ESTEEM Academic Journal
Vol. 11, Special Issue. 1, March 2014, 135-143

ANALYSIS OF PM_{10} USING EXTREME VALUE THEORY

Hasfazilah Ahmat[1], Ahmad Shukri Yahaya[2], Nor Azam Ramli[3], Ahmad Zia ul-Saufie Mohamad Japeri[4] and Hazrul Abdul Hamid[5]

[2,3] Clean Air Research (CARE) Group, School of Civil Engineering, Engineering Campus, Universiti Sains Malaysia, 14300 P.Pinang, MALAYSIA
[1,4] Department of Computer and Mathematical Sciences, Universiti Teknologi MARA Pulau Pinang, 13500 Permatang Pauh, P.Pinang, MALAYSIA
[5]Department of Mathematics, Kolej Matrikulasi Pulau Pinang, Pongsu Seribu, 13200 Kepala Batas, P.Pinang, MALAYSIA
[1]hasfazilah.ahmat@gmail.com; [2]shukri@eng.usm.my; [3]azam@eng.usm.my; [4]ziaulsaufie@gmail.com; [5]hazrul@kmpp.matrik.edu.my

ABSTRACT

The literature review had identified that the extreme value theory is widely used in hydrological studies. However, its contribution in air pollution is indisputably important. This paper assesses the use of extreme value distributions of the two-parameter Gumbel, two and three-parameter Weibull, Generalized Extreme Value (GEV) and two and three-parameter Generalized Pareto Distribution (GPD) on the maximum concentration of daily PM_{10} data recorded in the year 2005 in Shah Alam, Selangor. Parameters estimations for all distributions were evaluated using the method of Maximum Likelihood Estimator (MLE). The goodness-of-fit of the distribution was determined using six performance indicators namely; the accuracy measures which include Predictive Accuracy (PA), Coefficient of Determination (R^2), Index of Agreement (IA) and error measures that consist of Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and Normalized Absolute Error (NAE). The best distribution was selected based on the highest accuracy measures and the smallest error measures. This study reveals that the three-parameter Weibull was the best fit for daily maximum concentration for PM_{10}. The analysis also demonstrates that the number of days in which the concentration of PM_{10} exceeded the Malaysia Ambient Air Quality Guidelines (MAAQG) of 150 mg/m^3 for 2005 was 25 days as compared to the actual 15 days.

Keywords: Air pollution; Extreme Value Theory (EVT); PM_{10}; Prediction; Gumbel; Weibull; Generalized Extreme Value (GEV).

1. INTRODUCTION

Extreme value theory is among the most important statistics that has been used in the applied sciences and many other fields for several decades. The main feature of this analysis is to

p-ISSN 1675-7939; e-ISSN 2289-4934

quantify unusual or infrequent events (extreme) as the minimum or maximum concentration, high concentration or frequency data (Coles, 2001). Various studies on the application of extreme value theory in different fields have been published. For example in the field of risk management operations (Yao, Wen and Luan , 2013), exposure to volatile organic compounds (Su, Jia and Batterman, 2012), future market (Kao and Lin, 2010), capital requirements (Tsai and Chen, 2011), wind speed (Reynolds, 2012; Torrielli, Repetto and Solari, 2013), wave heights (Petrov, Guedes Soares and Gotovac, 2013) and storm (Reeve, Randell, Ewans and Jonathan, 2012).

Though the extreme value theory is widely used in hydrological studies, its contribution in air pollution is indisputably important. The extreme value distribution is a widely used method for assessing and estimating the concentrations of air pollution (Dasgupta and Bhaumik, 1995; Horowitz and Barakat, 1979; Hurairah, Ibrahim, Daud and Haron, 2005; Kuchenhoff and Thamerus, 1995; Lu, 2002; Quintela-del-Río and Francisco-Fernández, 2011; Reyes, Vaquera and Villasenor, 2010; Roberts, 1979; Smith, 1989; Surman, Bodero and Simpson, 1987). Researches which apply the extreme value theory on particulate matters with the aerodynamic diameter of less than 10 micrometers (mm) (PM_{10}) are (Md Yusof, Ramli and Yahaya, 2011) and (Sharma et al., 2012).

2. METHODOLOGY

2.1 Family theory of extreme values and model

This research undertakes the analysis of PM_{10} data using the extreme value distributions, namely: Gumbel, two and three-parameter Weibull, Generalized Extreme Value (GEV) and two and three-parameter Generalized Pareto Distribution (GPD). All the parameters of the distributions are estimated using the method of maximum likelihood estimators (MLE).

2.2 Performance indicators

This study uses six performance indicators to select the best distribution to represent the data. The accuracy measures are the prediction accuracy (PA), coefficient of determination (R^2) and Index of Accuracy (IA). The accuracy value is between 0 and 1 and as the value approaches 1, the model is appropriate. On the other hand, as the value of error measures approaching 0, the model is deemed to be the best model. The error measures used in this study are the root mean squared error (RMSE), the normalized absolute error (NAE) and the mean absolute error (MAE). The accuracy measures are dimensionless that is independent of the unit of data while the error measures are scale and unit-dependant (Ji and Gallo, 2006).

2.3 Data

The daily maximum data of PM_{10} from January to December 2005 was furnished by the Department of Environment, Malaysia. The data was collected through a continuous monitoring by Alam Sekitar Sdn. Bhd. (ASMA) from the air monitoring station located at Sek. Keb. Raja Muda, Shah Alam (coordinate : N03°04.636, E101°30.673) in the central of Peninsular Malaysia. Azmi, Latif, Ismail, Juneng and Jemain (2010) reported that a high density of vehicles contributes to high concentrations of PM_{10} in Shah Alam. Furthermore,

p-ISSN 1675-7939; e-ISSN 2289-4934

the Klang Valley area is continuously exposed to the problem of air quality due to its geographical location and rapid population growth, industrial as well as commercial activities. On top of that, the pollutants in the area are stagnant due to stable atmospheric conditions resulting from the presence of weak winds in the area (Department of Environment Malaysia, 2006).

The analysis of the data is completed using a high-level language and interactive environment for numerical computation, visualization, and programming package for engineers called MATLAB® (Chapman, 2004).

3. RESULT AND DISCUSSION

The descriptive statistics of the data is depicted in Table 1. The average concentrations of PM_{10} recorded was 74.3 μg/m^3. The number is below the Malaysia Ambient Air Quality Guidelines (MAAQG) for the daily average of 150 μg/m^3 (Department of Environment Malaysia, 2014). The PM_{10} concentrations were skewed to the right (skewness = 5.743) as illustrated in Figure 1 and hence indicates that there were extreme concentrations of PM_{10} exist during the year. The maximum concentration was 587 μg/m^3 while the lowest concentration was recorded at 17 μg/m^3.

Table 1 : Descriptive statistics of PM_{10} (μg/m^3)

N		365
Mean		74.32
Median		62.00
Std. Deviation		57.35
Variance		3288.51
Skewness		5.74
Kurtosis		44.81
Minimum		18.00
Maximum		587.00
Percentiles	50	62.00
	75	84.00
	95	139.40

Figure 1 demonstrates time series plot and histogram of daily maximum data of PM_{10} in 2005. The number of days where concentrations exceeding Malaysia Ambient Air Quality Guidelines (MAAQG) of 150 μg/m^3 in 2005 were 15 days. The concentration peaked in the month of August.

p-ISSN 1675-7939; e-ISSN 2289-4934

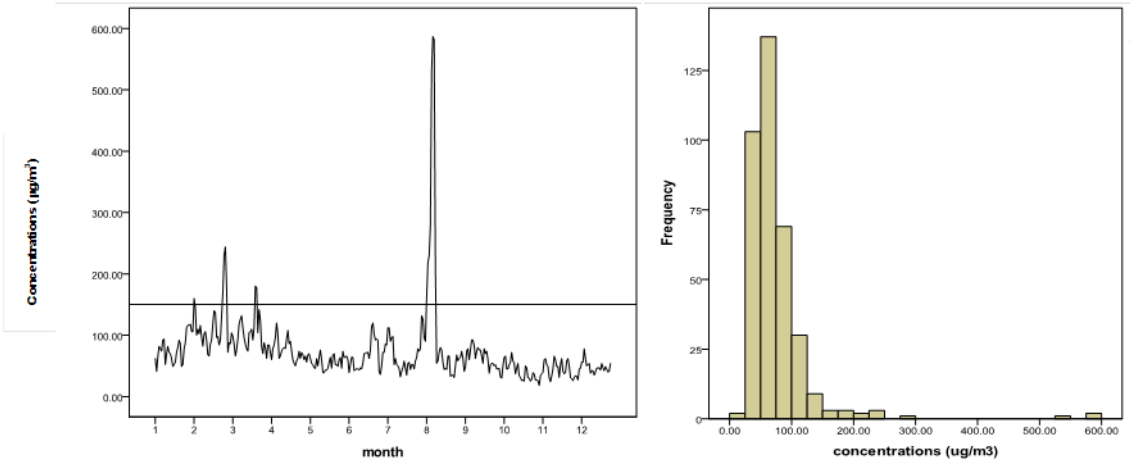

Figure 1: Time series plot and histogram of daily maximum of PM_{10}

The records of PM_{10} as illustrated in Figure 2 shows that the concentrations of PM_{10} were higher in certain months. The concentrations were particularly higher in the month of August as compared to the other months due to trans boundary sources of smoke from forest fires in Sumatera province of Indonesia which affected the central, eastern and northern parts of Peninsular Malaysia and particularly in the Klang Valley area (Department of Environment Malaysia, 2006). The following month recorded low concentrations as the wind had blown all the particulates further northwards.

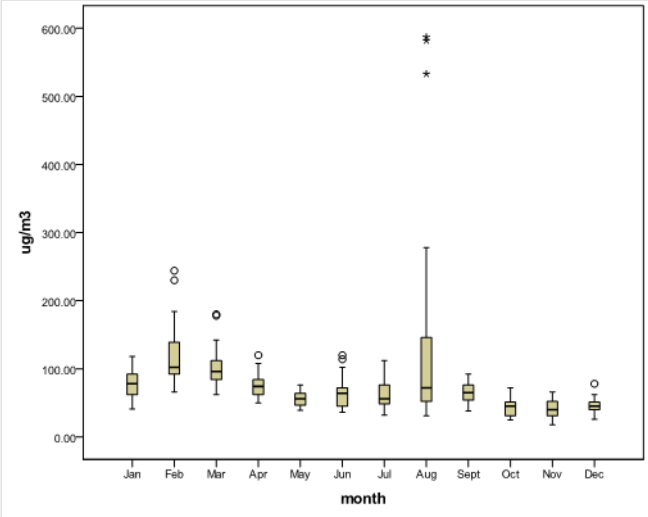

Figure 2 : Boxplot

Table 2 lists the values for the location parameter, μ, scale parameter, σ and shape parameter, λ for all distributions using the method of MLE.

p-ISSN 1675-7939; e-ISSN 2289-4934

Table 2 : Parameters estimation

Distributions	Parameters		
	μ	σ	λ
2-Gumbel	132.8827	113.9164	-
2-Weibull	-	83.6210	1.5706
3-Weibull	17.9698	61.5549	1.2860
GEV	53.5461	22.2272	0.2393
2-GPD	-	78.8026	-0.0666
3-GPD	18.0000	78.8026	-0.0666

Based on performance indicators in Table 3, the distributions were ranked as depicted in Table 4. GEV distribution recorded the lowest NAE and MAE. Two-parameter GPD on the other hand, recorded the highest in PA and R^2. Three parameter Weibull recorded the lowest RMSE and highest in IA. The best distribution was selected based on the highest accuracy measures and the smallest error measures as shown in Table 4.

Table 3 : Performance Indicators

Distributions	Performance Indicators					
	NAE	PA	R^2	RMSE	IA	MAE
2-Gumbel	1.2262	0.5931	0.3498	132.3609	0.5522	91.1317
2-Weibull	0.2153	0.8531	0.7238	30.0200	0.9173	16.0030
3-Weibull	0.1729	0.9184	0.8388	23.9527	0.9514	12.8538
GEV	0.0802	0.8201	0.6689	47.5452	0.8713	5.9622
2-GPD	0.4716	0.9237	0.8486	61.9631	0.8581	35.0472
3-GPD	0.3482	0.8967	0.7997	38.6023	0.9108	25.8772

Table 4 : Ranking of performance

Distributions	Ranking						Total	Overall Ranking
	NAE	PA	R^2	RMSE	IA	MAE		
2-Gumbel	6	6	6	6	6	6	36	5
2-Weibull	3	4	4	2	2	3	18	2
3-Weibull	2	2	2	1	1	2	10	1
GEV	1	5	5	4	4	1	20	3
2-GPD	5	1	1	5	5	5	22	4
3-GPD	4	3	3	3	3	4	20	3

Based on the ranking, the three-parameter Weibull was chosen to be the best distribution for daily maximum concentration for PM_{10} since it had better performance in accuracy and error measures than the other distributions.

p-ISSN 1675-7939; e-ISSN 2289-4934

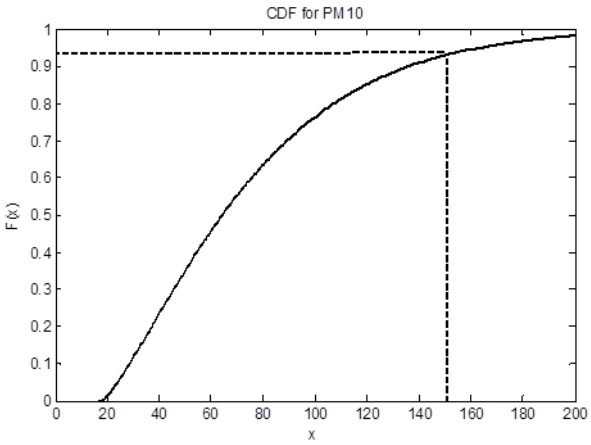

Figure 3: Cumulative distribution function of three parameter Weibull

Cumulative distribution function (CDF) of the three-parameter Weibull is presented in Figure 3. From this figure, the probability of the concentrations exceeding the levels of MAAQG of 150 μg/m^3 was estimated at 0.0694 (F (x) <150 = 0.9306). Hence, the estimated number of days in which PM_{10} concentrations exceed MAAQG was 0.0694 x 365 days = 25 days as compared to actual observed records of 15 days.

4. CONCLUSION

This paper discussed the probability and the number of days of the extreme concentrations which exceeded the permissible value of PM_{10} concentrations of 150 μg/m^3 in Shah Alam. The MLE was used to estimate the parameters for the distributions. All the daily maximum data in 2005 was used to analyse the efficiency of the extreme value distributions. The analyses of three accuracy measures, namely PA, R^2 and IA and three error measures – NAE, RMSE and MAE were acquired to indicate the efficiency or the performance indicators of the distributions.

From the findings, the average reading of the PM_{10} concentrations in Shah Alam was well below the stipulated MAAQG for the daily average of 150 μg/m^3. The highest concentration recorded in 2005 was due to trans-boundary smoke from forest fires in Sumatera which was transported by South westerly winds. The central region of Peninsular Malaysia was the most affected by the unfavourable weather conditions of hot and dry periods caused by the South westerly winds.

The three parameter Weibull gave the best estimators for Shah Alam monitoring station using the rank of accuracy and error measures. From the plots, the probabilities of the concentrations exceeded the levels of MAAQG of 150 μg/m^3 were estimated and the predicted number of days was calculated. The estimated number of days for Shah Alam were 25 days as compared to actual observed records of 15 days.

To conclude, the three parameter Weibull could be the most appropriate distribution to the PM_{10} concentrations analysis to effectively predict the exceedances of future extreme

p-ISSN 1675-7939; e-ISSN 2289-4934

concentrations of PM_{10}. As a result, it may help the policy makers in the respective field to plan suitable measures to curb the occurrence of PM_{10} extreme concentrations and eventually may reduce the effects on human health

REFERENCES

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 & Health*, *3*, 53–64. doi:10.1007/s11869-009-0051-1

Chapman, S. (2004). *MATLAB programming for engineers* (Third Edit.). Australia: Thomson.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Bristol: Springer series in statistics.

Dasgupta, R., & Bhaumik, D. K. (1995). Upper and lower tolerance limits of atmospheric ozone level and extreme value distribution. *Sankhya : The Indian Journal of Statistics*, *57*(B(2)), 182–199.

Department of Environment Malaysia. (2006). *Malaysia environmental quality report 2005*.

Department of Environment Malaysia. (2014). *Malaysia Environmental Quality Report 2013*. Kuala Lumpur.

Horowitz, J., & Barakat, S. (1979). Statistical analysis of the maximum concentration of an air pollutant: Effects of autocorrelation and non-stationarity. *Atmospheric Environment (1967)*, *13*(6), 811–818. doi:10.1016/0004-6981(79)90272-5

Hurairah, A., Ibrahim, N. A., Daud, I. Bin, & Haron, K. (2005). An application of a new extreme value distribution to air pollution data. *Management of Environmental Quality: An International Journal*, *16*(1), 17–25. doi:10.1108/14777830510574317

Ji, L., & Gallo, K. (2006). An Agreement Coefficient for Image Comparison. *Photogrammetric Engineering & Remote Sensing*, *72*(7), 823–833.

Kao, T., & Lin, C. (2010). Setting margin levels in futures markets: An extreme value method. *Nonlinear Analysis: Real World Applications*, *11*, 1704–1713. doi:10.1016/j.nonrwa.2009.03.025

Kotz, S., & Nadarajah, S. (2000). *Extreme-Value Distributions : Theory and Applications*. London: Imperial College Press.

Kuchenhoff, H., & Thamerus, M. (1995). Extreme value analysis of Munich air pollution data. *Sonderforschungsbereich*, *386*(4), 1–24. Retrieved from http://epub.ub.uni-muenchen.de/

p-ISSN 1675-7939; e-ISSN 2289-4934

Lu, H. (2002). The statistical characters of PM 10 concentration in Taiwan area. *Atmospheric Environment*, *36*, 491–502.

Md Yusof, N. F. F., Ramli, N. A., & Yahaya, A. S. (2011). Extreme Value Distribution for Prediction of Future PM 10 Exceedences. *International Journal of Environmental Protection*, *1*(4), 28–36.

Petrov, V., Guedes Soares, C., & Gotovac, H. (2013). Prediction of extreme significant wave heights using maximum entropy. *Coastal Engineering*, *74*, 1–10. doi:10.1016/j.coastaleng.2012.11.009

Quintela-del-Río, A., & Francisco-Fernández, M. (2011). Nonparametric functional data estimation applied to ozone data: prediction and extreme value analysis. *Chemosphere*, *82*, 800–808. doi:10.1016/j.chemosphere.2010.11.025

Reeve, D. T., Randell, D., Ewans, K. C., & Jonathan, P. (2012). Uncertainty due to choice of measurement scale in extreme value modelling of North Sea storm severity. *Ocean Engineering*, *53*, 164–176. doi:10.1016/j.oceaneng.2012.07.001

Reyes, H. J., Vaquera, H., & Villasenor, J. A. (2010). Estimation of trends in high urban ozone levels using the quantiles of (GEV). *Environmetrics*, *21*, 470–481. doi:10.1002/env

Reynolds, A. M. (2012). Gusts within plant canopies are extreme value processes. *Physica A: Statistical Mechanics and Its Applications*, *391*, 5059–5063. doi:10.1016/j.physa.2012.05.062

Roberts, E. M. (1979). Review of Statistics of Extreme Values with Applications to Air Quality Data. Part II. Applications. *Journal of the Air Pollution Control Association*, *29*(7), 733–740. doi:10.1080/00022470.1979.10470856

Sharma, P., Chandra, A., Kaushik, S. C., Sharma, P., & Jain, S. (2012). Predicting violations of national ambient air quality standards using extreme value theory for Delhi city. *Atmospheric Pollution Research*, *3*, 170–179. doi:10.5094/APR.2012.017

Smith, R. L. (1989). Extreme Value Analysis of Environmental Time Series : An Application to Trend Detection in Ground-Level Ozone. *Statistical Sciences*, *4*(4), 367–393.

Su, F.-C., Jia, C., & Batterman, S. (2012). Extreme value analyses of VOC exposures and risks: A comparison of RIOPA and NHANES datasets. *Atmospheric Environment*, *62*, 97–106. doi:10.1016/j.atmosenv.2012.06.038

Surman, P. G., Bodero, J., & Simpson, R. W. (1987). The Prediction of the Numbers of Violations of Standards and the Frequency of Air Pollution Episodes using Extreme Value Theory. *Atmospheric Environment*, *21*(8), 1843–1848.

Torrielli, A., Repetto, M. P., & Solari, G. (2013). Extreme wind speeds from long-term synthetic records. *Journal of Wind Engineering and Industrial Aerodynamics*, *115*, 22–38. doi:10.1016/j.jweia.2012.12.008

p-ISSN 1675-7939; e-ISSN 2289-4934

Tsai, M.-S., & Chen, L.-C. (2011). The calculation of capital requirement using Extreme Value Theory. *Economic Modelling*, *28*, 390–395. doi:10.1016/j.econmod.2010.08.010

Yao, F., Wen, H., & Luan, J. (2013). CVaR measurement and operational risk management in commercial banks according to the peak value method of extreme value theory. *Mathematical and Computer Modelling*, *58*(1-2), 15–27. doi:10.1016/j.mcm.2012.07.013

p-ISSN 1675-7939; e-ISSN 2289-4934