *m*. The weight function for the mth PC, the $\xi_m(t)$ is also required to satisfy $\int \xi_m(t)^2 dt = 1$ and $\int \xi_m(t) \xi_k(t) dt = 0, k < m$. Given the variance-covariance function such as:

$$v(s,t) = n^{-1} \sum_{i=1}^{n} z_i(s) z_i(t) \qquad (2)$$

we can obtain a set of eigen-weight functions $\xi(t) = [\xi_1, ..., \xi_m]$ and a set of eigenvalues $\lambda = [\lambda_1, ..., \lambda_m]$ by solving the eigen-equation problem given by:

$$\int v(s,t) \xi(t) dt = \lambda \xi(s) \qquad (3)$$

Each principal component accounts for a different proportion of the variability in the curves which is given by an eigenvalue. The first captures the greatest amount of the variation, the second captures the second greatest amount of the variation, and so on, and are independently indicating different information. In terms of the computational aspect, the eigen-analysis problem is transformed into matrix eigen-analysis task, either by discretizing the functions or using basis function expansion of the functions. In this study, a basis function expansion method was used by adopting the approach of [9]. The fda package in R or Mathlab software may help in the computation of FPCA.

2.3. Application of FDA and FPCA

For the purpose of the application of FPCA, an analysis is conducted based on historical data consisting of the daily by hourly PM10 concentration level from 30 air quality monitoring stations in Peninsular Malaysia. The data were obtained from the Air Quality Division, Malaysian Department of Environment (DOE) for the period of records from 2001-2010 focusing on the summer monsoon season between June and September. The data were recorded as part of a Malaysian Continuous Air Quality Monitoring using the β-ray attenuation mass monitor (BAM-1020) manufactured by Met-One Instruments Inc. Missing values in the data set were replaced using the column median value of the available data [12].

3. RESULTS AND DISCUSSION

Before FPCA was carried out on the 30 mean centered curves from the air quality monitoring stations, the optimal number of principal component that needs to be retained for the data set is first determined. The results given by the scree-plot in figure (2) shows that the first three log eigenvalues seem to be well above the linear trend. Thus, indicating that the leading three principal components are sufficient to explain the important variations in the diurnal curves.

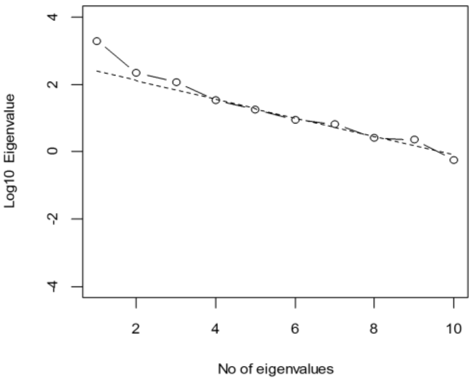


Figure (2) Scree-plots of the important number of PCs

Figure (3 a,b,c) depicts the features of the eigen- weight functions that explain the contribution hours to the PM10 variation for each PC. For PC1, the $\xi_1(t)$ defines the highest mode of variation in the PM10 curves that is positive throughout the day producing trend of a peak during the busy hours and a valley around 3.00 pm and starts to increase again during late evening hours until night. This means that the greatest variability between air quality monitoring stations will be found by heavily weighting morning and late evening rush time hours of PM10 concentration. Supported by [13], this pattern indicates contribution by vehicular activity.

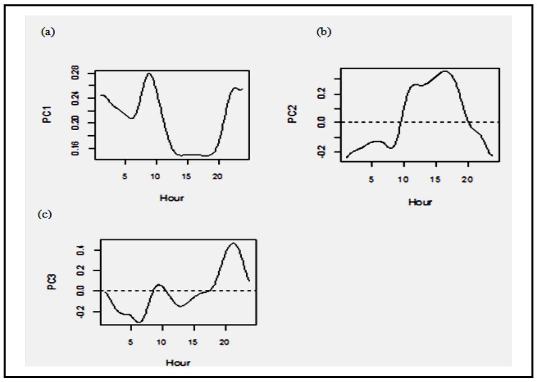


Figure (3) Contribution of time (hour) to the diurnal variation shown by (a) PC1, (b) PC2, and (c) PC3 eigen-weight function curve

The pattern of the second highest variation by eigen-weight function $\xi_2(t)$ consists of positive contribution for the day time hours which start after 10.00 am and end around 5.00 pm and negative contribution at other times. The trend of high contribution hours during the daylight hours given by PC2 coincides with the period of hours where photochemical activity takes place. This pattern of variation also matches the pattern discovered by [13]. Meanwhile, the third eigen-weight function $\xi_3(t)$ shows more cycles of feature giving a bimodal shape of contribution hours. The smaller peak is during rush time morning hours and the second peak is during night time. Positive contribution is identified only an hour before and after 10.00 am and after 5.30 pm, while the largest negative contribution is at 6.00 am in the morning. Supported by [13, 14], the variation described by PC3 is considered as the intensity of the variability in relation to the contribution of the so called mixing factors.

To help interpreting the variation given by each PC for the diurnal curves, the plot of the PC as the perturbations to the mean curve in figure (4 a,b,c) is used. This is done by adding and subtracting a multiple of each PC curve. Figure (4 a,b,c) shows the effect of systematically increasing and decreasing the score of each PC, the solid curve is the mean diurnal curves of the 30 stations. Changes in the score for the PC1 produced a primarily vertical shift with an overall increase in the PM10 concentration which occurred dominantly during morning and evening rush hours. This means that the major station-to-station PM10 diurnal cycle variability in Peninsular Malaysia was dominated by variability of PM10 due to vehicular activities. This pattern of variation and its association with vehicular emission contribution is also supported by the study of [14] and [15].

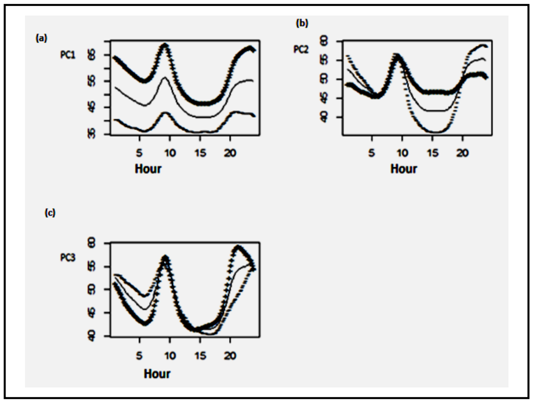
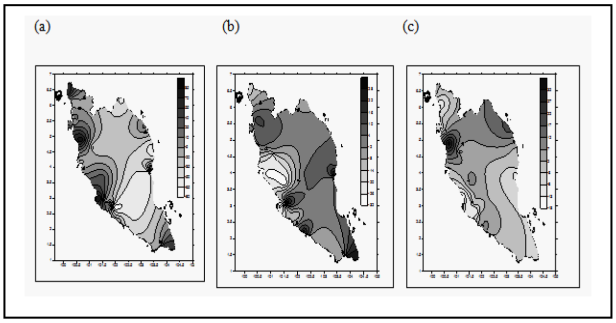


Figure (4) the average PM10 diurnal curves (middle line) and the effect of adding (+) and subtracting (-) a suitable multiple of each PC curve.

Changes in the score of the PC2 produce a time shift before and after daytime hours and tend to captures the day-night contrast in the concentration level while changes in the score of PC3 pick up the time shift (turning) effect that occurred before and after morning peak hours and before and after evening peak hours. A larger decrease in the PM10 concentration in the range between 1.00 to 7.00 a.m than between 11.00 am to 4.00 pm was also observed.

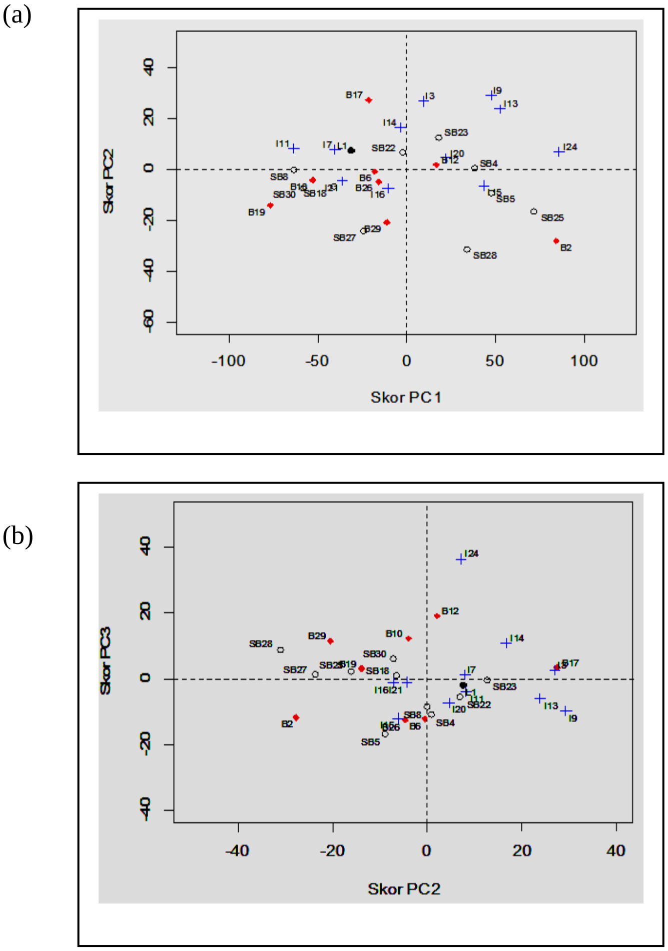
Figure (5 a,b,c) displays the map of each of the three principal component scores. Three different spatial patterns of the scores were revealed. Figure (5a) shows that the positive scores of the high component variation described by PC1 delineated mostly along the coastal region of the western part of Peninsular Malaysia. Two areas were identified to have been concentrated with PC1 variation: one is in the northern coastal areas which include Prai and Sungai Petani stations. Another area is the Klang Valley region which is located in the western coastal area of Peninsular Malaysia. Two spots of positive scores covering the area surrounding the Kuala Terengganu and Balok stations in the east were also observed. Noticeably, those areas with high positive scores are busy urban areas and industrial activities. In particular, the Klang Valley region is known as the Malaysia's center of economy. Klang is also known as the busiest urban site in the Klang Valley region which is located near the main road, industrial area, and a port [15]. On the other hand, the area at the latitude between 6.0^o to 6.2^o in the state of Kedah on the northern part of Peninsular Malaysia and another area at the central part of the Peninsular Malaysia which is located in Pahang with latitude between 3.5^o to 4.0^o have the smallest negative scores. These areas with negative scores are identified to be the rural area, surrounded by agricultural activities such as padi field and rain forest.



Fig(5) The spatial pattern of (a) PC1 (b) PC2 and (c) PC3 scores over Peninsular Malaysia region

Figure (5b) shows that PC2 identified high variation mostly at the central and northern part of Peninsular Malaysia with the highest score concentrated at the western part of Perak and Kedah. The area to the east of the southern Johor state experienced high positive PC2 variation. Another revealing positive score also appeared near Shah Alam and expanding towards Kuala Lumpur area in the middle of Selangor and a location at the further southwestern area located near the Bukit Rambai station in Malacca. In the eastern part, positive score was also visible at the nearby Balok Baru station in Pahang. Those affected areas are categorized as areas with industrial background. Figure (6 a,b) provides evidence of the influence of industrial activities to the high variations of diurnal PM10 as described by the PC2 pattern. Figure (5c) also shows that the third principal component (PC3) explained the highest variation mostly over the northern part of the Malaysian Peninsular. The highest of the positive score centered in the northern coastal region within the area of Perak.

The plot of principal component scores for PC1, PC2, and PC3 given by figure (6 a,b) illustrates how the curves cluster or distribute themselves according to their background category. It can be seen that the majority of stations with industrial background have high positive PC2 scores. Station I24 (Taiping) was identified as the industrial station with the highest PC1, PC2, and PC3 scores, while an urban station B2 (Klang) has the highest PC1 score. The results based on the plot of the components' scores shown by figure (6 a,b) provide a strong basis to relate the variation contributed by PC2 with the influence of photochemical activity. For further explanation, let's relate the behavior into the product of the photochemical activity where a portion of secondary PM10 is produced when O_3 is high [16]. In addition, [17] has cited that the secondary PM10 products are sulfates, nitrates, and organic compounds. Those are the components emitted by industrial activities. The larger variation contributed by PC2 in the industrial areas indicates that Peninsular Malaysia may have been experiencing air pollution due the formation of secondary PM10. The findings, thus, suggest that other than the vehicular emission that has been identified as the major local source, sulfate and nitrate PM10 could be the second major local sources.



Fig(6) Plots of scores of PC1 versus PC2 (a) and PC2 versus PC3 (b) with different symbols and colors indicating different background of stations: cross (blue)-Industrial, full circle (black)-Background, diamond (red)-Urban, and empty circle (black)- Sub-urban.

4. CONCLUSIONS

In this paper, problems to investigate several distinct and important patterns of diurnal variation of PM10 and its spatial distribution across Peninsular Malaysia region during the summer monsoon are considered. Statistical methods to analyze the variation in the curves or functional data namely FDA and FPCA were used. The application of FPCA gives several advantages: it provides the ability to visualize, evaluate, and describe continuous variation of PM10 over a day period. The diurnal PM10 variations provide meaningful insight on the sources of the diurnal variation.

The FPCA eigen-weight function has been found to be a useful visualization tool in understanding the variation, while the eigenvalues for each of the component tell how much (percentage) of emission has been contributed by the dominant sources. The results have indicated that motor vehicles emission is the main contributing source to the diurnal variation in Peninsular Malaysia, which dominantly exists at the western part of Peninsular Malaysia, followed by industrial sources (dominantly exist at industrial locations) other than meteorological condition and mixing factors which dominantly affect at the northern region.

In conclusion, the results of this study have provided the evidence that FDA and FPCA can be used as visualization tools to understand the temporal variation of pollutant concentration level as well as to provide insights on possible sources of PM10 emission. Furthermore, the FPCA eigenvalues can also be suggested as alternative approaches for source apportionment study.

Acknowledgment

The authors would like to thank the Department of Environment Malaysia and Prof. Dr Mohd Talib Latif from Universiti Kebangsaan Malaysia (UKM) for providing the information and data.

5. REFERANCE

[1] Tangang, F.T., Latif, M.T., Juneng, L., “The roles of climate variability and climate change on smoke haze occurrences in Southeast Asia region” *LSE Ideas, SPR 004*, 36-49(2010).

[2] Mahiyuddin, W.R.W., Sahani, M., Aripin,R., Latif, M.T., Thatch, T.Q., Wong, C.M., “Short-term effects of daily air pollution on mortality” *Atmospheric Environment*, **65**: 69-79 (2013).

[3] Costabile, F., Bertoni,G., Santis, F.D., Bellagotti, R., Ciuchini, C., Vichi, F., Allegrini, I., “Spatial distribution of urban air pollution in Lanzhou China” *Environmental Pollution and Toxicology Journal*, **2**: 8-15 (2010).

[4] Staniswalis, J.G., Parks, N.J., Bader, J.O., Maldonado, Y.M., “Temporal analysis of airborne particulate matter reveals a dose -rate effect on mortality in El Paso: Indications of differential toxicity for different particle mixtures” *Air & Waste Management Association*, **55**: 893-903 (2005).

[5] Zheng, J., Wenwei, C., Zhuoyun, Z., Liangfu, C., Liuju, Z., “Analysis spatial and temporal variability of PM10 concentrations using MODIS Aerosol Optical Thickness in the Pearl River Delta region, China” *Aerosol and Air Quality Research Journal*, **13**: 862-876 (2013).

[6] Li, R., Wiendinmyer, C., Baker, K.R., Hannigan, M.P., “Characterization of course particulate matter in the western United States: a comparison between observed and modeling” *Atmospheric Chemistry Physics*, **13**: 1311-1327 (2013).

[7] Bohm, M., McCune, B., Vandetta, T., “Diurnal curves of tropospheric ozone in the western United States” *Atmospheric Environment*, **25**A: 1577-1590 (1991).

[8] Juneng, L., Talib, M.T., Tangang, F.T., Mansor, H., “Spatial assessment of air quality patterns in Malaysia using multivariate analysis” *Atmospheric Environment*, **43**: 4584-4594 (2009).

[9] Ramsay, J.O., Silverman, B.W., “Functional data analysis” (second ed.), *New York, Springer* (2006)

[10] Huang, J.Z., Sheng, H., “Functional coefficient regression models for non-linear time series: a polynomial spline approach” *Scandinavian Journal of Statistics*, **31**: 515-534 (2004).

[11] Ramsay, J.O., Hooker, G., Graves, S., “Functional data analysis with R and Mathlab” *New York, Springer* (2009).

[12] Acuna, E., Rodriguez, C., “The treatment of missing values and its effect in the classifier accuracy; Classification, Clustering and Data Mining Applications”, *Springer-Verlag Berlin-Heidelberg*: 639-648 (2004).

[13] Chang, S.C., Lee, C.T., “Secondary aerosol formation through photochemical reactions estimated by using air quality monitoring data in Taipei City from 1994-2003” *Atmospheric Environment*, **41**(19): 4002-4017 (2007).

[14] Morawska, L., Vishvakarman, D., Swanson, C.E., “Diurnal variation of PM10 concentrations and its spatial distribution in the South East Queensland airshed” *Clean Air and Environmental Quality*, **41**(4): 19-25 (2007).

[15] Dominick, D., Juahir, H., Latif, M.T., Zain, S.M., Aris, A.Z., “Spatial assessment of air quality patterns in Malaysia using multivariate analysis” *Atmospheric Environment*, **60**: 172-181 (2012).

[16] Chang, S.C., Lee, C.T., “Evaluation of the temporal variations of air quality in Taipei City, Taiwan from 1994 to 2003” *Journal of Environmental Management*, **86**: 627-635 (2008).

[17] Shukla, S.P., Sharma, M., “Source apportionment of atmospheric PM10 in Kanpur, India” *Environmental Engineering Science*, **25**(6): 849-862 (2008).

[*] For correspondence; Tel. + (60) 0355435323, E-mail:shahida@tmsk.uitm.edu.my