Proc. of the Intl. Conf. on Advances In Applied Science and Environmental Engineering - ASEE 2014.

ISBN: 978-1-63248-004-0 doi: 10.15224/ 978-1-63248-004-0-02

Reducing Autocorrelation Effect in the Control Chart for PM10 Curves

Norshahida Shaadan, Abdul Aziz Jemain and Sayang Mohd Deni

***Abstract*—The traditional control chart for industry often requires identical and independent (i.i.d) assumptions. Thus, some modification on the control chart used in the industry need to be done before it can be used in the environmental application due to the difference in the nature of the data. The recorded environmental data are usually multivariate, correlated, and non-stationary and can be expressed as a function of time. In this present study, a control chart for monitoring curves data is to be applied where the hourly recorded data within a one day period is treated as multivariate point data and the occurrence is treated as a function of time. The control chart is constructed using the Functional Principal Component Analysis (FPCA) model. In the monitoring of quality indices overtime, autocorrelated data often negatively affects the performance of a control chart resulting into increase in the number of false alarm. Data pre-whitening approach is proposed during the monitoring phase of analysis to tackle the problem of data autocorrelation. Based on the seven years (2004-2010) recorded data from the Shah Alam air quality monitoring station which is located in the Selangor state of Peninsular Malaysia, after data pre-whitening has been conducted, the analysis results have shown that the false alarm rate in the control chart for the PM10 indices has reduced and the existence of lag 1 autocorrelation AR(1) effect is found insignificant. Thus, it is indicated that the employment of data pre-whitening approach able to improve the control chart performance.**

***Keywords*—control chart, PM10, monitoring, autocorrelation effect**

I. Introduction

The application of the quality control chart for the monitoring of product manufacturing performance is well known in the industry. In the environmental field, the application of the control chart has also gained interest for environmental management [1, 2, 3]. The use of the control chart as a visualization tool is proposed to check for data quality as well as to assist in controlling and managing environmental performance.

The control chart is used as a tool to monitor pollutant emission so that abnormal changes can be detected in a timely manner. When and where abnormal occurrence which is often associated with high pollutant levels can be identified, the possible sources can be further investigated and identified.

A high PM10 level which is often associated with haze incidences has become a common problem in Southeast Asian countries including Malaysia. Globally, a lot of researches have proven that exposure to the contaminated levels of PM10 has caused short and long term effects on human health [4-7]. As for Malaysia, a recent study by [8] has shown that PM10 is significantly related with respiratory mortality in the Klang Valley region of Peninsular Malaysia. Thus, in considering the negative impact, the effort of monitoring and investigating the air pollution problem is important. To help in synthesizing and analyzing environmental monitoring, the control chart is proposed as a universal format to summarize data sets and to point out any possible need for management action [9].

There are several types of control chart used to monitor environmental data in the literature. The most basic control chart is the Shewhart control chart for monitoring univariate data through the data transformation method [10]. When measured data consist of multiple criteria in which each criteria is inter-correlated, the multivariate control chart is used using the Principal Component Analysis (PCA) [11, 12] including the distance based method by [13].

The pollutant data such as PM10 in particular are normally taken on a daily basis and discretely recorded at every hour of a day period. The data can be treated as multivariate data where hours are considered as the variables. Naturally, it can be thought that pollutant data are continuous and arise as a function of time within the day process. Such kind of data is called functional data [14] which in this case they consist of daily PM10 pollutant curves. In order to monitor pollutant curves, a previous work by [15] has proposed a control chart which is based on the method used for product quality profile studied by [16].

Environmental observations or indices are often correlated with time. The situation is known as autocorrelation problem. Auto correlated data negatively affects the performance of a control chart resulting in increases in the number of false alarm [17-19]. False alarm refers to false signal; the detected abnormal points (indices) although they are actually not. Moreover, too many signals in the control chart would not be acceptable by the management because it requires more time and cost for investigation activity. Thus, a technique to overcome or at least reduce the impact of autocorrelation is necessary in the control chart establishment as well as using the control chart to monitor pollutant indices overtime.

To further enhance the usage of the proposed control chart in the previous work by [15], the auto

Norshahida Shaadan, Sayang Mohd Deni

Center for Statistical & Decision Science Studies, Faculty of Computer &
Mathematical Sciences, Universiti Teknologi MARA (UiTM),
40450, Shah Alam,
Malaysia

Abdul Aziz Jemain

DELTA, School of Mathematical Sciences, Faculty of Science & Technology,
Universiti Kebangsaan Malaysia (UKM), 43600, Bangi, Selangor,
Malaysia

the monitoring phase is now taken into consideration. Data pre-whitening approach is suggested in this study to tackle the problem of data autocorrelation. Specifically this study aims to examine the capability of data pre-whitening approach in reducing the lag 1 autocorrelation AR(1) effect in the monitoring the performance of the daily PM10 curves over time.

II. Methodology

Process control using control charts typically involves two phases [21]. Phase 1 is known as the base period analysis where the objective is to obtain a baseline data that represent the normal or stable state of pollutant process. A set of 'in control' data set that represents the norm or common process of the population is obtained. The subsequent period of the data analysis is known as phase 2. The objective of phase 2 is to conduct an online monitoring for future data using the information obtained at the phase 1 analysis as the benchmark.

A. *The Control Chart for PM10 Curves*

In [15], the phase 1 control starts with data conversion from discretely recorded form into curve form. Given a set of *N* daily recorded PM10 (y_j) data at discrete time points t_j with *j*=1,...,*n* , the data are then converted into continuous measured data at time *t* such that $y_j = x(t_j)+\varepsilon$. Therefore, over the interval [1, *n*] there exist a continuous value of PM10 level. The function $x(t_j)$ is estimated using B-spline basis with *K* number of appropriate basis and for any day (*i*) where i=1,...,*N* the curve is defined as follows:

$$x_i(t)=\sum_{k=1}^{K} c_{i,k}\varphi_k(t) \qquad (1)$$

After data conversion is completed, now a set of *N* daily PM10 curves denoted as matrix $\mathbf{X}$ are available. Functional Principal Component Analysis (FPCA) is used to compute the daily indices. FPCA is an important functional data analysis (FDA) method to explore the mode of variation in the curves data. To prevent the influence of outliers on the FPCA model, any outlying curves from the present data set are preliminary removed. The next step is to randomly select the remaining data set to allow for independently distributed data without the influence of time. The FPCA model is performed to compute the daily indices IND using the Hotelling T^2 statistics to represent the condition of any curve *i* such that

$$\text{IND}_i=\frac{(s_{i1})^2}{\lambda_1}+\ldots+\frac{(s_{ip})^2}{\lambda_p} \qquad (2)$$

where λ is the sample variance of the principal component (PC) and s_{ip} is the *p*th principal component score. The upper and lower control chart limit for the indices can be computed as $\text{UCL}=\chi^2_{\alpha,p}$ and $\text{LCL}=0$ where $\chi^2_{\alpha,p}$ is the $100(1-\alpha)$ percentile of the Chi-squared distribution with *p* degrees of freedom [20]. In this control chart only UCL is used since IND are positive values. The daily PM10 process is considered stable or normal if $\text{IND}\le\text{UCL}$. Using FPCA, the variation of curves is summarized into a smaller number of independent components (the PC) without the loss of the original information. The score for each PC is given as follows:

$$S_{ip}=\int_0^T \xi_p(t)z_i(t)dt \qquad (3)$$

The term $\xi_p(t)$ refers to the pth eigenfunction. A set of *m* required eigenfunction with *m*=1,..,*p* and the component variance λ are obtained by solving the eigen equation $\int_0^T v(r,t)\xi(t)dt=\lambda\xi(r)$ with respect to the constraints $\int_0^T \xi_m{}^2(t)dt=1$ and $\int_0^T \xi_{m-1}(t)\xi_m(t)dt=0$. The term $v(r,t)$ is the corresponding variance-covariance function for the sample of mean centered curves $z_i(t)$ where $z(t)=x(t)-\bar{x}(t)$. FPCA computation is based on the matrix approach by means of basis expansion approach [14]. Using the UCL, the IND indices obtained are plotted on the control chart. The day with indices which lie above the UCL is removed from the data set. Again a new set of indices is obtained for the remaining data and the indices are plotted onto the control chart and once again days with points above the UCL are removed. The same procedure is repeated until no more outlying points are detected which produces a set of 'in control' data that represent the stable state of process. Hotelling T^2 statistics assumes that data are multivariate normal; therefore, in conducting the phase 2 analysis, the multivariate normality assumption is required.

For phase 2 analysis, if the multivariate normality assumption is satisfied, given a set of future data $\mathbf{X}_{new}$, $\mathbf{X}_0$ as the in control data set, the in control component variance $\boldsymbol{\lambda}_0$ and the in control eigenfunction $\boldsymbol{\xi}_0$, the new indices for series of future curves to be monitored can be predicted as follows:

$$\text{IND}_{i.new}=\frac{(s_{i1.new})^2}{\lambda_{o1}}+\ldots+\frac{(s_{ip.new})^2}{\lambda_{op}} \qquad (4)$$

The component score for the new data set is estimated using $S_{ip.new}=\int_0^T \xi_{o.p}(t)z_{i.new}(t)dt$ where $z_{new}=x_{new}(t)-\bar{x}_o(t)$ are the functional data objects. Now, considering the autocorrelation issue in the phase 2 control chart, the autocorrelation effect is then proposed to be reduced from the new indices using the data pre-whitening approach.

B. *Method to Overcome Autocorrelation Effect*

Two commonly used methods to overcome the autocorrelation effect are identified. The first one is the application of time series model where the residual control chart is used [19, 22, 24, 25]. The second approach suggests the modification of the control limits [23]. The resulting control charts are referred to as the modified control chart.

In this study, a modified version control chart is also used to reduce the problem of the well-known AR(1) problem. Unlike the previous approach, this time the pollutant indices are adjusted by removing the AR(1) effect from the data (indices) while still maintaining the existing control limit. Data pre-whitening approach is proposed to be used in order to reduce the autocorrelation effect before the monitoring takes place. In this context, data pre-whitening is defined as removing the AR(1) component from the series of the indices (IND $_{new)}$ using the following formula:

$$\mathrm{RIND}_{\mathrm{new}(t)} = \mathrm{IND}_{\mathrm{new}(t)} - \rho \mathrm{IND}_{\mathrm{new}(t-1)} \qquad (5)$$

where $\mathrm{RIND}_{\mathrm{new}(t)}$ is the new indices after the autocorrelation effect is removed and $\hat{\rho}$ belongs to the lag-g serial correlation coefficient whereby in this study, for AR(1), g is equivalent to 1. The general formula for $\hat{\rho}$ is given as follows:

$$\hat{\rho} = \frac{\frac{1}{N-g}\sum_{t=1}^{N-g}\left[\mathrm{IND}_{\mathrm{new}(t)} - \mathrm{IN}\bar{\mathrm{D}}_{\mathrm{new}}\right]\left[\mathrm{IND}_{\mathrm{new}(t+g)} - \mathrm{IN}\bar{\mathrm{D}}_{\mathrm{new}}\right]}{\frac{1}{N}\sum_{t=1}^{N}\left[\mathrm{IND}_{\mathrm{new}(t)} - \mathrm{IN}\bar{\mathrm{D}}_{\mathrm{new}}\right]^2} \qquad (6)$$

For large sample size *N*, the asymptotic distribution for the sample AR(1) autocorrelation is approximately normally distributed with mean zero and variance 1/*N* [27]. To determine whether a serial correlation AR(1) exists or not, the $\hat{\rho}$ will be investigated at $\alpha = 0.05$ using the estimation such that $-2/\sqrt{N} \le \rho \le 2/\sqrt{N}$; the serial correlation significantly exists if the estimated $\hat{\rho}$ lies outside the interval values.

III. Application

A. *Data and Description*

To examine the capability of the data pre-whitening approach in reducing the effect of serial correlation in the performance of the control chart, a data set from Shah Alam air quality monitoring station which is located in the Selangor state of Malaysia Peninsular is utilized. The site is nearby a residential and industrial area and is situated in the Klang Valley Region. The data consist of daily hourly recorded data from the years 2004 to 2010. After data cleaning has been done, the data are then divided into two sets. Data within the years (2004-2008) are used for the phase 1 analysis and the years (2009-2010) are considered as monitoring data at phase 2. Before the phase 1 and phase 2 are conducted, the data are first converted into curves form using 17 basis functions following (1). The analysis computation is conducted using the free R-software [28].

B. *Results of Analysis*

At phase 1 analysis, the 'robMah' method [26] was used to preliminary remove 106 functional outliers from the data set that consist 1827 daily curves. Next, a sample of 1500 curves was randomly selected from the remaining (1721) curves in order to allow for independency of data. It was decided that about four numbers of retained principal component would be used in the FPCA model and thus the iterative phase 1 control chart was conducted using the 1500 randomly selected samples until a stable state or 'in control' process was obtained. In this case it was achieved after six iterations. The phase 1 analysis results are summarized in Table I.

TABLE I. PHASE 1 RESULTS OF FPCA

Iteration	**PC variation**					**UCL**
	PC1	PC2	PC3	PC4	Total	
6	0.57	0.12	0.07	0.05	0.81	13.28

The reference data contained data of the stable process and was produced based on historical information of a large size [21], thus supporting the sufficiency of the data in the control chart usage. The 99% quartile of the chi-squared distribution with four degree freedom ($\chi^2_{0.01,4}$) gave the final upper control limit (UCL) equivalent to 13.28. Using FPCA by four number of PC, about 81% of the total variation explained in the data was considered in the formation of the index IND (i.e. representing the condition of daily PM10 curves as compared to the average process). The most dominant PC1 explained more than half of the total variation while 12%, 7% and 5% of other variation came from PC2, PC3 and PC4 respectively.

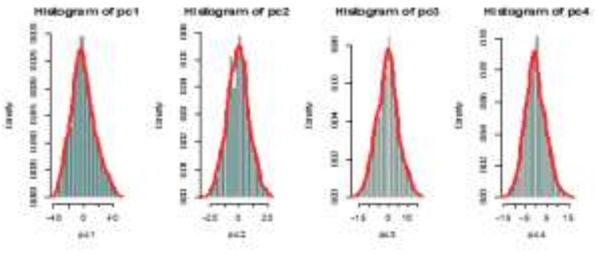

Figure 1. The histogram of PCs score

The histogram of all the PCs depicted in Fig. 1 showed that multivariate normality assumption was satisfactory. The results indicated that the control limit constructed using the Hotelling T^2 in phase 1 could be used in phase 2. Otherwise the non-parametric or bootstrapping approach was the alternative in order to establish the new control limit in phase 2.

The results of phase 2 analysis are shown and tabulated in Table II. The indices for daily PM10 curves for data set (2009-2010), the IND $_{new}$ were computed and were further tested to identify whether AR(1) existed using the confidence interval in (6). It was found that the lag-1 serial autocorrelation was significant with the value of the estimated correlation coefficient $\hat{\rho}$ equals 0.368. When the indices were plotted onto the control chart, about 5.07% of abnormal days (signal rate) was detected.

TABLE II. PHASE 2 RESULTS: BEFORE AND AFTER DATA PRE-WHITENING

Before data pre-whitening			
$\hat{\rho}$ value	CI of ρ	AR(1) status	Signal rate
0.364	(-0.074, 0.074)	Significant	**5.07**
After data pre-whitening			
$\hat{\rho}$ value	CI of ρ	AR(1) status	Signal rate
-0.057	(-0.074, 0.074)	Insignificant	**3.15**

a. CI-confidence interval

In dealing with the existence of AR(1), the data pre-whitening approach was then applied to the IND $_{new}$ giving new series of removed AR(1) effect known as the RIND. The RIND was then plotted on the same control chart. This time it showed that the signal rate was reduced. About 3.15% days that were considered abnormal from the average process was identified. The confidence interval of $\hat{\rho}$ has shown that the new estimated value lie within the interval limit. Thus, after data pre-whitening, the indices were shown to be free of the AR(1) effect. The results of Table 2 indicated that the data pre-whitening approach was capable in handling AR(1) autocorrelation problem. This was evidenced by the increase in the performance of the control chart where the percentage of the false alarm rate was reduced upon the application. The graphical representations of the control chart before and after data pre-whitening are shown in Fig. 2.

The control chart in Fig. 2 shows the trend of the daily PM10 where the indices represented the variation of PM10 levels over the entire day of the PM10 process (the curve data). The behavior and trend of the daily PM10 curve could be directly visualized. Over the two year monitoring period, the control chart showed that air quality of PM10 has improved in 2010 as compared to in 2009. Both years experienced a higher number of signals during the middle of the year in the months of June to September but with more frequent and higher indices during the year 2009. During this period, Peninsular Malaysia was experiencing the dry season which fell under the south west monsoon. After data pre-whitening was conducted, the number of detected signal was reduced. This was indicated by the reduction in the number indices above the UCL. The shape of the identified abnormal indices was slightly changed. As for example, before pre-whitening there were two days with similar extreme levels of abnormality in 2009 but after data pre-whitening, only one day was recognized as the most extreme case. The other day was actually false in the behavior before the AR(1) effect was removed. This was due to the masking impact of serial correlation.

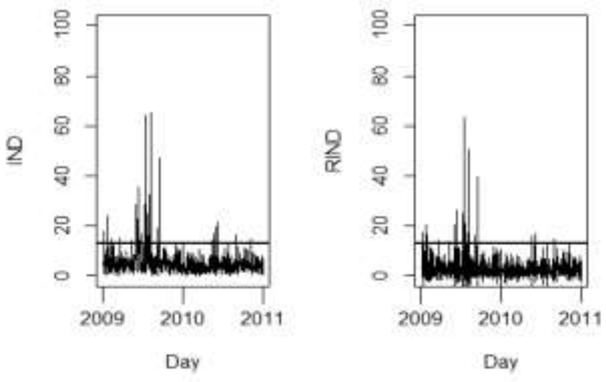

Figure 2. The control chart before and after data pre-whitening

IV. Conclusion

This paper proposed the application of data pre-whitening approach to enhance the performance of control chart for monitoring the daily PM10 indices when the indices were correlated over time. In this study, the indices were the representation of the daily PM10 curves overtime. The FPCA model was used in the control chart establishment to develop the daily indices by means of Hotelling T^2 statistics given named as IND. The phase 1 control chart via an iterative procedure resulted into a set of 'in control' data that were subsequently to be used for the phase 2 monitoring analysis. Data randomization was applied at the phase 1 analysis while data pre-whitening approach was used to remove the AR(1) effect. The aim was to handle data independency when monitoring series of future PM10 indices at phase 2 analysis. By examining the value of the AR(1) coefficient using the 95% confidence interval, the study has shown that the AR(1) effect was reduced. Consequently, using the new daily indices RIND after the AR(1) effect was removed, the control chart resulted into less numbers of detected abnormal days (signal rate) compared to using the IND indices. In conclusion, the data pre-whitening approach was capable in handling the AR(1) autocorrelation problem. This was evidenced by the increase in the performance of the control chart where the percentage of false alarm rate was reduced upon the application.

Acknowledgment

The authors would like to thank the Department of Environment Malaysia and Prof. Dr. Mohd Talib Latif from University 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

[1] C.N. Madu, Managing Green Technologies for Global Competitiveness. Quorum Books, Westport CT, 1996.

[2] L.C. Angell, R.D. Klassen, "Integrating environmental issues into the mainstream: an agenda for research in operations management," Journal of Operations Management, 17, pp. 575-598, 1999.

[3] C.J. Corbett, J. Pan, "Evaluating environmental performance using statistical process control techniques," European Journal of Operation Research, 139, pp. 68-83, 2002.

[4] C.I. Davidson, R.F. Phalen, P.A. Solomon, "Airbone particulate matter and human health: a review," Aerosol Science and Technology, 39, pp.737-749, 2005.

[5] G. Polichetti, S. Cocco, A.Spinali, V. Trimarco, A. Nunziata, "Effects of particulate matter (PM10, PM2.5, and PM1) on the cardiovascular system, " Toxicology, 261, pp. 1-8, 2009.

[6] A. Herman, "Ambient air pollution and adverse health effects," Procedia Social and Behavioral Sciences, 2, pp. 7333-7338, 2010.

[7] M. Adam, D. Felber-Dietrich, E. Schaffner, D. Carballo, J.C. Barthelemy, J.M. Gaspoz, M.Y. Tsai, R. Rapp, H.C. Phuleria, C. Schindler, J. Schwartz, N. Kunzli, N.M. Probst-Hensch, "Long-term exposure to traffic-related PM10 and decreased heart rate variability: Is the association restricted to subjects taking ACE inhibitors?," Environment International, 48, pp. 9-16. 2012.

[8] W.R.W. Mahiyuddin, M. Sahani, R. Aripin, M.T. Latif, T.Q. Thach, C.M. Wong, "Short-term effects of daily air pollution on mortality," Atmospheric Environment, 65, pp. 69-79, 2013.

[9] L.W. Morrison, "The use of control chart to interpret environmental monitoring data, "Natural Areas Journal, 28, 1, pp. 66-73, 2008.

[10] D. Maurer, M. Mengel, G. Robertson, T. Gerlinger, A. Lissner, "Statistical process control in sediment pollutant analysis," Environmental Pollution, 104, pp. 21-29,1999.

[11] C.V. Palau, F. Arregui, A. Ferrer, "Using multivariate principal component analysis of injected water flows to detect anomalous behaviours in a water supply system. A case study," Water Science and Technology: Water Supply, 4,3, pp. 169-181, 2004.

[12] C.K. Yoo, K. Villez, S.W.H. Van Hulle, P.A. Vanrolleghem, "Enhanced process monitoring for wastewater treatment systems," Environmetrics, 19, 6, pp. 602-617, 2007.

[13] M.J. Anderson, A.A. Thompson, "Multivariate control charts for ecological and environmental monitoring," Ecological Applications, 14, 6, pp. 1921-1935, 2004.

[14] J.O. Ramsay, B.W. Silverman, Functional data analysis, 2nd Edition, Springer, New York, 2006.

[15] N. Shaadan, A.A. Jemain, S.M Deni, "The construction of control chart for PM10 functional data," *Proceedings of The 3rd International conference on Mathematical Sciences*, 2013. (to be published).

[16] B.M. Colosimo, M. Pacella, "A comparison study of control chart for statistical monitoring of functional data," International Journal of Production Research, 48, 6, 1575-1601, 2010.

[17] J.J.H. Shiau, Y.C. Hsu, "Robustness of the EWMA control chart to non-normality for autocorrelated processes," Quality Technology & Quantitative Management, 2, 2 pp. 125-146, 2005.

[18] T.J. Harris, W.H. Ross,"Statistical process control procedures for correlated observations," The Canadian Journal of Chemical Engineering, 69, pp. 48-57, 1991.

[19] S. Elevli, N. Uzgoren, M. Savas, "Control chart for autocorrelated colemanite data," Journal of Scientific and Industrial Research, 68, pp. 11-17, 2009.

[20] J.E. Jackson, A user guide to Principal Component, Wiley, New York, 2003.

[21] D.C. Montgomery, Statistical quality control: a modern introduction, 6th edition, John Wiley, New York, 2009.

[22] C.W. Lu, M.R. Jr. Reynolds, "Control charts for monitoring the mean and variance of autocorrelated processes," Journal of Quality Technology, 31, 3, 259-274, 1999.

[23] A.V. Vasilopoulos, A.P. Stamboulis, "Modification of control limits in the presence of data correlation," Communication in Statistics - Simulation and Computation, 26, pp. 979-1008, 1978.

[24] J.N. Pan, S.T.Chen," Monitoring long-memory air quality data using ARFIMA model,"Environmetrics, 19, pp.209-219, 2008.

[25] J.C. Garcia-Diaz, "Monitoring and forecasting nitrate concentration in the groundwater using statistical process control and time series analysis: a cased study," Stochastic Environmental Research and Risk Assessment, 25, pp. 331-339, 2011.

[26] J.Z. Huang, H. Shen, "Functional coefficient regression models for non-linear time series:a polynomial spline approach," Scandinavian Journal of Statistics, 31, pp. 515-534, 2004.

[27] C. Chatfield, The analysis of time series, an introduction, 6th edition, Chapman & Hall/ CRC, New York, 2004.

[28] R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

About Author (s):

Norshahida Shaadan - currently a senior lecturer at UiTM Shah Alam Malaysia. Her research activities focused on performance index, environmental and computational statistics.

Abdul Aziz Jemain - a professor of statistics at UKM Bangi Malaysia -PhD holder from the University of Reading, U.K. in 1989 - research area: social, medical and environmental statistics, modelling and multi criteria decision making.

Sayang Mohd Deni - an associate professor of statistics at UiTM Shah Alam Malaysia - PhD holder in rainfall modeling-an active researcher with several reputable publications - research areas: rainfall modeling, environmental and computational statistics.