PREDICTING HOURLY PM10 CONCENTRATION IN SEBERANG PERAI AND PETALING JAYA USING LOG-NORMAL LINEAR MODEL

[1]ROSSITA M.YUNUS, [2]MASUD M. HASAN

University of Malaya, Kuala Lumpur, MALAYSIA
E-mail: [1]rossita@um.edu.my, [2]mmhusc@gmail.com

Abstract- Prediction of particulate matter (PM10) concentration is useful to assist future planning in the context of environmental problems. The aim of this study is to predict hourly PM10 concentration by considering the probability distribution, the serial correlation between subsequent observations and the seasonal pattern of the PM10 concentration. We propose to analyze the data using the log-normal linear models with potential predictors. In the initial study, three probability distributions (Weibull, gamma, and log-normal) were considered to fit the hourly PM10 concentration for two stations of Peninsular Malaysia – Seberang Perai and Petaling Jaya from 2008-2009 and 2014-2015. Within the distributions, the log-normal distribution was found appropriate. Then, the log-normal linear model with sine and cosine terms, and lagged as predictors were fitted to the PM10 concentration data. The Likelihood Ratio Test (LRT) and the Akaike Information Criterion (AIC) were used to assess model appropriateness. Diagnostic QQ plots indicate that the models fit the data well except in the extreme tails. The models were fitted using the first 75% of data and validated using the remainder.
Using the model, prediction of inadequate hourly PM10 records is possible.

Index Terms- air quality, lagged, log-normal linear model, regression, statistical modelling.

I. INTRODUCTION

Particulate matter (PM10) is among the most harmful pollutants to the environment and respiratory system, and a lot of research were conducted to develop strategies to reduce the emission of the PM10 concentration. Predicting the PM10 concentration will have important application to health [1]-[5], water acidity [6], and visibility [3],[7],[8]. However, predicting accurately the hourly PM10 concentration is not an easy task, and it lies in the statistical model that is used.

The studies of PM10 concentration in Malaysia are more focusing on determining the distribution that fit well the data [9]-[11]. The log-normal distribution has been widely used for application in the environmental data pollution, since over three decades ago, since the papers by Larsen [12]-[14], and this may due to its theoretical properties and its relationship with the normal distribution [15]. The distribution is found suitable for modeling the PM10 1 R.M. Yunus thanks University of Malaya (Research Grant no. RG369-15AFR) for funding this research concentration in many current case studies for most places worldwide [16]-[18]. The first step to data modeling is to determine the distribution of PM10. However, such study alone does not help in predicting the PM10 concentration at a given hour.

The multiple linear regression models have been widely used to model air concentration in many studies [19]-[23] for a variety purposes and these include for predicting the hourly PM10 concentration. The aim of this study is to predict hourly PM10 concentration levels, which are recorded as time series, on the basis of the probability distribution, the serial correlation of observations and seasonal pattern of the PM10 concentration. The seasonal pattern or other cyclical pattern (repeat over time in a cyclic manner) of air quality data is evident for many studies and the pattern is often modeled using models with sine and cosine terms [24]-[26]. The seasonal pattern has an impact on the PM10 concentration in Malaysia. Based on descriptive statistics and graphical plots, the concentration of PM10 in Peninsular Malaysia is higher in June - August, as compared to November – February [27]-[28].

As noted by [29], a model should take into account on the serial correlation exhibited from air quality observations that are close together in time. According to [30], air quality values in preceding time steps are similar to the initial conditions for the air quality prediction in the following time steps, and they have often been considered as a predictor. Many studies show that including lagged PM10 as a predictor improved modeling results [31]. For instance, [32] showed that lagged PM10 is more important than temporal or meteorological variables in predicting the PM10 concentration in urban areas.

In this paper, we fit the log-normal linear model to hourly PM10 concentration using the lagged, and the sine and cosine terms as potential predictors. The models are useful for simulating and predicting future events. In this paper, the focus is on the hourly PM10 concentration from two stations in Peninsular Malaysia. First, the properties of the PM10 concentration were studied to determine an appropriate theoretical probability distribution within some skewed distributions. Then, considering this

distribution, the log-normal models were fitted to the data with potential predictors. Various model selection criteria are used to obtain the best-fit model. This paper is organized as follows. Section 2 discusses the data used in this study. The log-normal linear models which are used to model the data are given in Section 3. The models considered for the analysis of PM10 concentration data are described in Section 4. Section 5 discusses the results of fitting the models to PM10 concentration data, and includes the diagnostic analysis that evaluates how well the model fits the data. Finally, concluding remarks are given in Section 6.

II. STUDY AREAS, DATA AND PRELIMINARIES

The study considers hourly PM10 concentration data (in $\mu g/m^3$) from two case study stations in Peninsular Malaysia. The station Seberang Perai is located in the northern part of the country with geographical coordinates of 5° 21' north latitude and 100° 24' east longitude. Petaling Jaya is located at the west coast with geographical coordinates, 3° 06' north latitude and 101° 39' east longitude. The studied data series (available for the period from Jan 1998 to Dec 1999, and from Jan 2014 to Dec 015) were obtained from the Department of Environmental Malaysia, and from the website of Data Terbuka Malaysia, respectively.

Petaling Jaya is an industrial city with about half a million populations and more than 2,780 industrial projects. The industries consist of electronic and electrical industry, machine manufacturing and fabricated metal products, and chemical production (Majlis Bandaraya Petaling Jaya 2011). In 2013, total population of Seberang Perai, a district of Penang state was estimated at about 0.94 million (Kerajaan Negeri Pulau Pinang). The district has more than 600 industrial projects from various sectors such as textile, electric and electronic, and manufacturing.

The air quality of both cities has been affected by rapid growth of industrial activities and traffic densities. Pollutants emitted from these sources into the atmosphere led to increasing level of air pollution index (API), as measured by PM10 concentration level.

Table I shows summary statistics of the PM10 concentration of the studied stations. On an average, hourly PM10 concentration of Petaling Jaya (47 $\mu g/m^3$) is larger than Seberang Perai (39.87). The highest score of PM10 concentration for Petaling Jaya and Seberang Perai were recorded as 257 $\mu g/m^3$ and 197 $\mu g/m^3$, respectively. The stations usually receive higher PM10 concentration during the northeast monsoon (November –February) and the southwest monsoon (June – August), and lower PM10 concentration during the inter-monsoon (March-May, September – October) (see Fig. 1). Exceptionally high PM10 concentration levels were recorded from September to October in 2015 for both stations.

To determine an appropriate distribution for hourly PM10 concentration of the studied stations, this study considers three skewed probability distributions – Weibull, gamma, and log-normal. The distribution of hourly PM10 concentration is skewed to the right (Fig. 2). The lognormal and gamma distributions seem more appropriate for the hourly PM10 concentration data, than the Weibull distribution, and they fit well to Petaling Jaya station than Seberang Perai station.

Table I. Summary statistics of the hourly PM10 concentration for Seberang Perai and Petaling Jaya stations.

	Mean	Standard deviation	IQR*	Minimum	Maximum
Seberang Perai	39.87	14.93	16	10	197
Petaling Jaya	47.28	20.06	21	12	257

*IQR, interquartile range.

III. THE LOG-NORMAL LINEAR MODEL

The probability density function of the log-normal random variable , is given by

$$f(y;\mu,\sigma) = \frac{1}{y\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\left(\frac{\log(y)-\mu}{\sigma}\right)^2\right],$$

where μ and σ are mean and standard deviation of $\log(Y)$.

Important properties of the log-normal distribution that make it appealing for use in PM10 concentration modeling include: it is capable of modelling skewed continuous data and can be incorporated into the linear model.

For modelling hourly PM10 concentration data using the log-normal distribution, it is assumed that any hourly record t produces an amount of concentration y_t, and y follows a log-normal distribution with mean e^{μ} and variance e^{σ^2}.

Consequently, the log-normal linear models [33] of the form $\log(Y) = X^T\beta + e$, where X is a vector of predictors, β is a vector of regression parameters, and $e \sim N(0,\sigma^2)$. The models were fitted to the hourly PM10 data with a set of predictors.

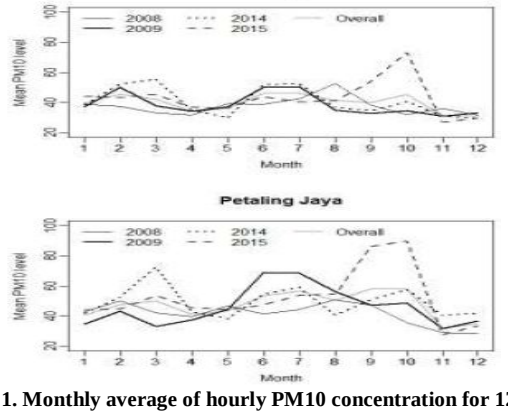


Fig. 1. Monthly average of hourly PM10 concentration for 12 months.

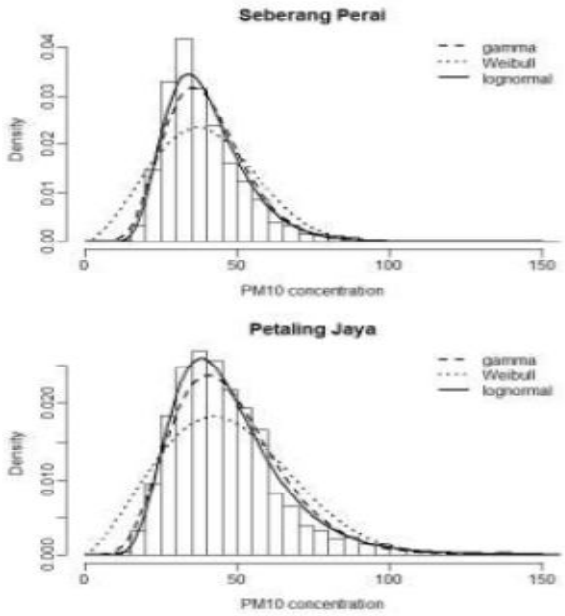


Fig. 2. Histogram of hourly PM10 concentration for Seberang Perai and Petaling Jaya. Curve lines are the theoretical density function from estimated distributional parameters.

IV. MODEL FOR THE HOURLY PM10 DATA

Hourly PM10 concentration shows an annual cycle, and thus to adjust for the seasonal variability (see Fig. 1), sine and cosine terms were included into the model. The partial autocorrelation function (PACF) of the quantile residuals from this model (Fig. 3), with just the sine and cosine terms as predictors, shows significant coefficients at lag 24 (hour). For this reason, the logarithm of the previous 24th hour PM10 concentration is included as one of the predictors. Finally, the following fitted models were compared:

$$\text{Model 1:} \quad \log \hat{Y}_t = \hat{\beta}_{10} + \hat{\beta}_{11} \sin\left(\frac{2\pi d}{365}\right) + \hat{\beta}_{12} \cos\left(\frac{2\pi d}{365}\right);$$

$$\text{Model 2:} \quad \log \hat{Y}_t = \hat{\beta}_{20} + \hat{\beta}_{21} \sin\left(\frac{2\pi d}{365}\right) + \hat{\beta}_{22} \cos\left(\frac{2\pi d}{365}\right) + \hat{\beta}_{23} \log(Y_{t-24});$$

where d =day of the year (1=1st January, 2=2nd January, ... and so on), t represents the individual hour, Y_{t-24} is the PM10 concentration of the previous 24th hour.

Model 1 is the simplest one with only sine and cosine terms as predictors and is nested within the Model 2. Model 2 includes PM10 concentration from the previous 24 hour (on log scale) in addition to the sine and cosine terms.

V. MODEL RESULT

Results of the models fitted to the hourly PM10 concentration from the studied stations are presented in Tables II and III. The cyclical sine and cosine terms have significant impact on hourly log PM10. The partial autocorrelation function of quantile residuals for model with only sine and cosine terms indicates significant lag 24 correlation of PM10 events. Including PM10 of previous day in the model significantly reduced the correlation.

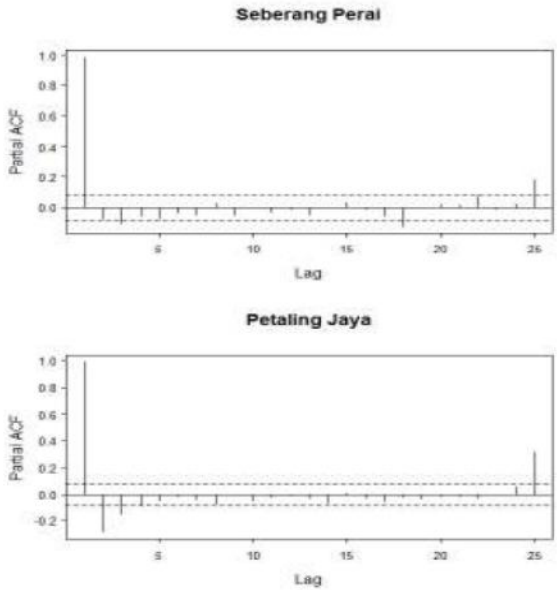


Fig. 3. Partial autocorrelation of the quantile residuals from model with only sine and cosine terms shows significant deviations from zero correlation at lag 1, 25, suggesting the same 24 hours pattern. The dashed line represents the 95% confidence level.

Table II . The estimated values of the coefficient parameters of the predictors in the fitted models for Seberang Perai.

	Model 1		Model 2	
F-value			34156*	
AIC	17280		-5776.2	
	$\hat{\beta}_j$	t-value	$\hat{\beta}_j$	t-value
Intercept	2.06	11.76*	0.49	4.02*
Sine	0.48	9.87*	0.13	3.86*
Cosine	1.51	8.84*	0.52	4.40*
Lagged			0.72	184.83*

* $p < 0.001$

Table III. The estimated values of the coefficient parameters of the predictors in the fitted models for Petaling Jaya.

	Model 1		Model 2	
F-value			40187*	
AIC	27994		2022.5	
	$\hat{\beta}_j$	t-value	$\hat{\beta}_j$	t-value
Intercept	2.49	12.19*	0.58	4.24*
Sine	0.36	6.32*	0.11	2.83*
Cosine	1.25	6.28*	0.36	2.72*
Lagged			0.75	200.48*

* $p < 0.001$
No superscript means not significant.

The t-test shows that sine and cosine terms are significantly important in Model 1. When considering the AIC, the model with lagged performs substantially better than the model with the sine and cosine terms. In addition, the F-test indicates that the improvement is statistically significant. Model 1 is nested in the Model 2. The F-test is used to compare the fit of Model 2 with Model 1. Based on the AIC, F-test, and t-test, the model with sine, cosine, and lagged is preferred for both stations.

The model can be used to predict PM10 concentration at a given hour. For example, the mean amount of PM10 concentration for Petaling Jaya (with Model 2) for hour t is:

$$\hat{Y}_t = \exp(0.58) \times \exp\left(0.11 \sin\left(\frac{2\pi d}{365}\right)\right) \times \exp\left(0.36 \cos\left(\frac{2\pi d}{365}\right)\right) \times \exp(0.75 \log(Y_{t-24}))$$

To predict the PM10 concentration at 8 p.m. on the 1st February 2014 for Petaling Jaya station, we substitute $t = 18308$ and $d = 32$ in the above equation. The amount of PM10 concentration at 8 p.m. on the 31st January 2014 $Y_{18284} = 46\ \mu g/m^3$. Thus, for 1st February 2014, the predicted PM10 concentration amount for Petaling Jaya at 8 p.m. is 45 $\mu g/m^3$ which is close to the observed value 42 $\mu g/m^3$.

A. Diagnostic Check

The QQ plots of the quantile residuals [34] were used to examine whether the log-normal linear models (preferred) fit the data adequately. Fig. 4 shows that all the residuals are located on or close to the straight line, suggesting well-fitting models. No large deviations were observed, except at the tails of the QQ plot for both stations. The diagnostic plots suggest that the log-normal distribution fits the hourly PM10 concentration reasonably well for all stations.

B. Model Validation

For validation purpose, the preferred model was fitted to the first 75% of the data and the rest was retained for validation purposes. Means of hourly PM10 concentration for different months for observed, validated, and simulated data are presented in Fig. 5. From the figure, the models fit reasonably well to the hourly PM10 concentration of the studied stations. However, discrepancies exist for average August – November PM10 concentration for both stations, due to haze from forest fires burning in neighboring regions. Further study may be required to improve the models used.

CONCLUSION

The log-normal linear models were proposed to model hourly PM10 concentration data with a set of predictors. Models were fitted and tested to hourly PM10 concentration from two stations from Peninsular Malaysia as case studies. Distinct seasonal trend on PM10 concentration is observed on all the studied stations. Higher concentration is observed during the northeast monsoon (November – February) and the southwest monsoon (June – August), and lower concentration during the inter-monsoon (March-May, September – October).

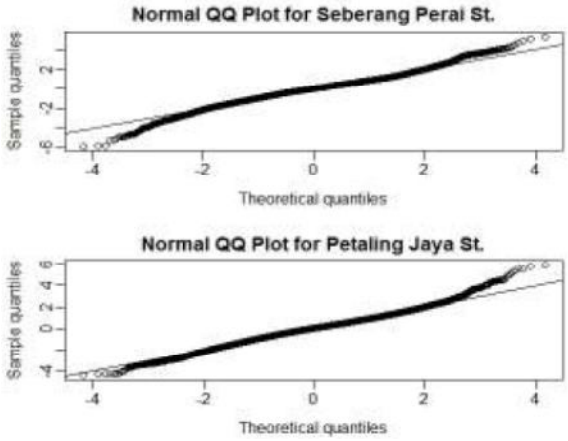


Fig. 4. QQ plots of quantile residuals for Model 2 of studied stations.

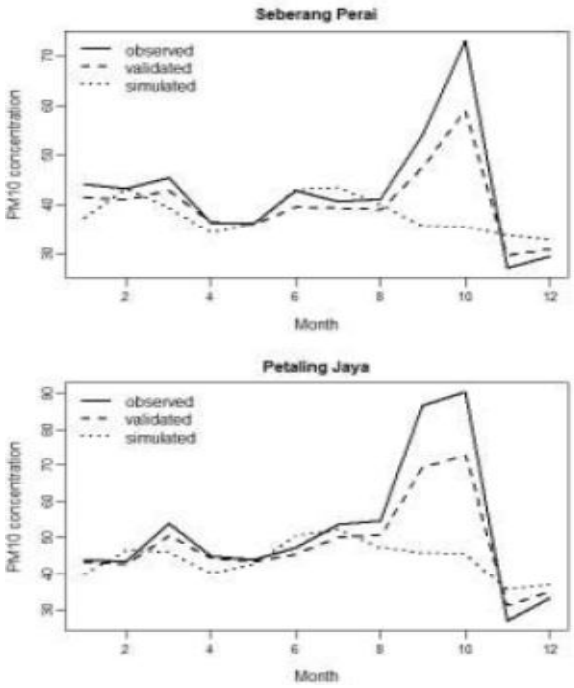


Fig. 5. Observed, validated, and simulated mean hourly PM10 concentration for various months.

Haze episode from late August to early November in 2015 is an exception to the usual trend of the PM10 concentration. The models can be used to generate synthetic series of hourly PM10 concentration. In addition, based on the log-normal linear models, one-day lead prediction of mean of hourly PM10 concentration can be used in environmental and health management, for example, to control the PM10 emissions from Malaysia industry and traffic. Including data from more stations, and adding climatologically or meteorological variables in the modelling framework may improve the performance of the model.

REFERENCES

[1] M. Gustafsson, B. Forsberg B., H. Orru, S. Åström, H. Tekie, and K. Sjöberg, "Quantification of population exposure to NO2, PM2.5 and PM10 and estimated health

impacts in Sweden 2010", IVL Report B 2197, Stockholm Sweeden: IVL Svenska Miljöinstitutet, Dec 2014.
[2] T.M. Habeebullah, "Modeling particulate matter (PM10) in Makkah, Saudi Arabia - A View Point of Health Impact", Journal of Clean Energy Technologies, vol. 2, no. 3, July 2014.
[3] R. Huang, G. Bian, T. He, L. Chen, and G. Xu, "Effects of meteorological parameters and PM10 on the incidence of hand, foot, and mouth disease in children in China", Int. J. Environ. Res. Public Health vol. 13, 481, May 2016.
[4] M.P. Keuken and P. Zandveld, S. van den Elshout, G. Hoek, and N. Janssen, "Health impact assessment of PM10 and EC in the city of Rotterdam (the Netherlands) in the period 1985-2008", Netherlands Research Program on Particulate Matter BOP II Report TNO-060-UT-2011-01068, 2011.
[5] M. McCarthy, J. P. Biddulph, M. Utley, J. Ferguson, and S. Gallivan, "A health impact assessment model for environmental changes attributable to development projects", J Epidemiol. Community Health, vol. 56, pp. 611-616, 2002.
[6] S. P. Shukla, and M. Sharma, "Neutralization of rainwater acidity at Kanpur, India", Tellus B., vol. 62 (3), pp.172-180, 2010.
[7] E.R. Jayaratne, S. Clifford, and L. Morawska "Atmospheric visibility and PM10 as indicators of new particle formation in an urban environment", Environ. Sci. Technol., vol. 49(21), pp. 12751-12757, Oct 2015.
[8] N. Vajanapoom, C.M. Shy, L.M. Neas, and D. Loomis, "Estimation of particulate matter from visibility in Bangkok, Thailand", Journal of Exposure Analysis and Environmental Epidemiology, vol. 11, 97 – 102, 2001.
[9] N.A. Razak, Y.Z. Zubairi, and R.M. Yunus, " Imputing missing values in modeling the PM10 concentrations", Sains Malaysiana, vol. 43 (10), pp. 1601-1609, 2014.
[10] N.F.F. Md Yusof, N.A. Ramli, A.S. Yahaya, et al., "Monsoonal differences and probability distribution of PM10 concentration", Environ. Monit. Assess., vol. 163, pp. 655-667, 2010.
[11] N. Sansuddin, N.A. Ramli, A.S. Yahaya, N.F.F. Md Yusof, N.A. Ghazali, and W.A. Al Madhoun, "Statistical analysis of PM10 concentrations at different locations in Malaysia", Environ. Monit. Assess., vol.180, pp.573-588, 2011.
[12] R.I. Larsen, "An air quality data analysis system for interrelating effects, standards and needed source reductions-Part 2", Journal of the Air Pollution Control Association, vol. 24 (6), pp. 551-558, 1974.
[13] R.I. Larsen, "A new mathematical model of air pollutant concentration averaging time and frequency", Journal of the Air Pollution Control Association, vol. 19, pp. 24-30, 1969.
[14] R.I. Larsen, "An air quality data analysis system for interrelating effects, standards and needed source reductions", Journal of the Air Pollution Control Association, vol. 23 (11), pp. 933-940, 1973.
[15] C. Marchant, V. Leiva, M. F. Cavieres, and A. Sanhueza, "Air contaminant statistical distributions with application to PM10 in Santiago, Chile", Reviews of Environmental Contamination and Toxicology, vol. 223, pp. 1-31, 2013
[16] N.M. N, M.M.A. Abdullah, C. Tan, N.A. Ramli, A.S. Yahaya, and N.F.M.Y. Fitri, "Modelling of PM10 concentration for industrialized area in Malaysia: A case study in Shah Alam", Physics Procedia, vol. 22, pp. 318-324, 2011.
[17] D.K. Papanastasiou, and D. Melas, "Application of PM10s Statistical Distribution to Air Quality Management—A Case Study in Central Greece", Water Air Soil Pollut., vol. 207, pp. 115-122, 2010.
[18] X. Wang, R.J. Chen, B.H. Chen, and H.D. Kan, "Application of statistical distribution of PM10 concentration in air quality management in 5 representative cities of China", Biomed. Environ. Sci., vol. 26(8), pp. 638-646, Aug 2013.
[19] A. Afzali , M. Rashid, B. Sabariah, and M. Ramli, "PM10 pollution: its prediction and meteorological influence in PasirGudang, Johor." IOP Conf. Series: Earth and Environmental Science, vol. 18(2014), 012100, 2014.
[20] A. K. Paschalidou, S. Karakitsios, S. Kleanthous, and P. A. Kassomenos, "Forecasting hourly PM10 concentration in Cyprus through artificial neural networks and multiple regression models: implications to local environmental management", Environmental Science and Pollution Research, vol. 18(2), pp. 316–327, 2011.
[21] J. C. M. Pires, F. G. Martins, S. I. V. Sousa, M. C. M. Alvim-Ferraz, and M. C. Pereira, "Prediction of the daily mean PM10 concentrations using linear models", American Journal of Environmental Sciences, vol. 4 (5), pp. 445-453, 2008.
[22] A.S. Sayegh, S. Munir, and T.M. Habeebullah, "Comparing the performance of statistical models for predicting PM10 concentrations", Aerosol and Air Quality Research, vol. 14, pp. 653–665, July 2014.
[23] A.Z. Ul-Saufie, A.S. Yahya, and N.A. Ramli, "Improving multiple linear regression model using principal component analysis for predicting PM10 concentration in Seberang Prai, Pulau Pinang", International Journal of Environmental Sciences, vol. 2, no. 2, Nov 2011.
[24] C. S. Cheng, M. Campbell, Q. Li, G. Li, H. Auld, N. Day, D. Pengelly, S. Gingrich, and D. Yap, "A Synoptic climatological approach to assess climatic impact on air quality in South-central Canada. Part I: Historical analysis", Water Air Soil Pollut, vol. 182, pp. 131–148, Jan 2007.
[25] V. Gvozdić, E. Kovač-Andrić, and J. Brana, "Influence of meteorological factors NO2, SO2, CO and PM10 on the concentration of O3 in the urban atmosphere of Eastern Croatia", Environ. Model. Assess., vol. 16, pp 491-501, Oct 2011.
[26] N. Otero , J. Sillmann , J. L. Schnell, H.W. Rust, and T. Butler, "Synoptic and meteorological drivers of extreme ozone concentrations over Europe", Environ. Res. Lett., vol. 11 (2), 024005, 2016.
[27] N. A. Abdullah, S. H. Shuhaimi, T.Y. Ying, A.H. Shapee, M. Mohamad, "The study of seasonal variation of PM10 concentration in Peninsular, Sabah and Sarawak", Malaysia Meteorological Department Ministry of Science, Technology and Innovation, Research Publication, No. 9/2011.
[28] N. Amil, M.T. Latif, M. F. Khan, and M. Mohamad, "Seasonal variability of PM2.5 composition and sources in the Klang Valley urban-industrial environment", Atmos. Chem. Phys., vol. 16, pp. 5357-5381, Apr 2016.
[29] W.R. Ott, "Environmental statistics and data analysis", Lewis Publishers, Boca Raton, FL, pp. 222, 1995.
[30] J. Hooyberghs, C. Mensink, G. Dumont, F. Fierens, and O. Brasseur, "A neural network forecast for daily average PM10 concentrations in Belgium", Atmos. Environ., vol. 39(18), pp. 3279–3289, 2005.
[31] T. Slini, A. Kaprara, K. Karatzas, N. Moussiopoulos, "PM10 forecasting for Thessaloniki, Greece", Environ. Model. Softw., vol. 21 (4), pp 559–565, Apr 2006.
[32] E. Stadlober, S. Hörmann, and B. Pfeiler, "Quality and performance of a PM10 daily forecasting model", Atmos. Environ., vol. 42, pp. 1098–1109, 2008.
[33] H. Shen and Z. Zhu, "Efficient mean estimation in log-normal linear models", Journal of Statistical Planning and Inference, vol. 138, pp. 552-567, 2008.
[34] P.K. Dunn and G.K. Smyth, "Randomised quantile residuals", Journal of Computational and Graphical Statistics, vol. 5, pp. 236–244, 1996.

★★★