Pertanika J. Sci. & Technol. 26 (1): 299 - 308 (2018)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Investigating the Randomness and Duration of PM10 Pollution Using Functional Data Analysis

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

[1]Center for Statistical and Decision Science Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi Mara, 40450 UiTM, Shah Alam, Selangor, Malaysia
[2]School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM, Bangi, Selangor, Malaysia

ABSTRACT

Information on situation of air pollution is critically needed as input in four disciplines of research including risk management, risk evaluation, environmental epidemiology, as well as for status and trend analysis. Two normal practices were identified to evaluate daily air pollution situation; first, pollution magnitude has been treated as the common indicator, and second, the analysis was often conducted based on hourly average data. However, the information on the magnitude level alone to represent the pollution condition based on a rigid point data such as the average was seen as insufficient. Thus, to fill the gap, this study was conducted based on continuously measured data in the form of curves, which is also known as functional data, whereby pollution duration is emphasised. A statistical method based on curve ranking was used in the investigation. The application of the method at Klang, Petaling Jaya and Shah Alam air quality monitoring stations located in the Klang Valley, Malaysia, has shown that pollution duration decreases as the magnitude increases. Shah Alam has the longest pollution duration at low and medium magnitude levels. Meanwhile, all the three stations experienced quite a similar length of average pollution duration for the high magnitude level, that is, about 2.5 days. It was also shown that the occurrence of PM10 pollution at the area is significantly not random.

Keywords: Air pollution, functional data analysis, PM10, curve ranking, Malaysia

ARTICLE INFO

Article history:
Received: 03 March 2017
Accepted: 28 September 2017

E-mail addresses:
shahida@tmsk.uitm.edu.my (Shaadan, N.),
sayang@tmsk.uitm.edu.my (Deni, S. M.),
azizj@ukm.edu.my (Jemain, A. A.)
*Corresponding Author

Current Affiliation:
Shaadan, N. and Deni, S. M.
Advanced Analytics Engineering Center,
Faculty of Computer and Mathematical
Sciences, Universiti Teknologi Mara,
40450 UiTM, Shah Alam, Selangor, Malaysia

ISSN: 0128-7680

INTRODUCTION

Having exposed to air pollution has been shown to affect the life quality of human beings, reduce health and decrease life expectancy to a certain extent (Bae, Pan, Kim, Park, Kim, & Kim, 2010; Policheti, Cocco, Spinali, Trimarco, & Nunziata, 2009; Mahiyuddin, Sahani, Aripin, Latif, Thach, & Wong, 2013). Thus, understanding the situation of air pollution including the intensity or the magnitude, as well as their characteristics of occurrences such as their duration and arrival is necessary for preparing and choosing suitable adaptation and mitigation strategies. By any means, the information regarding the situation must be made available. Furthermore, the information is often of a great importance to be used as an input in four disciplines, namely, risk management, risk evaluation, environmental epidemiology, as well as for status and trend analysis (IPCS, 2008).

PM10, a particulate matter with a size of less than 10 micrometer, has been known as one of the important air pollutants other than Ozone (O_3), Nitrogen Oxide (NO_x), Sulphur Dioxide (SO_2) and Carbon Monoxide (CO). In Malaysia, PM10 pollution has become an important environmental problem of concern that needs higher attention as compared to other kinds of environmental pollution in the country; in fact, it has been identified as one of the dominant pollutants in the country other than Ozone (DOE, 2006; Field, Werf & Shen, 2009). The presence of high levels of PM10 in the atmosphere has been reported to significantly associate with haze, which has become a repeated typical problem in the country since 1980s. There have been several haze episodes occurring in Malaysia in the years of 1983, 1987, 1991, 1994, 1997, 2002, 2004, 2005 and 2006 (Afroz, Hassan, & Ibrahim, 2003; DOE, 2006; Field et al., 2009). Although those in 1997 and 2005 were the two extreme years of haze reported in the history of Malaysia, the incidence in 1997 was the most serious as it had caused a huge socioeconomic impact in various sectors including health, production, tourism, transportation and fisheries (Afroz et al., 2003; Abas, Oros, & Simoneit, 2004). For instance, about 83.2% of the population were exposed to health risk during the episode and there was a loss of RM802 million between August and October 1997 due to haze damages (Othman & Shawahid, 1999). The Klang valley region has been reported to be the area most frequently affected by PM10 pollution almost every year, which is regularly caused by transboundary PM10 from Sumatera Island and PM10 emitted from mobile sources, particularly transport, and both are known as the major sources of PM10 in Malaysia (Azmi, Latif, Ismail, Juneng, & Jemain, 2010).

In normal practices, daily air quality data are often recorded discretely by an hourly basis. In the majority of pollution investigation analysis conducted, average data have been popularly used in order to represent the pollutant concentration level for a day period. In a particular perspective, however, the nature of pollutant behaves continuously with time (Gao, 2007; Pudasainee, Sapkota, Shreshta, Kaga, Kondo, & Inoue, 2006). Therefore, it is more ideal to have the continuous evolution of PM10 concentration within the day process taken into consideration in the analysis part. Furthermore, using average data in the analysis is not sufficient enough due the fact of losing information or data reduction as a consequence of averaging or summarising process. To fill the gaps, this paper discusses and proposes the employment of functional data in the study analysis in order to study a situation of PM10 pollution at an area. Specifically, this study aims to investigate the randomness and duration of PM10 pollution. Therefore, to

achieve the study objective, functional data analysis (FDA), which is defined as the statistical methods to analyse functional data or curve data, is used.

METHODOLOGY

Several steps are needed to enable the analysis to be conducted using functional data. The analysis starts with data conversion process, whereby original observed data are transformed into a day curve. Next, the process is followed by the determination of the situation between polluted (abnormal) and normal day based on the concept of curve ranking. Further investigation on the randomness of the pollution occurrences is then conducted.

Data Conversion Process

Converting discrete original recorded data into functional data or curves is the first step that needs to be done. Various approaches can be used either by parametric or non-parametric. In this study, a famous and the most flexible method namely the basis expansion method is chosen (Ramsay & Silverman, 2006). Using the method, the discrete points of PM10 level (y_j) recorded at time point t_j for j=1, ..., 24 of day i are converted into a continuous function $x_i(t)$ using the following formula:

$$x_i(t) = \sum_{k=1}^{K} c_k \phi_k(t), \quad t \in [1,24] \tag{1}$$

The term $\phi(t)$ is the chosen basis system consisting of K number of appropriate basis functions, while c is the corresponding basis coefficient. In this study, K is determined using Bayesian Information Criteria (BIC) and the function's bases are the cubic B-spline. The coefficient c is determined using the least square methods by minimising the error terms based on the following equations:

$$SSE = \sum_{j=1}^{24} (y_j - x(t_j))^2 = \sum_{j=1}^{24} \left(y_j - \sum_{k=1}^{K} c_k \phi_k(t_j) \right)^2 \tag{2}$$

Determining the Status of PM10 Condition Based on Curves Ranking

Consider a set of functional data consisting of n daily PM10 curves, represented by functions $x_i(t)$ for i=1, ..., n. The next step is to determine which day is considered as polluted (abnormal) and which one is considered as normal. PM10 pollution refers to the concentration level in the atmosphere that exceeds a specific allowable level or a threshold level. For this study, a polluted (abnormal) day is defined as any day curve that is higher than a given threshold curve.

The three threshold curves that represent the three different degrees of pollution magnitude are the 50th, 75th and the 90th percentile curves. The 50th percentile curve is used to represent pollution with low magnitude; the 75th percentile is aimed for pollution with medium magnitude level, while the 90th percentile represents pollution with high magnitude. The identification of the percentile curves and determination of the pollution condition of any day is conducted based on curve ranking. In the ranking process, the curves position can be assessed using an

index named SNJF. The index is positive in the value and defined as the ratio of distances and constructed using the following formula:

$$\text{SNJF}(x_i) = \frac{A}{B} \tag{3}$$

Where:

$$A = \frac{d(x_i, x_{\min})}{d(x_i, x_{\min}) + d(x_i, x_{\max})},$$

$$B = \frac{d(x_{\text{median}}, x_{\min})}{d(x_{\text{median}}, x_{\min}) + d(x_{\text{median}}, x_{\max})}$$

$d(x_i, x_{\min})$ is the distance between a curve and the minimum curve
$d(x_i, x_{\max})$ is the distance between a curve and the maximum curve
$d(x_{\text{median}}, x_{\min})$ is the distance between median and minimum curve
$d(x_{\text{median}}, x_{\max})$ is the distance between median and the maximum curve

The distance between the two curves, namely, curve x_1 and x_2, which is obtained using the L_2 norm as in the following equation.

$$\begin{aligned} d(x_1, x_2) &= \|x_1 - x_2\|_2 \\ &= \left(\int_1^{24} |x_1(t) - x_2(t)|^2 \, dt \right)^{1/2} \end{aligned} \tag{4}$$

The minimum curve is defined as $\check{x}_j = \underset{i}{\text{minimum}}\,(x_i(t_j))$ and the maximum curve is $\hat{x}_j = \underset{i}{\text{maximum}}(x_i(t_j))$, while the median curve is obtained based on Fraiman and Muniz (2001). Any day with SNJF exceeding the SNJF of a given threshold curve indicates that the day is considered as a polluted or an abnormal day.

Obtaining the Duration of PM10 Pollution

A coding system is used to obtain a data set that contains the polluted (abnormal) and normal days such that; 0 for normal and 1 for polluted (abnormal) day. In order to investigate the duration of pollution, the data (0,1) are used. Examples of the data are given in Table 1. Based on Table 1, within the 15-day period, there are 2 runs of polluted days. The first run occurred in January 2001 within one-day period and the second run occurred within 2-day period on the 12th and 13th January 2001. In the context of this study, the pollution duration is defined as the number of consecutive days in the state of being polluted or abnormal (i.e., the length of run).

Table 1
Examples of data on the occurrences of polluted and normal day given the SNJF for a threshold curve (the 50th percentile) as 1.037 for a 15-day period

Year	Month	Day	SNJF	Occurrences	
2001	1	1	1.7717	1	} 1st run
2001	1	2	0.5736	0	
2001	1	3	0.7699	0	
2001	1	4	0.5590	0	
2001	1	5	0.7392	0	
2001	1	6	0.9931	0	
2001	1	7	0.5751	0	
2001	1	8	0.5699	0	
2001	1	9	0.6175	0	
2001	1	10	0.5450	0	
2001	1	11	0.9262	0	
2001	1	12	1.3545	1	} 2nd run
2001	1	13	1.1356	1	
2001	1	14	0.9694	0	
2001	1	15	0.8775	0	

Checking the Randomness of Pollution Occurrences

A non-parametric run test known as the Bradley run test was conducted to investigate the behaviour of PM10 pollution occurrences (Bradley, 1968). The aim is to determine whether the occurrences of PM10 pollution were random. Since the analysis involved a large size of data, and thus based on the occurrences and using normal approximation, the run test was conducted to examine the hypothesis whether the distribution of the pollution occurrences was random (null hypothesis) or otherwise (alternative hypothesis) using statistic z such that:

$$z_{calc} = \frac{\left| NR - \overline{NR} \right| - 0.5}{S_R} \tag{5}$$

Where, R is the number of runs, $\overline{NR}$ is the expected number of runs and S_R is the standard deviation of the number of runs, which are given by the following equation:

$$\overline{NR} = \frac{2n_a n_b}{n_a + n_b} + 1 \tag{6}$$

$$S_R = \sqrt{\frac{2n_a n_b (2n_a n_b - n_a - n_b)}{(n_a + n_b)^2 (n_a + n_b - 1)}} \tag{7}$$

where, n_b is the number of normal days and n_a is the number of polluted (abnormal) days. At 5% significance level ($\alpha = 0.05$), for large sample size, the null hypothesis is rejected if the z_{calc} statistic in equation (5) lies outside the interval $[Z_{\alpha/2}, Z_{1-\alpha/2}]$, whereby z_{calc} follows the standard normal distribution with mean 0 and variance 1 (Mendenhall, 1982). Alternatively,

using p-value, which is defined as the probability of $Z < -z_{calc}$ or $Z > z_{calc}$, the null hypothesis is also rejected when the p-value is less than the significance level, α. The run test can only be conducted when there is at least one run for normal and polluted days over the considered period of time such that $n_a, n_b \neq 0$.

APPLICATION OF THE METHOD

The methods explained in the previous section were applied to investigate the PM10 pollution situation at several locations in the Klang valley of Peninsular Malaysia. The analysis was conducted using daily curves data for a period of 10 years (i.e., from 2001 until 2010). The original observed data (daily by hourly discrete data) were obtained from the Department of Environment, Malaysia, involving three air quality monitoring stations, namely, Klang (S1), Petaling Jaya (S2), and Shah Alam (S3). Based on equation (4), the observed data were converted into daily curves using the pre-determined number of basis function K which equals 15, 19 and 17 at each station (S1, S2 and S3) respectively by means of BIC values.

RESULTS OF THE ANALYSIS

For illustration purposes, we can considered a set of five-day PM10 data that were recorded at an hourly basis, as shown in Table 2. The computed SNJF index for each curve is shown in Table 3, while the physical form of the five-day curves is represented in Figure 1. Using the SNJF index, the position of the curves in term of their magnitude can be determined. For this sample of the curves, based on SNJF index, day 4 was found to be the lowest curve, while day 5 is the highest curve.

Table 2
Examples of the 5-day recorded PM10 data

Day	Hour									
	1	2	3	4	5	…	21	22	23	24
1	38	29	23	44	49	…	31	43	37	48
2	51	28	26	29	19	…	29	30	35	44
3	50	51	27	46	39	…	57	44	45	43
4	29	24	16	16	17	…	48	39	38	34
5	19	28	26	30	26	…	59	69	66	80
$\bar{y}$	37.4	32.0	23.6	33.0	30.0	…	44.8	45	44.2	49.8

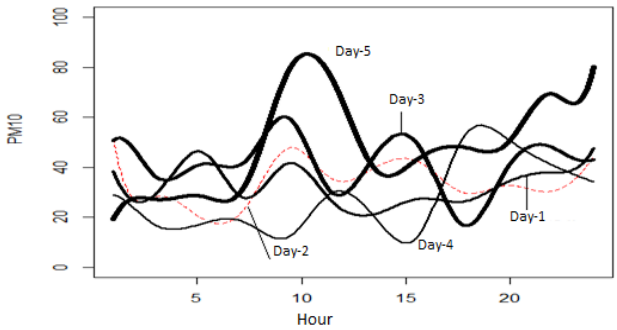

Figure 1. Data set for the five days curve

Table 3
Curve (day) ranking based on SNJF (1-lowest, 5- highest)

Curve (day)	1	2	3	4	5
SNJF Index	1.00	1.11	1.54	0.74	2.09
Position	2	3	4	1	5

Results of the analysis conducted for the data from the three air quality monitoring stations are recorded in Table 4. The results indicate that PM10 pollution duration decreased as the magnitude increased. Among the three stations, with respect to their own local environment, Shah Alam experienced the longest pollution duration at low and medium magnitude levels, while for the high magnitude level, the three stations experienced quite a similar length of pollution duration that was about 2.5 days.

Table 4
The mean duration of PM10 pollution (in days) at three different magnitude levels

Station Code	Name	Threshold Curve		
		P50	P75	P90
		Pollution duration (days)		
S1	Klang	3.92	2.94	2.54
S2	Petaling Jaya	5.20	3.36	2.56
S3	Shah Alam	4.76	3.71	2.47

At the 5% significance level, after the randomness (run) test was conducted, the analysis revealed that the p-value for the test at each level of magnitude; the low (50^{th} percentile), medium (75^{th} percentile) and high (90^{th} percentile) were near to zero (0.00). Therefore, we can conclude that the occurrence of PM10 pollution at the three stations is not random, thus the results suggest that PM10 pollution in the area is not generated by a random process.

CONCLUSION

This study has highlighted a statistical methodology on how to investigate the situation of air pollution using functional data as the input in the study analysis. In contrast to many previous research, functional data have been chosen to incorporate full evolution of the pollutant concentration level within a day period of time. Instead of preventing insufficient information contained in the day data, the application was found to be more appropriate due to the continuous nature of the pollutant process with respect to time.

The identification of polluted (abnormal) and normal day was made based on the simplest approach, that is, via curve ranking procedure. The magnitude level of each PM10 curve (curve) is represented by an index called the SNJF. The three situations of PM10 pollution were considered with respect to three categories of magnitude levels; low, medium and high. SNJF was used to determine the threshold curve (i.e. the 50^{th}, 75^{th} and 90^{th} percentile), as well as indicate whether a day curve is considered as polluted or normal. A day with the SNJF index higher than SNJF index of a threshold curve is detected as being in a polluted (abnormal) situation. Further investigations to study the duration and randomness behaviour of the PM10 pollution occurrences have also been conducted.

The application of the discussed method for the data set at the three air quality monitoring stations located in the Klang Valley revealed that with respect to their respective environment, Shah Alam experienced the longest pollution duration at low and medium magnitude levels, whereas for the high magnitude level, the three stations experienced quite a similar length of pollution duration of 2.5 days. The occurrence of PM10 pollution in the area is significantly not random.

In conclusion, as the application of functional data is becoming popular in many areas including environmental research, this paper would help researchers to conduct a study analysis using curves or functional data.

ACKNOWLEDGEMENT

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. This work is supported by the UKM's Research University Grant [UKM-AP-2011_19].

REFERENCES

Abas, M. R. B., Oros, D. R., & Simoneit, B. R. T. (2004). Biomass burning as the main source of organic aerosol particulate matter in Malaysia during haze episodes. *Chemosphere, 55*(8), 1089-1095.

Afroz, R., Hassan, M. N., & Ibrahim, N. A. (2003). Review of air pollution and health impacts in Malaysia. *Environmental Research, 92*(2), 71-77.

Azmi, S. Z., Latif, M. T., Ismail, A. S., Juneng, L., & Jemain, A. A. (2010). Trend and status of air quality at three different monitoring stations in Klang Valley, Malaysia. *Air Quality Atmospheric Health, 3*(1), 53-64.

Bae, S., Pan, X., Kim, S., Park, K., Kim, Y., & Kim, H. (2010). Exposures to particulate matter and polycyclic aromatic hydrocarbons and oxidative stress in schoolchildren. *Environmental Health Perspective, 118*(4), 579-583.

Bradley, J. V. (1968). *Distribution-free Statistical Tests*. Englewood Cliffs, New Jersey: Prentice-Hall.

DOE. (2006). Department of Environment. *Malaysia Environmental Quality Report, 2005*. Ministry of Environment and Natural Resources, Putrajaya.

Field, R. D., Werf, G. R. V., & Shen, S. S. P. (2009). Human implication of drought-induced biomass burning in Indonesia since 1960s. *Nature Geoscience, 2*(3), 185-188.

Fraiman, R., & Muniz, G. (2001). Trimmed means for functional data. *Test, 10*(2), 419-440.

Gao, H. O. (2007). Day of week effects on diurnal ozone, NOx cycles and transportation emissions in Southern California. *Transportation Research Part D, 12*(4), 292-305.

IPCS. (2008). *Human Exposure Assessment, Environmental Health Criteria 214 WHO, Geneva, 2000.* International Program on Chemical Safety (IPCS) and INCHEM (Chemical Safety Information from Intergovernmental Organisation).

Mahiyuddin, W. R. W., Sahani, M., Aripin, R., Latif, M. T., Thach, T. Q., & Wong, C. M. (2013). Short-term effects of daily air pollution on mortality. *Atmospheric Environment, 65*, 69-79.

Mendenhall, W. (1982). *Statistics for Management and Economics*. Boston: Duxbury Press.

Othman, J., & Shahwahid, O. H. M. (1999). Cost of Trans-boundary haze externalities. *Jurnal Ekonomi Malaysia, 33*, 3-19.

Policheti, G., Cocco, S., Spinali, A., Trimarco, V., & Nunziata, A. (2009). Effects of particulate matter (PM10, PM2.5 and PM1) on cardiovascular system. *Toxicology, 261*(1), 1-8.

Pudasainee, D., Sapkota, B., Shreshta, M. L., Kaga, A., Kondo, A., & Inoue, Y. (2006). Ground level ozone concentrations and its association with NOx and meteorological parameters in Kathmandu valley Nepal. *Atmospheric Environment, 40*(40), 8081-8087.

Ramsay, J. O., & Silverman, B. W. (2006). *Functional data analysis*. New York: Springer.