Atmospheric Pollution Research 6 (2015) 365-375

Atmospheric Pollution Research

www.atmospolres.com

Anomaly detection and assessment of PM_{10} functional data at several locations in the Klang Valley, Malaysia

Norshahida Shaadan [1], Abdul Aziz Jemain [2], Mohd Talib Latif [3], Sayang Mohd Deni [1]

[1] *Center for Statistical and Decision Science Studies, Faculty of Computer & Mathematical Sciences, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia*
[2] *DELTA, School of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia*
[3] *School of Environmental and Natural Resource Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia*

ABSTRACT

In environmental data sets, the occurrence of a high concentration of an unusual pollutant, more formally known as an anomaly, may indicate air quality problems. Thus, a critical understanding of the behavior of anomalies is increasingly becoming very important for air pollution investigations. This study was conducted to detect anomalies in daily PM_{10} functional data, to investigate the patterns of behavior as well as to identify possible factors that determine PM_{10} anomalies at three selected air quality monitoring stations (Klang, Kuala Selangor and Petaling Jaya) in the Klang Valley, Malaysia. The statistical method employed to detect these anomalies consisted of a combination of the robust projection pursuit and the robust Mahalanobis distance methods using air quality data recorded from 2005 to 2010. Analysis of obtained anomalous PM_{10} profiles showed that data recorded during El Nino years (2005, 2006 and 2009) contained the highest frequency of anomalies. More frequent anomalies appeared during the southwest (SW) monsoon which occurs in the months of July and August as well as during the northeast (NE) monsoon in February. A lesser number of anomalies were also observed during weekends compared to weekdays. The weekend and monsoonal effect phenomena were shown to be significantly existent at all stations while wind speed was positively associated with extreme PM_{10} anomalies at the Klang and Petaling Jaya stations. In conclusion, anomalies detection was found useful for air pollution investigation in this study. The findings of this study imply that the location and background of a station, as well as wind speed, seasonal (monsoon) and weekdays–weekend variations play important role in influencing PM_{10} anomalies.

Keywords: *PM_{10}, functional data, anomaly detection, air quality monitoring*

Corresponding Author:
Norshahida Shaadan
☎ : *+603-55435323*
🖷 : *+603-55435501*
✉ : *shahida@tmsk.uitm.edu.my*

Article History:
Received: 24 April 2014
Revised: 30 September 2014
Accepted: 26 October 2014

doi: 10.5094/APR.2015.040

1. Introduction

Particulate matter is the main pollutant in ambient air, particularly in urban areas (Wrobel et al., 2000; Acero et al., 2012). In normal conditions, particulate matter usually originates from natural and anthropogenic sources such as sea spray, road dust, soil, motor vehicle usage, industrial activities, domestic activities, and biomass burning (Schauer et al., 1996; Keuken et al., 2013). Particulate matter could also be generated through the accumulation of small–sized particles or via secondary interactions between gases and ions (Shon et al., 2012). Measurement of particulate matter is usually based on its aerodynamic diameter. Generally, particulate matter with a diameter size below 10 µm is referred to as PM_{10} while particulate matter with a diameter size of less than 2.5 µm is referred to as $PM_{2.5}$. PM_{10} has long been recognized as the main parameter for determining particulate matter in Malaysia. Additionally, it has been consistently used as the principal specification for the calculation of the Malaysian Air Pollution Index (API) (Afroz et al., 2003; Awang et al., 2000). Elevated concentrations of PM_{10} have also been implicated in respiratory mortality, particularly in busy areas such as the Klang Valley (Mahiyuddin et al., 2013).

PM_{10} has been found to have a significant connection with haze episodes from biomass burning, a challenge which has become both typical and reoccurring in Southeast Asia since the 1980s (Abas et al., 2004; Field et al., 2009). Local anthropogenic activities involving, for example, motor vehicle usage and industrial activity are the major sources of PM_{10} pollution during non–haze periods, particularly in the Klang Valley region. Meteorological factors also contribute to the amount of particulate matter in the region (Juneng et al., 2011; Dominick et al., 2012). Higher concentrations of particulate matter have been recorded during the dry season, notably during the El Nino/Southern Oscillation (ENSO) events (Matsueda et al., 1999; Mahmud, 2009). The wind direction from the southwest which comes from Sumatra between July and September brings a high amount of particulate matter to the Klang Valley. The concentration of particulate matter is also influenced by local wind direction e.g. sea and land breezes and the movement of wind within the valley. The amount of rain during the rainy season (northeast monsoon) plays an important role in reducing the quantity of particulate matter in the ambient air. Other factors, such as activities on weekdays and weekends also influence the amount of particulate matter in the Klang Valley (Azmi et al., 2010).

Understanding the behavior of anomaly occurrences is becoming more important in air pollution investigation. The term "anomalies" refers to a small portion of the data set that is unusual or dissimilar to the rest of the data. Anomalies may consist of noisy data due to random errors; alternatively, they may be irregular items of data resulting from unusual or unexpected events which may indicate abnormal behavior (Torres et al., 2011). Intensive study of anomalies helps in identifying potential sources of the

occurrences. "Anomaly or outlier detection" refers to statistical techniques used to detect abnormal data or outliers (Muniz et al., 2012). Basically, in environmental research and other fields, outlier detection is among the most important tasks in data analysis (Filzmoser, 2005; Garces and Sbarbaro, 2011). The practical application of this technique ranges from its usage in detecting financial fraud (Sharma and Panigrahi, 2012), network intrusion (Garcia–Teodoro et al., 2009; Davis and Clark, 2011), system health monitoring (Hauskrecht et al., 2013), criminal incidence analysis (Lin and Brown, 2006) and many other aspects. On the other hand, anomaly detection is less applied to environmental data particularly during the monitoring of air pollution whereas the application is important because it could be used to evaluate polluted air in an area (Torres et al., 2011). Furthermore, as supported by Hawkins et al. (2002), it is reasonable to assume values for possibly polluted air behaving as outliers or anomalies.

Several studies that focuses on air pollutant variations in the Klang Valley region have been conducted (Azmi et al. 2010; Juneng et al., 2011; Ahamad et al., 2014). Noticeably in those studies, the employment of functional data was less applied. A previous study by Shaadan et al. (2012) highlighted the advantage of functional data approach in assessing and comparing the PM_{10} behavior during and between the two extreme haze years (1997 and 2005) that have been reported in Malaysia. Nevertheless, for this study, besides aiming to provide a complementary technique for the evaluation of air pollution problem, functional data were further extrapolated to increase understanding for PM_{10} anomalies and the associated influential factors.

In specific, the objective of this study is to detect and analyze the profiles of anomalies in daily PM_{10} functional data as well as to investigate possible factors associated with the existence of anomalies at three air quality monitoring stations (Klang, Kuala Selangor and Petaling Jaya) with differing locational backgrounds in the Klang Valley region of the Malaysian Peninsular.

2. Methodology

2.1. Description of data and study location

The Klang Valley region is considered to be the heartland of Malaysia's industrial and commercial sectors with a high–density multi–racial population. The climate of Malaysia is very much influenced by two major types of monsoon seasons; the southwest (SW) and the northeast (NE), as well as another two inter–monsoon periods. During the SW monsoon, which is reported to occur from late May until September, drier weather conditions are normally experienced. Meanwhile, the NE monsoon which takes place between November and March receives a higher precipitation level, particularly during the first few months of the season.

The air pollutant and meteorological variable of concern in this study are PM_{10} and wind speed, respectively. The data were made available by the Air Quality Division, Department of Environment Malaysia (DOE). To ensure for reliability of the measurement process, continuous monitoring and calibration of the equipment was carried out by Alam Sekitar Sdn Bhd (ASMA), a private company mandated by the DOE for this purpose. Daily by hourly PM_{10} in ($\mu g\ m^{-3}$) was recorded using β– ray attenuation mass monitor (BAM–1020) while wind speed data in ($km\ h^{-1}$) was determined using Met One 010C sensor.

Three selected air quality monitoring stations; Klang (S1), Kuala Selangor (S2) and Petaling Jaya (S3) which are located within the Klang Valley region of Peninsular Malaysia were involved in this study. Table 1 describes background information and the data while Figure 1 shows the location of the sampling stations. Klang air monitoring station is located in the city centre and is in close proximity to a busy, traffic–laden industrialized area, surrounded by main roads and a busy port (Port Klang). Kuala Selangor air monitoring station is located in a residential area, on the outskirts of a small town which is both near the coast and to the main road. Meanwhile, Petaling Jaya air monitoring station is the nearest station to Kuala Lumpur city centre and is surrounded by industries, residential and commercial areas as well as a heavily congested road.

Figure 1. *Location of air quality monitoring stations.*

Missing data were treated using the column median value computed from the available data. The method utilized was chosen due to a considerably small percentage of missing values (<5%). This approach is supported by Acuna and Rodriguez (2004). Since the study set out to detect anomalies in the form of functional data (curves), 2 192 (N) daily PM_{10} curves were used for analysis of data obtained at each station. The curve data at time [hour (t)] and at any point (j) is defined as follows:

$$x_i(t_j);\ i = 1, \dots, N;\ j = 1, \dots, n \tag{1}$$

Data processing and analysis were conducted using the free R software (R Development Core Team, 2008) together with "rainbow" (Shang and Hyndman, 2013) and "fda" packages (Ramsay et al., 2013).

2.2. Data analysis

Data analysis in this study involved several stages. The first stage involved data conversion from point values into functional or curve forms. The second stage focused on the detection of anomalies in PM_{10} functional data at the three selected stations. This was subsequently followed by an assessment of the effectiveness of the preferred detection method and profile construction wherein detected anomalies were extracted and summarized. Finally, a statistical test was conducted to ascertain a phenomenon that may be indicated by the anomaly profile and also to investigate the association between anomalies and wind speed.

Data conversion from points to functional data. Hourly recorded data were converted into daily i functional data, $x_i(t)$ using basis function expansion given by the following equation:

$$x_i(t) = \sum_{k=1}^{K} \beta_{i,k}\, \varphi_k(t) \tag{2}$$

Table 1. *Data and the station information*

Station	Background	Longitude (°)	Latitude (°)	Missing Data (%)
Klang (S1)	Urban	N03°00.597'	E101°24.507'	2.17
Kuala Selangor (S2)	Sub-urban	N03°19.592'	E101°15.532'	1.94
Petaling Jaya (S3)	Industry	N03°06.553'	E101°38.322'	2.62

which, consists of a linear combination of K, independent basis function $\varphi_k(t)$ and the basis coefficient β_k. Although various kinds of basis function can be used in the modeling process, determining which basis is the best is dependent on the nature of the data. Fourier–basis, for example, is suitable for periodic data while spline is more appropriate for non–periodic data (Ramsay and Silverman, 2006). The appropriate value of K is determined using Bayesian Information Criteria (BIC) based on the construction of the mean functional data from the data set (Huang and Shen, 2004). The appropriate K is the one that gives the minimum BIC. In this study, data conversion was conducted using the b–spline basis with the number of basis, K, equal to 15 for Klang and 17 for both Kuala Selangor and Petaling Jaya air monitoring stations. Further information on the theory and application of the statistical approaches in functional data can be obtained from Ramsay and Silverman (2002; 2006).

Anomaly detection. The multivariate robust Mahalanobis distance method reported by Hyndman and Shang (2010) was used for the detection of anomalies in this study. Since computation using the robust Mahalanobis distance method adopts a multivariate approach, n equally spaced discretized points on the curve that span across the curves interval are needed to represent a functional data. As required, some trade off were considered in choosing the appropriate value of n. A small n ensures stability in the algorithm computation while a large n better approximates the curve. However, a too large n incurs problem in the computation of multivariate statistics due to the singularity of covariance matrix (Liebl, 2013).

The robust, multivariate Mahalanobis distance approach consists of two sub–procedures. First is the projection pursuit procedure and second is the computation of the measures for curve outlyingness. The aim of the first procedure is to search for specific linear projections of the discretized curves. Using principal component analysis (PCA), the discretized curves were projected into p dimensional space. Considering a matrix X of size N rows by n columns, the search for the projected curves follows the eigen equation:

$$Vu = \lambda u \tag{3}$$

where, V is the sample variance–covariance matrix, that is V=N^{-1} X^T X, the term u is an eigenvector of V and λ is an eigenvalue of V. The solution can be obtained by finding the first p projected weight vector (i.e. eigenvector) that maximizes the variance in the data. The projected scores is given by $s_1 = u_1^T X$, $s_2 = u_2^T X$, ...$s_p = u_p^T X$. Eigenvactors are orthogonal to each other; the first is contributed to by the largest variation in the data set, the second by the second largest variation and so on.

In the second procedure, by considering the first two projections (i.e. principal component) that describes two major proportions of variation in the data set, the squared robust Mahalanobis distance for each curve, $D^2(x_i)$ was computed. This was achieved using the projected scores obtained from the first procedure where the new matrix data set was defined to be X=[s_1, s_2] of bivariate covariates. The term $D^2(x_i)$ is a measure of the outlyingness of a curve. The larger the value, the more outlying a curve is from the centre of the group. The computation for formula $D^2(x_i)$ for each curve is as follows:

$$D^2(x_i) = (x_i - \bar{x})^T \Sigma^{-1} (x_i - \bar{x}) \tag{4}$$

where, x_i is a vector of measured points for curve i, $\bar{x}$ is the location estimator (i.e. mean vector) and matrix Σ is the robust estimate of the covariance matrix of X. Given the assumption that the data were generated from a chi–squared distribution, the cut–off point to differentiate between anomalous and non–anomalous curve was based on the critical value of the $\chi_{1-\alpha,p}^2$, that is, predefined α quantile of the distribution with p degree of freedom (Filzmoser, 2005). The cut–off point was determined based on the choice of α. Lower α values indicated higher cut–off points which resulted into lower percentages of detected anomalies. In this study, anomalies were detected when their distance exceeded the critical value $\chi_{0.99,p}^2$ which is a measure of the outlyingness of a curve with α=0.01. The larger the squared value of the robust Mahalanobis distance, the more outlying the curve was from the centre of the group.

The robust estimate of the covariance matrix was taken as the covariance of the optimal subsample h. The subsample was considered optimum if it had the minimum determinant of the covariance matrix, more formally known as the minimum covariance determinant (MCD) approach (Rousseeuw and Van Driessen, 1999). The value of h was assumed to be the minimum number of curves which must not be outlying. Using MCD, the outlying points were ignored in the computation process. MCD was defined as follows:

$$MCD = (\bar{x}_L^*, s_L^*) \tag{5}$$

where, L is the matrix of the subsample h that has the minimum determinant of the covariance matrix with h=[(N+p+1)/2], $\bar{x}_L^* = \frac{1}{h}\sum_{i\epsilon L}(x_i)$ and $s_L^* = \frac{1}{h}\sum_{i\epsilon L}(x_i - \bar{x}_L^*)(x_i - \bar{x}_L^*)^T$, thus $\Sigma = s_L^*$ is the sample covariance estimate. The discussed method was favored due to the convincing application results for detecting functional outliers as reported by Hyndman and Shang (2010), and also due to the efficiency of the approach in handling large data sets (Rousseeuw and Van Driessen, 1999).

3. Results and Discussion

The main goal of the analysis was to detect anomalies. Anomalies with all hours of the day which lie above the median levels when detected have increasingly become an issue of concern due to their potential impact on human health. These anomalies are known as red anomalies (RA). The second goal was to investigate the effectiveness of the employed anomaly detection method followed by an establishment of the anomaly profiles and finally an examination of the possible existence of a phenomenon that might be indicated by the profile as well as the influence of wind speed on PM_{10} anomalies.

3.1. Descriptive statistics of anomalies

A prior analysis was conducted to determine the appropriate number of n discretized function to be used in the multivariate computation for detecting anomalies. Based on the computation for the percentage of RA (P_{red}), the results in Table 2 show that the choice of n=40 was appropriate for the data in the Klang and Petaling Jaya stations, while n=50 was suitable for the Kuala Selangor station. These values yielded identical results or very

small difference in deviance when n (the size of discretized points) increased. Furthermore, ranging from 20 to 70 in size, there was not much difference in the deviance of the overall mean ($\bar{x}$). The anomaly detection for the data set was then conducted using the determined size of the discretized function.

Based on the analysis conducted during the six year study period, the percentage of frequency of anomaly occurrence was found to be 6.80% for Klang, 9.04% for Kuala Selangor and 5.20% for Petaling Jaya monitoring station. Figure 2a shows that the maximum level for all of the detected anomalies was above the maximum level of the median curve. Thus, the results indicate that none of the anomaly curves was totally below the median curve. On the other hand, the percentage of RA was found to be the highest at the sub–urban site (Kuala Selangor) (3.15%) followed by the industrial site (Petaling Jaya) (2.46%), while the lowest percentage was recorded at the urban site (Klang) (1.14%). The larger percentage of abnormal days of PM_{10} levels at a quieter area such as Kuala Selangor, suggests a stronger influence from non–local sources, particularly transboundary pollution. Both Kuala Selangor and Klang are located downwind and near to the Island of Sumatra, while Petaling Jaya is further downwind towards the central part of the Klang Valley region. Even though Abas et al. (2004) stated that transboundary pollution is the major source of pollution during haze incidences, background sources and meteorological factors are also believed to relate to high anomaly occurrences (Lee et al., 2011). Therefore, with respect to the station's background, the results suggest that apart from the transported haze, emissions from heavy traffic along with industrial activity at Klang and Petaling Jaya exacerbate the conditions. Ultimately, this makes them poorer in terms of air quality.

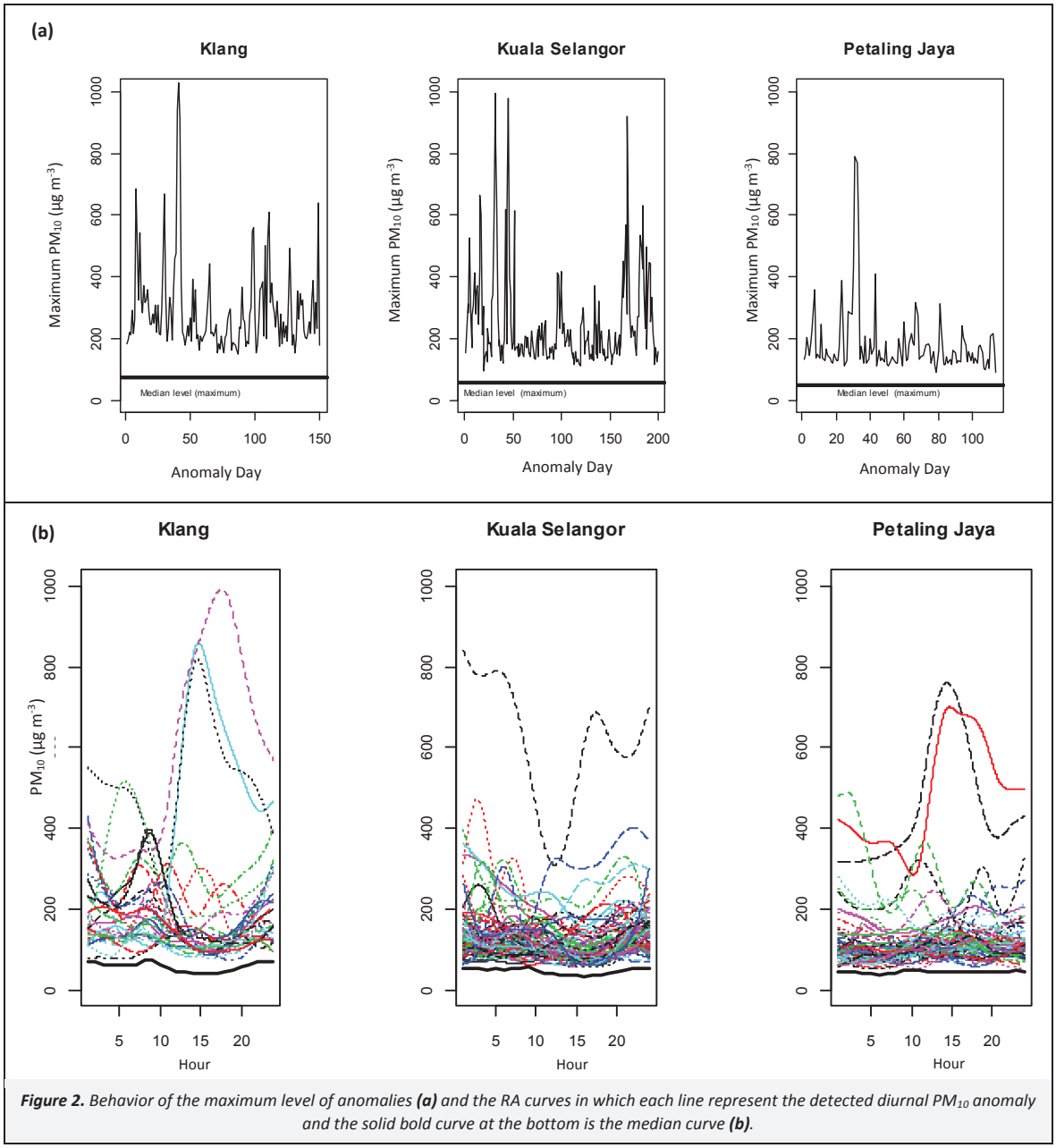

Figure 2. *Behavior of the maximum level of anomalies* ***(a)*** *and the RA curves in which each line represent the detected diurnal PM_{10} anomaly and the solid bold curve at the bottom is the median curve* ***(b)****.*

Table 2. *Different sizes of (n) discretized points on the curve and the estimated statistics mean and results*

	Station								
	Klang			Kuala Selangor			Petaling Jaya		
Size (n)	$\bar{x}$	Deviance	P_{red}	$\bar{x}$	Deviance	P_{red}	$\bar{x}$	Deviance	P_{red}
20	67.323			57.985			51.462		
30	67.029	0.293	1.32	57.808	0.178	3.33	51.421	0.041	2.65
40	66.883	0.146	1.14	57.721	0.087	3.24	51.401	0.020	2.46
50	66.795	0.088	1.14	57.670	0.051	3.15	51.390	0.012	2.51
60	66.736	0.059	1.14	57.636	0.034	3.15	51.381	0.008	2.46
70	66.694	0.042	1.14	57.612	0.024	3.15	51.376	0.006	2.46
Size (n)		n=40			n=50			n=40	
Frequency of anomaly		149 (6.80)			198 (9.04)			114 (5.20)	

Figure 2b shows the functional form of the detected RA. The diurnal levels of RA anomalies fluctuate with unstable direction with the majority of them exhibiting peaks during day time at Klang and Petaling Jaya stations. Meanwhile, the peak at Kuala Selangor occurred during the night time. It is also shown that a few of the severest RA at Klang and Petaling Jaya shared the same diurnal pattern of maximum peak that occurred after midday at around 3:00 to 5:00 pm. Kuala Selangor experienced the most extreme RA that reached a peak at midnight and a minimum at 12:00 noon.

3.2. Examining the effectiveness of the method used in detecting anomalies

The effectiveness of the anomaly detection method employed in this study was also investigated so that further analysis and conclusions could be drawn. Here, the effectiveness of the method is defined as the ability of the method to detect anomalies with date matches with the period of the reported haze incidences that have occurred in the country. Not only is the date, the magnitude of the concentration levels also considered to show that those days are anomaly. According to Tangang et al. (2010), several serious haze episodes have occurred in various parts of the Malaysian region, including those that occurred in 1982–83, 1987, 1991, 2002, 2004, 2005, 2006 and 2009. Due to the limited information on the reported incidences available, only a group of several top anomaly curves that lies at the upper percentile of the distribution were sampled and fed into the analysis. These detected anomalies were believed to be the consequence of abnormal events. As reported in Afroz et al. (2003), high PM_{10} levels often associated with haze incidences in Malaysia. Thus, we inferred that the most significant anomalous behavior occurred as a result of the severe haze incidents that were reported. From the data used in the analysis, some dates and periods of reported incidences were taken from the Malaysian Environmental Quality Report for year 2005, 2006 and 2009 (DOE, 2006; DOE, 2007; DOE, 2010). The information was recorded and summarized in the first three columns in Table 3. The dates of occurrence of the top 20 anomalies detected from the data (2005–2010) at all the three air quality monitoring stations are reported in the rest of the columns (Table 3) including the maximum (point value) of the concentration level.

Remarkably, the results in Table 3 have shown that the detected time of anomaly occurrences match with the recorded time and period of the haze incidences that had been reported. The severity ranking (number in bracket) of the detected anomaly curves was determined using the computed Mahalanobis distance value $D(x_i)$ while the maximum level ever achieved indicated those detected days were anomalies since all of them contained high PM_{10} concentration level whereby their maximum level was far above the maximum of the median curve (i.e. Klang=73 μg m^{-3}, Kuala Selangor=55 μg m^{-3} and Petaling Jaya=49 μg m^{-3}). The ranking result has shown that 10th August 2005 (indicated by number 1 in brackets) was the most severely polluted day at the Klang and Kuala Selangor stations while it was 11th August 2005 at the Petaling Jaya station. Based on these findings, the results support the effectiveness of the applied anomaly detection method used in this study. These results revealed that the significant top 20 anomalous level is mostly dominant during the SW monsoon season with the majority of the severe incidences occurring in the year 2005. The drier climate during the SW monsoon and a more serious forest fire in Sumatra, are believed to be the main reasons for this. In accordance with the study conducted by Fuller and Murphy (2006), the forest clearing fire is strongly linked to the monsoonal system, where the dry season is the favored season for burning activity.

3.3. The influence of wind variable on the severity of anomaly curves

Variations in air pollutant concentrations are strongly related with variations in meteorological changes (Chang and Lee, 2008). Juneng et al. (2011) using the regression model showed that meteorological factors, including: temperature, humidity and wind speed are significant in modulating the variation of PM_{10} over the Klang Valley region during the southwest (SW) monsoon. The relationship between local wind speed and average PM_{10} level was found to be negative at all air monitoring stations, namely: Klang, Kuala Selangor and Petaling Jaya. However, since the model used focused on the average PM_{10} data, it does not explain the extreme value observations.

In this study, in order to examine the possible contribution of meteorological factors, particularly focusing on wind and the diurnal fluctuations of the extreme anomaly, two graphs were plotted (as depicted in Figure 3); a sample of 10 most extreme PM_{10} curves and the graph of the corresponding wind speed curves. Obtained results indicated that wind speed positively influenced the extreme PM_{10} anomalies at Klang and Petaling Jaya. On the contrary, the relationship was negative at Kuala Selangor. At the 5% significance level, Spearman Correlation Coefficient analysis provided evidence of a positive relationship between wind speed and extreme PM_{10} anomalies at the Klang with a coefficient value r=0.39 and a corresponding p–value=0.03. A positive relationship (r=0.66, p–value=0.00) between wind speed and extreme PM_{10} anomalies was also observed for Petaling Jaya. On the other hand, a negative correlation was observed between wind speed and extreme PM_{10} anomalies at Kuala Selangor (r=–0.74, p–value=1.00).

3.4. The profile of anomaly occurrences with respect to an annual, monthly and day–of–the–week basis

All of the anomalies detected in the data set have also been extracted and investigated to identify and study the patterns of abnormal behavior of daily PM_{10}. Hence, several profiles of anomaly occurrences were used to describe and summarize the changes, both relating to time (temporal: on an annual, monthly

and day–of–the–week basis) and stations (spatial), see Figure 4. Noticeably, the trend of the frequency of anomaly occurrences at all three stations indicated that the years 2005, 2006 and 2009 were the most affected and that 2010 was the least affected by abnormal PM_{10} levels (Figure 4a). In terms of individual stations, Klang and Petaling Jaya were observed to have experienced the most frequent anomaly occurrences in the year 2005, while Kuala Selangor encountered them in the year 2006. In general, Kuala Selangor could be said to be the most prone station in terms of the possibility of being affected by anomalous PM_{10} behavior, followed by Klang and the least prone station; Petaling Jaya. The reason could be due to the influential factor of warmer and drier temperature during the El Nino period. The increase in the forest fires activity in the Southeast Asian Maritime Continent during the dry season (Reid et al., 2012) consequently exacerbated the PM_{10} pollution to degrading levels.

The highest frequency of occurrence was also observed in the drier weather period, namely the SW monsoon. In the case of Klang, this was in August, whilst for Kuala Selangor and Petaling Jaya, this occurred in July as is shown in Figure 4b. A secondary peak of occurrence took place during the NE monsoon in Klang and Kuala Selangor, while Petaling Jaya station exhibited two secondary peaks; one in the month of February and the other in the month of April during the inter–monsoon period. The graph also shows that almost zero anomalies were obtained in November. This could be attributed to the wash–out effect from the higher precipitation levels during this time (at the early NE monsoon). The results clearly indicate the "monsoonal effect" on the frequency of anomaly occurrences.

On a weekly scale, Figure 4c shows that the frequencies of anomalies fluctuate with an increasing pattern from Monday to Friday at Kuala Selangor and Petaling Jaya station. The frequencies however drop on Saturday and Sunday. On the other hand, only a slight increase in frequency between Friday and Saturday was observed at Klang station, this was then followed by a drop in the value on Sunday. The main road to Kuala Lumpur (the capital city of Malaysia), which links many districts, such as Kuala Langat, Banting, Kuala Selangor, etc., is located in Klang. Since Saturday is a school holiday, the high number of anomalies could be due to short vacations or mini trip activity. It is possible that the background of the station may lead to the depletion rate of PM_{10} at Klang station being lower than was observed at the other stations. Based on these results, the increase in frequency during weekdays (Monday to Friday) and the decrease during the weekend (Saturday and Sunday) may indicate the existence of the "weekend effect" phenomenon.

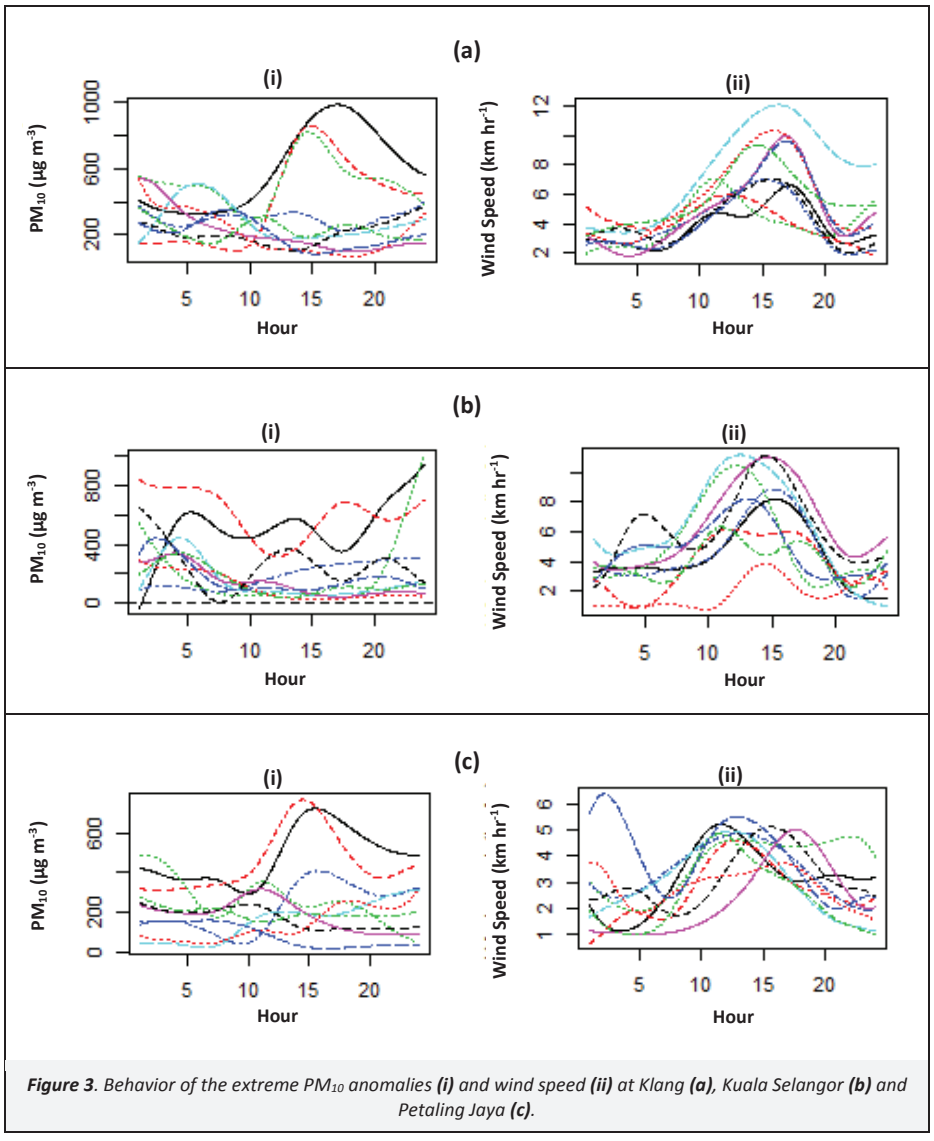

***Figure 3**. Behavior of the extreme PM_{10} anomalies **(i)** and wind speed **(ii)** at Klang **(a)**, Kuala Selangor **(b)** and Petaling Jaya **(c)**.*

Table 3. *Analysis of matching date between the top 20 detected anomalies and the reported (historical) haze incidences*

Reported Haze Incidences from 2005–2009			Monsoon	Results of the Detected Date of Anomalies and the Maximum Concentration Level (μg m^{-3}) from the Data					
Period of Occurrences	Description	Influential Factors		Klang	Max.	Kuala Selangor	Max.	Petaling Jaya	Max.
1 August–15 August 2005	The central part (including Klang Valley), eastern and northern part of Peninsular Malaysia experienced severe haze incidences.	Transboundary pollution due to land and forest fire in Sumatra island.	SW	7/7/2005 (13)	356	9/7/2005 (6)	978	2/8/2005 (8)	463
				8/7/2005 (8) [a]	353	11/7/2005 (4)	524	6/8/2005 (14)	169
				6/8/2005 (9)	669	7/8/2005 (7) [a]	321	7/8/2005 (9) [a]	283
				7/8/2005 (4) [a]	460	8/8/2005 (19) [a]	600	8/8/2005 (12) [a]	271
				8/8/2005 (15)	412	10/8/2005 (1)	995	9/8/2005 (4)	434
				9/8/2005 (2) [a]	901	11/8/2005 (2) [a]	880	10/8/2005 (2) [a]	806
				10/8/2005 (1) [a]	994	12/8/2005 (10)	444	11/8/2005 (1) [a]	759
11 August–13 August 2005	Malaysia has declared haze emergency in Pelabuhan Klang and Kuala Selangor.			11/8/2005 (3) [a]	905	11/9/2005 (17) [a]	604	12/8/2005 (3) [a]	542
				12/8/2005 (11) [a]	275	14/9/2005 (5)	857	8/6/2006 (5)	314
				11/6/2009 (16) [a]	262	30/7/2009 (11)	404	5/8/2009 (15)	287
				12/6/2009 (17) [a]	475	1/8/2009 (9)	799		
				15/7/2009 (18)	421	5/8/2009 (16)	290		
Between February and March 2005	Some areas in Klang Valley experienced short periods of slight to moderate haze.	Local peat–land fires.		16/7/2009 (10)	413				
				4/8/2009 (20) [a]	325				
				5/8/2009 (19)	457				
Between mid–May until mid–October 2005	Several parts of Malaysia experience short term mild to severe haze episodes.	Transboundary pollution and partly local peat–land fires.	NE	17/2/2005 (7)	687	26/2/2005 (13)	326	24/2/2005 (19) [a]	188
				23/2/2005 (14) [a]	536	19/3/2005 (20)	433	25/2/2005 (11) [a]	222
				25/2/2005 (5) [a]	533	20/3/2005 (14)	365	26/2/2005 (6) [a]	360
				20/2/2008 (6)	522			22/3/2005 (13) [a]	256
				8/10/2006 (12)	511	8/10/2005 (3) [a]	418	5/10/2006 (17)	225
From July until October 2006	Malaysia experienced short periods of slight to moderate haze.	Transboundary pollution.	Inter–monsoon			4/10/2006 (18) [a]	341	6/10/2006 (18) [a]	183
						6/10/2006 (12) [a]	641	7/10/2006 (10)	178
						5/10/2006 (8)	418	9/10/2006 (7) [a]	279
						9/10/2006 (15) [a]	574	16/10/2006 (16)	181
During June to August 2009	The number of healthy days has slightly decreased compared to 2010.	Transboundary pollution and partly local peat–land fires.						18/10/2006 (20) [a]	173

[a] Red anomalies (RA) of functional data

Additively, the monthly and day–of–week profile of the anomaly occurrences indicated the potential existence of the "monsoonal effect" and the "weekend effect" phenomena at the study locations. Thus, using the mean distribution of PM_{10} concentration levels, the hypotheses of significant differences in the mean PM_{10} levels between the SW and NE monsoons and between weekdays and weekends were tested. In this study, PM_{10} "weekend effect" phenomena was defined as the difference in the PM_{10} level between weekdays (Monday to Friday) and the weekend (Saturday and Sunday). During the weekend, the emissions of anthropogenic precursors are believed to decrease from weekday values because major sources of precursors, such as motor vehicles and power plants, may be less active on weekends.

On the whole, t–test analysis of the mean level ($p<0.05$) provided evidence that the average diurnal PM_{10} level during the SW monsoon was significantly higher than during the NE monsoon. The p–value obtained for the test of hypothesis on the "weekend effect" also produced the same results. It was thus established that the phenomenon exists at all considered stations. The t–test results only represent the difference in the overall hours. Specifically, by an hourly scale, as shown by the functional mean of the PM_{10} concentration level in Figure 5a, significantly higher levels were observed at all hours of the day except hours between 10:00 am and 12:00 noon at the Klang station. For the Kuala Selangor and Petaling Jaya stations, the levels were always lower during the NE monsoon as compared to the SW monsoon at all hours of the day.

On a temporal basis, the functional, descriptive statistical mean of the diurnal level between the weekdays and weekend (see Figure 5b) has shown that the dominant difference in the level occurring after dawn until midnight (i.e. during anthropogenic activity time) was always higher on weekdays than weekends at both Klang and Petaling Jaya stations. On the other hand, the same pattern was observed to occur across the hours of the day at Kuala Selangor station. Of the three stations, the "weekend effect" phenomenon in Petaling Jaya was far more significant. It is believed that this was due to the active emission sources on weekdays as compared to weekends.

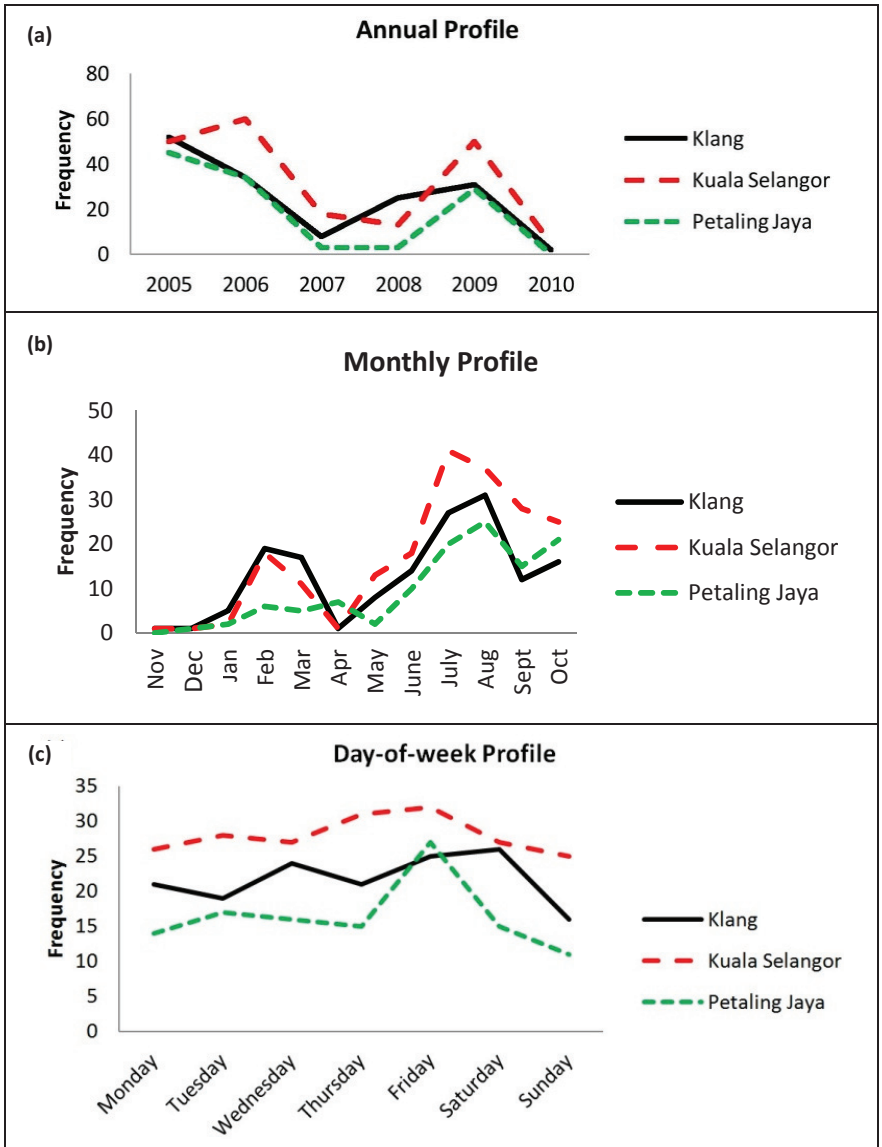

Figure 4. *Annual* ***(a)****, monthly* ***(b)*** *and day–of–week* ***(c)*** *profile of anomaly occurrences.*

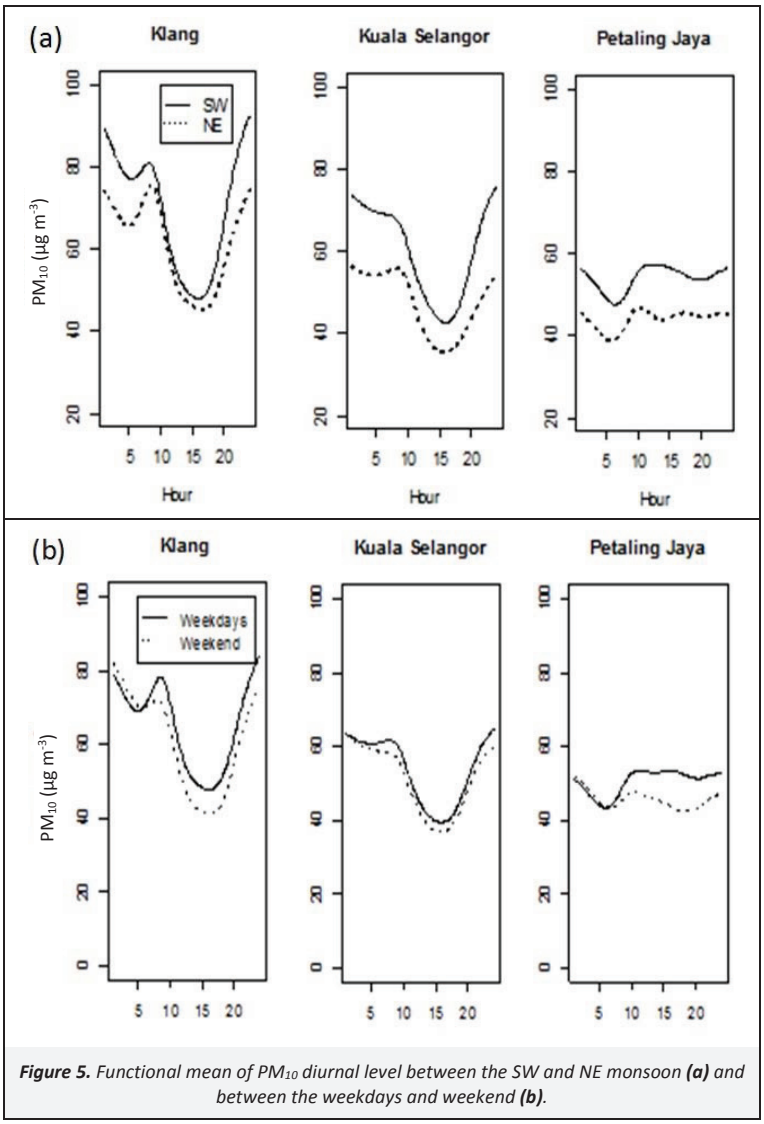

Figure 5. *Functional mean of PM_{10} diurnal level between the SW and NE monsoon* ***(a)*** *and between the weekdays and weekend* ***(b)***.

4. Summary and Conclusion

A combination of the robust projection pursuit and Mahalanobis distance method used by Hyndman and Shang (2010) were employed to identify the daily PM_{10} anomalies in the form of curves or functional data at three selected air quality monitoring stations in the Klang Valley region of the Malaysian Peninsular. This study shows that anomalies detection was a useful statistical technique in studying and investigating abnormalities in the daily PM_{10} process system. Using functional data analysis, the whole structure of daily diurnal patterns of anomalies could be visualized. It is also shown that functional data for extreme anomalies and wind speed offers a solution to investigate the relationship between two extreme data. The approach could overcome the problem facing by Juneng et al. (2011) due to the incapability of regression method used.

The detected anomalies from the data set represent interesting annual, monthly and day–of–the–week patterns of behavior in their frequency of occurrence. Years with El Nino events, such as 2005, 2006 and 2009, resulted in the highest frequency of occurrences. The dry season characterized by the SW monsoon was the dominant period of anomalies, with the months of July and August being the most frequent months where anomalies occurred. Transboundary sources were identified as being a major influence. Another interesting peak was in the month of February during the NE monsoon season where the causes were attributed to local sources. The increasing pattern in the frequency of anomalies during weekdays compared to week-ends indicated the impact of active sources of PM_{10} such as motor vehicles.

The study has also provided evidence to demonstrate the existence of the "monsoonal effect" and "weekend effect" phenomena at the study locations. Of the three stations, Kuala Selangor was found to experience the most significant "monsoonal effects" while Petaling Jaya experienced the most significant "weekend effect". Wind speed was shown to positively influence the extreme anomalies at the Klang and Petaling Jaya stations.

Based on the study findings, it was found that the stations' location and background, wind speed along with seasonal (monsoon) and weekdays–weekend variation play important role in influencing PM_{10} anomalies. In addition, the profile of anomalies could be utilized as a guideline for analyzing the effectiveness of current air quality control regulations or even for the planning of new mitigation policies. Given the appropriateness of the application, we suggest the incorporation of anomaly detection as an important step in data quality control systems as well as in efforts aimed at air pollution monitoring.

Acknowledgments

The authors would like to thank the Department of Environment Malaysia for providing the information and data. The work is supported by the UKM's Research University Grant [UKM–AP–2011_19]. The authors are also grateful to the three anonymous reviewers for their helpful comments.

References

Abas, M.R., 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, 1089–1095.

Acero, J.A., Simon, A., Padro, A., Coloma, O.S., 2012. Impact of local urban design and traffic restrictions on air quality in a medium–sized town. *Environmental Technology* 33, 2467–2477.

Acuna, E., Rodriguez, C., 2004. The treatment of missing values and its effect in the classifier accuracy, in *Classification, Clustering and Data Mining Applications*, edited by Banks, D., House, L., McMorris , F.R, Arabie, P., Gaul, W., Springer–Verlag Berlin–Heidelberg, New York, pp. 639–648.

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

Ahamad, F., Latif, M.T., Tang, R., Juneng, L., Dominick, D., Juahir, H., 2014. Variation of surface ozone exceedance around Klang Valley, Malaysia. *Atmospheric Research* 139, 116–127.

Awang, M.B., Jaafar, A.B., Abdullah, A.M., Ismail, M.B., Hassan, M.N., Abdullah, R., Johan, S., Noor, H., 2000. Air quality in Malaysia: Impacts, management issues and future challenges. *Respirology* 5, 183–196.

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 the Klang Valley, Malaysia. *Air Quality Atmosphere and Health* 3, 53–64.

Chang, S.C., Lee, C.T., 2008. Evaluation of the temporal variations of air quality in Taipei City, Taiwan, from 1994 to 2003. *Journal of Environmental Management* 86, 627–635.

Davis, J.J., Clark, A.J., 2011. Data preprocessing for anomaly based network intrusion detection: A review. *Computers & Security* 30, 353–375.

DOE (Department of Environment Malaysia), 2010. Malaysia Environ Quality Report 2009. Ministry of Science, Technology and Environment, Kuala Lumpur.

DOE (Department of Environment Malaysia), 2007. Malaysia Environ Quality Report 2006. Ministry of Science, Technology and Environment, Kuala Lumpur.

DOE (Department of Environment Malaysia), 2006. Malaysia Environ Quality Report 2005. Ministry of Science, Technology and Environment, Kuala Lumpur.

Dominick, D., Juahir, H., Latif, M.T., Zain, S.M., Aris, A.Z., 2012. Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric Environment* 60, 172–181.

Field, R.D., van der Werf, G.R., Shen, S.S.P., 2009. Human amplification of drought–induced biomass burning in Indonesia since 1960. *Nature Geoscience* 2, 185–188.

Filzmoser, P., 2005. Identification of multivariate outliers: A performance study. *Australian Journal of Statistics* 34, 127–138.

Fuller, D.O., Murphy, K., 2006. The ENSO – Fire dynamic in insular Southeast Asia. *Climatic Change* 74, 435–455.

Garces, H., Sbarbaro, D., 2011. Outliers detection in environmental monitoring databases. *Engineering Applications of Artificial Intelligence* 24, 341–349.

Garcia–Teodoro, P., Diaz–Verdejo, J., Macia–Fernandez, G., Vazquez, E., 2009. Anomaly–based network intrusion detection: Techniques, systems and challenges. *Computers & Security* 28, 18–28.

Hauskrecht, M., Batal, I., Valko, M., Visweswaran, S., Cooper, G.F., Clermont, G., 2013. Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics* 46, 47–55.

Hawkins, S.J., Gibbs, P.E., Pope, N.D., Burt, G.R., Chesman, B.S., Bray, S., Proud, S.V., Spence, S.K., Southward, A.J., Southward, G.A., Langston, W.J., 2002. Recovery of polluted ecosystems: The case for long–term studies. *Marine Environmental Research* 54, 215–222.

Huang, J.H.Z., Shen, H.P., 2004. Functional coefficient regression models for non–linear time series: A polynomial spline approach. *Scandinavian Journal of Statistics* 31, 515–534.

Hyndman, R.J., Shang, H.L., 2010. Rainbow plots, bag plots and box plots for functional data. *Journal of Computational and Graphical Statistics* 19, 29–45.

Juneng, L., Latif, M.T., Tangang, F., 2011. Factors influencing the variations of PM_{10} aerosol dust in Klang Valley, Malaysia during the summer. *Atmospheric Environment* 45, 4370–4378.

Keuken, M.P., Moerman, M., Voogt, M., Blom, M., Weijers, E.P., Rockmann, T., Dusek, U., 2013. Source contributions to $PM_{2.5}$ and PM_{10} at an urban background and a street location. *Atmospheric Environment* 71, 26–35.

Lee, S., Ho, C.H., Choi, Y.S., 2011. High – PM_{10} concentration episodes in Seoul, Korea: Background sources and related meteorological conditions. *Atmospheric Environment* 45, 7240–7247.

Liebl, D., 2013. Modeling and forecasting electricity spot prices: A functional data perspective. *Annals of Applied Statistics* 7, 1562–1592.

Lin, S., Brown, D.E., 2006. An outlier–based data association method for linking criminal incidents. *Decision Support Systems* 41, 604–615.

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.

Mahmud, M., 2009. Simulation of equatorial wind field patterns with TAPM during the 1997 haze episode in Peninsular Malaysia. *Singapore Journal of Tropical Geography* 30, 312–326.

Matsueda, H., Inoue, H.Y., Ishii, M., Tsutsumi, Y., 1999. Large injection of carbon monoxide into the upper troposphere due to intense biomass burning in 1997. *Journal of Geophysical Research – Atmospheres* 104, 26867–26879.

Muniz, C.D., Nieto, P.J.G., Fernandez, J.R.A., Torres, J.M., Taboada, J., 2012. Detection of outliers in water quality monitoring samples using functional data analysis in San Esteban Estuary (Northern Spain). *Science of the Total Environment* 439, 54–61.

R Development Core Team, 2008. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. http://www.R–project.org, accessed in January 2013.

Ramsay, J.O., Silverman, B.W., 2006. *Functional Data Analysis*, Springer, New York.

Ramsay, J.O., Silverman, B.W., 2002. *Applied Functional Data Analysis: Methods and Case Studies*, Springer, New York.

Ramsay, J.O., Wickham, H., Graves, S., Hooker, G., 2013. http://www.functionaldata.org, accessed in June 2013.

Reid, J.S., Xian, P., Hyer, E.J., Flatau, M.K., Ramirez, E.M., Turk, F.J., Sampson, C.R., Zhang, C., Fukada, E.M., Maloney, E.D., 2012. Multi–scale meteorological conceptual analysis of observed active fire hotspot activity and smoke optical depth in the Maritime Continent. *Atmospheric Chemistry and Physics* 12, 2117–2147.

Rousseeuw, P.J., Van Driessen, K., 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41, 212–223.

Schauer, J.J., Rogge, W.F., Hildemann, L.M., Mazurek, M.A., Cass, G.R., Simoneit, B.R.T., 1996. Source apportionment of airborne particulate matter using organic compounds as tracers. *Atmospheric Environment* 30, 3837–3855.

Shaadan, N., Deni, S.M., Jemain, A.A., 2012. Assessing and comparing PM_{10} pollutant behaviour using functional data approach. *Sains Malaysiana* 41, 1335–1344.

Shang, H.L., Hyndman, R.J., 2013. http://sites.google.com/site/hanlinshangswebsite/, accessed in July 2013.

Sharma, A., Panigrahi, P.K., 2012. A Review of financial accounting fraud detection based on data mining techniques. *International Journal of Computer Application* 39, 37–47.

Shon, Z.H., Kim, K.H., Song, S.K., Jung, K., Kim, N.J., Lee, J.B., 2012. Relationship between water–soluble ions in $PM_{2.5}$ and their precursor gases in Seoul megacity. *Atmospheric Environment* 59, 540–550.

Tangang, F.T., Latif, M.T., Juneng, L., 2010. Climate change: Is Southeast Asia up to the Challenge?: The roles of Climate Variability and Climate Change on Smoke Haze Occurrences in Southeast Asia region, IDEAS Reports – Special Reports, edited by Kitchen, N., LSE IDEAS, London School of Economics and Political Science, London, UK.

Torres, J.M., Nieto, P.J.G., Alejano, L., Reyes, A.N., 2011. Detection of outliers in gas emissions from urban areas using functional data analysis. *Journal of Hazardous Materials* 186, 144–149.

Wrobel, A., Rokita, E., Maenhaut, W., 2000. Transport of traffic–related aerosols in urban areas. *Science of the Total Environment* 257, 199–211.