Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 30 (2015) 309 – 314

Procedia Environmental Sciences

International Conference on Environmental Forensics 2015 (iENFORCE2015)

Bayesian Extreme for modeling high PM_{10} concentration in Johor

Nor Azrita Mohd Amin[a,b]*, Mohd Bakri Adam[b], Ahmad Zaharin Aris[c]

[a]*Institute of Engineering Mathematics, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600 Arau, Malaysia*
[b]*Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Malaysia*
[c]*Environmental Forensics Research Centre, Faculty of Environmental Studies, Universiti Putra Malaysia, 43400 UPM Serdang, Malaysia*

Abstract

The aim of this study is to determine the behavior of extreme PM_{10} levels monitored at three air monitoring stations in Johor using frequentist and Bayesian technique. Bayesian allows priors or additional information about the data into the analysis which expectedly improve the model fit. The generalized extreme value distribution is fitted to the monthly maxima PM_{10} data. The results obtained show that the Bayesian posterior inferences perform at least as trustworthy as maximum likelihood estimates but considerably more flexible and informative. The return levels for 10, 50 and 100-years were computed for future prediction.

Peer-review under responsibility of organizing committee of Environmental Forensics Research Centre, Faculty of Environmental Studies, Universiti Putra Malaysia.

Keywords: extreme value theory; PM_{10}; generalized extreme value distribution; Bayesian inference; maximum likelihood estimates

1. Introduction

The field of extreme values (EV) studies is motivated by the occurrence of extreme events such as atmospheric pollutions, high rainfalls, floods and many others that arise due to physical processes and also human activities. The impacts of these extreme phenomena not only kill populations but also cause serious injuries, material damages as well as affect a country’s economic developments. The relation of these catastrophic events with statistical analysis of EV theory had been developed some decades earlier. EV theory is unlike other statistical approaches since the focus is on the tail of the distribution either on maxima or minima values. The scope of EV theory has been widely explored in various fields. Recently, it has become a vigorous research area due to its significance in many applications. Literatures on EV theory among others are by Coles [1], Behrens et al. [2] and Haan and Ferreira [3].

* Corresponding author. Tel.: +6019-7768386
E-mail address: norazrita@unimap.edu.my

Peer-review under responsibility of organizing committee of Environmental Forensics Research Centre, Faculty of Environmental Studies, Universiti Putra Malaysia.
doi:10.1016/j.proenv.2015.10.055

Environmental quality management is more concerned about extreme situations than average values due to its various hazardous impacts. However, most statistical methods are concerned primarily with what goes on in the center of a statistical distribution, and do not pay particular attention to the tails of the distribution. In order to control the impacts of various air pollutants, statistical models are commonly used. High PM_{10} levels have been a common problem in Malaysia especially in the dry season. During the haze periods, PM_{10} was found as the main pollutant while the other air quality parameters remained within the permissible healthy standards [4]. Study conducted by Dominick et al. [5] found that air pollution in eight selected air monitoring stations in Malaysia including Johor Bahru station based on year 2008 to 2009 are predominantly influenced by PM_{10}. Yusof et al. [6] claimed that the hourly average PM_{10} data for Seberang Perai area (industrialized area) fit to lognormal distribution for 2000, 2001, and 2002 while 2003 and 2004 data fit to Weibull distribution model. Sharma et al. [7] fitted the Gumbel distribution to make predictions of the expected number of violations on the monthly maxima for sulfur dioxide, nitrogen dioxide and suspended particulate matter data. Lu [8] revealed that the monthly maxima for PM_{10} data are well fitted to Gumbel and exponential distribution. This study focuses on extreme PM_{10} concentrations based on generalized extreme value (GEV) model.

2. Methodology

The classical approach based on maximum likelihood method is a very convenient and widely applicable method for estimation. However, there are a lot of opinions concerning more advanced and flexible alternative approaches. The currently developed approach is based on Bayesian techniques, constructed with the facility to incorporate additional information about the process which is known as a prior. Knowledge from an expert on the process may be relevant to extremal behaviors which are independent of the available data. This concept is naturally connecting with the Bayesian framework [9]. Bayesian analysis has become a standard statistical technique in recent years due to the advances of computational technologies. This chapter highlights the potential of Bayesian method in EV context based on GEV model.

2.1. Extreme value theory

A common approach of EV theory is based on the block maxima method. Block maxima refer to the maximum value of observations in a length of interval, *T*. The choice of block is often in yearly, monthly or seasonal bases which refer to one and only one observation per block. This naturally leads to independent and identically random variables. The limiting distribution of the maxima is a GEV distribution given by Equation (1),

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)^{-1/\xi}\right]\right\}. \tag{1}$$

There are three parameters in GEV distribution which are location, $-\infty < \mu < \infty$, scale parameter, $\sigma > 0$ and shape parameter, $-\infty < \xi < \infty$. The $\xi > 0$ corresponds to Frechet distribution, $\xi < 0$ corresponds to Weibull distribution and $\xi = 0$ corresponds to Gumbel distribution. Predictions of the future extreme pollutant levels are extremely important in order to make preparations to face their dangerous impacts. In EV theory, the return level, z_p is used as a value that assumes to be exceeded once every 1/p years. Return level for GEV is given by Equation (2),

$$Z_p = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1-\{-\log(1-p)\}^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma\log\{-\log(1-p)\}, & \xi = 0. \end{cases} \tag{2}$$

2.2. Bayesian inference

The explosion of interests in Bayesian methods over the last decades has been the result of the convergence of modern computing advances and the efficiency of Markov chain Monte Carlo (MCMC) algorithms for sampling from posterior distribution [10]. MCMC methods offer a great statistical tool and have been explored in diverse areas. Chib and Greenberg [11] and Gamerman and Lopes [12] provide comprehensive preliminary details as well as intensive developments and applications of MCMC. Coles and Tawn[9] illustrated how the careful elicitations of prior expert information could enhance the data information and lead to improve the prediction of extreme cases. Basically the idea of Bayesian MCMC arises when the target distribution, say $\pi(x)$ is complex such that it is difficult to sample from it directly. The simulated values from the long enough chains can be treated as dependent samples from the target distribution and used as a basis for summarizing $\pi(x)$ [13]. The posterior analysis obtained using Bayesian approach appears at least as trustworthy as maximum likelihood estimates, but considerably more flexible and informative [14]. Metropolis-Hastings (MH) routine is capable of simulating a series from an arbitrary density as a basis for summarizing features of the equilibrium distribution which is a Bayesian posterior distribution for an unknown parameter, θ.

2.3. Metropolis-Hastings algorithm

Currently there is a wide variety of MCMC algorithms developed and practiced. But it is important to understand that each idea has its own distinct advantages and drawbacks. MH method [15,16] is a very famous and most practical MCMC technique. MH is the fundamental algorithm for many MCMC approaches. The derivation of MH algorithm is discussed in Chib and Greenberg [11] by exploiting the notion of reversibility. Basically, the MH algorithm in algorithmic form for GEV model can be summarized as follows.

- Initialize the parameters, (μ^0,σ^0,ξ^0) .
- Given that the chain is currently at (μ^j,σ^j,ξ^j), draw a candidate value $\mu^{can} \sim N(\mu^j,\upsilon_\mu)$, $\sigma^{can} \sim N(\sigma^j,\upsilon_\sigma)$ and $\xi^{can} \sim N(\xi^j,\upsilon_\xi)$ for some suitably chosen variance υ_θ .
- Compute the acceptance probability of the proposed values. *p1*, *p2* and *p3* which are the acceptance probability for parameter μ , σ and ξ respectively,

$$p1=\min\left\{1,\frac{\pi\left(\mu^{can}|\sigma^j,\xi^j\right)}{\pi\left(\mu^j|\sigma^j,\xi^j\right)}\right\};\ p2=\min\left\{1,\frac{\pi\left(\sigma^{can}|\mu^{j+1},\xi^j\right)}{\pi\left(\sigma^j|\mu^{j+1},\xi^j\right)}\right\}\ ;\ p3=\min\left\{1,\frac{\pi\left(\xi^{can}|\mu^{j+1},\sigma^{j+1}\right)}{\pi\left(\xi^j|\mu^{j+1},\sigma^{j+1}\right)}\right\}\ .$$

$\pi(\theta\,|\,x)$ is the conditional posterior distribution for θ.

- Update the parameters according to the following condition.

$$\theta^{j+1}=\begin{cases}\theta^{can} & \text{with probability } p,\\ \theta^j & \text{with probability } 1-p.\end{cases}$$

- Iterate the updating procedure.

The variance of the candidate value υ_θ is typically chosen by trial and error and aiming at an acceptance probability roughly around 20% to 50%. Prior elicitation is an important issue in Bayesian analysis. This study applies the non-informative priors of Normal distribution with large variance to indicate that the expert information on extreme PM_{10} data in Johor is unavailable at the moment.

3. Description of data

Based on Malaysian Ambient Air Quality Guidelines, PM_{10} is the most influencing pollutant towards the air quality in Malaysia. PM_{10} notation is used to describe aerosol particles with diameter less than $10\mu m$ in the form of solids or liquids found suspended in the atmosphere [17]. Malaysia guideline for PM_{10} concentrations for 24-hours average and 12-months average are $150\mu g/m^3$ and $50\mu g/m^3$ respectively. High PM_{10} levels in Malaysia are particularly due to the haze and biomass burning as well as industrial and vehicle emissions. This situation of unease has been an annual problem across Malaysia. High PM_{10} level is a prominent issue which triggers various impacts on human health and material damage. Prolonged exposure to high concentrations of PM_{10} could be harmful to health especially on eye and throat irritations, and associated with numerous respiratory problems such as decreased lung function among sensitive groups.

In this study, the analyzed data consist of daily maxima PM_{10} data obtained from Department of Environment, Malaysia from January 1, 2000 to December 31, 2010 from three air monitoring stations in Johor. The first station is located in Johor Bahru. Johor Bahru is the main city centre of Johor in the southern portion of Peninsular Malaysia and is located north of Singapore. It is surrounded by main roads, highly developed industrial, commercial areas, tourist attractions and has a high population density. The second station is located in Pasir Gudang, an industrial area where the main industries are transportation and logistics, shipbuilding, petrochemicals and other heavy industries, and palm oil storage and distribution. Therefore, PM_{10} concentration is expected to originate mostly from industrial emissions as well as vehicles emissions. Muar station can be considered as a residential area located in northwestern Johor. The monthly maxima series were extracted from the original data to satisfy the independency in EV theory. Therefore only one data for each month was considered involving 132 monthly maximum from 2000 to 2010 data. The extremes data were fitted to the GEV model and the results obtained using maximum likelihood and Bayesian approaches are presented in the next section.

4. Results and discussion

Fig. 1 (a), (b) and (c) show the histogram and the density plot of monthly maxima PM_{10} data for Johor Bahru, Pasir Gudang and Muar stations. The plots show that the extreme PM_{10} concentrations are positively skewed and supported by the skewness and kurtosis in Table 1. The skewness value is greater than zero and kurtosis is greater than three, implying the presence of extreme values. Johor Bahru experienced only one day in 2006 with high PM_{10} level which exceeded 200 (very unhealthy). Pasir Gudang station had not experienced any day of very unhealthy PM_{10} level and four days for Muar in 2004, 2006, 2009 and 2010. Table 1 provides the summary statistics of the monthly maxima PM_{10} concentrations. The highest recorded PM_{10} concentration was found in Muar which is 532 in year 2010. The mean of the monthly maxima PM_{10} for the three stations are at moderate level.

Using, MH algorithm 15 000 iterations were carried out, of which the first 5000 iterations were discarded. The chains were satisfactory converged after the 5000 iterations for all parameters of the distribution and the remaining 10000 iterates formed stationary sequences of the marginal posterior distribution. The extreme PM_{10} concentrations for all the three stations fit to GEV distribution with $\xi > 0$ giving the heavy tailed case (Frechet type). The maximum likelihood estimates and the posterior means for GEV parameters of each site are given in Table 2. It is expected that the posterior means would be very close to the maximum likelihood estimates since non-informative priors are used and they did not much influence the likelihood. However, the entire Bayesian posterior means for all stations are higher than the maximum likelihood estimates.

Table 1 Summary statistics for monthly maxima PM_{10} for Johor Bahru, Pasir Gudang and Muar stations

Stations	minimum	maximum	mean	Standard deviation	skewness	kurtosis
Johor Bahru	36	236	70.90	24.48	3.06	19.00
Pasir Gudang	43	192	83.90	28.48	1.61	6.07
Muar	45	532	95.90	55.65	4.73	33.16

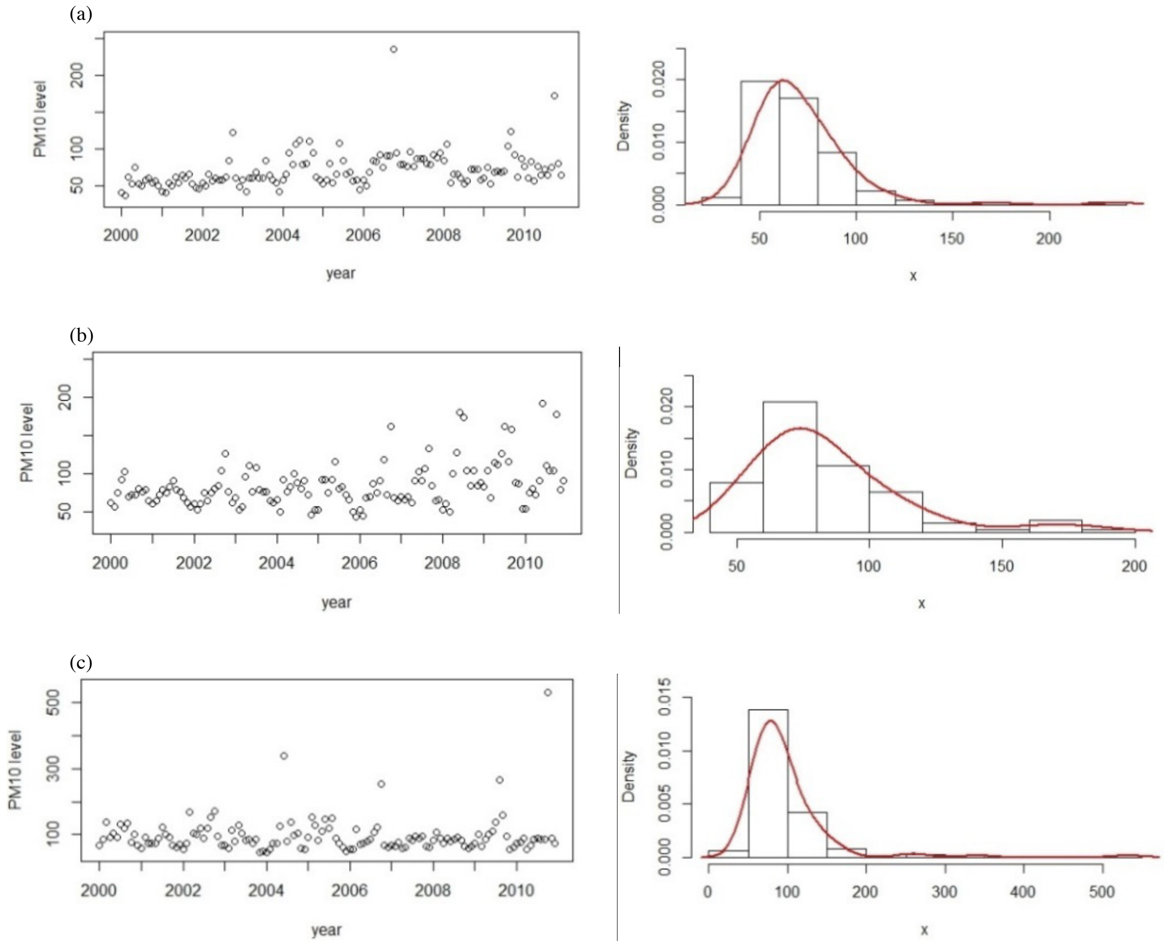

Fig. 1. Monthly maxima and density plots of PM_{10} data for (a) Johor Bahru, (b) Pasir Gudang and (c) Muar stations

Table 2 Maximum likelihood estimates and posterior means of GEV model for monthly maxima PM_{10}

Stations	Maximum likelihood estimates			Posterior means		
	μ	σ	ξ	μ	σ	ξ
Johor Bahru	60.47 (1.42)	14.70 (1.06)	0.12 (0.06)	60.52 (1.44)	15.08 (1.14)	0.12 (0.06)
Pasir Gudang	70.67 (1.82)	18.51 (1.38)	0.12 (0.07)	70.73 (1.81)	18.93 (1.44)	0.13 (0.07)
Muar	74.43 (2.12)	21.55 (1.75)	0.28 (0.07)	74.43 (2.11)	22.07 (1.83)	0.29 (0.07)

The predicted return levels for 10, 50 and 100 years in Table 3 show that Muar has higher return level of PM_{10} in the future compared to the other stations. We also noticed that the return level of PM_{10} concentration for Pasir Gudang station is higher compared to Johor Bahru although it did not experience PM_{10} higher than 200 during the study period. In comparison between Bayesian and maximum likelihood methods, the predicted return levels obtained by computation based on Bayesian approach is higher. The differences increase with the length of periods for the return levels.

Table 3 Maximum likelihood estimates (95% confidence intervals) and posterior mean (95% credibility intervals) for the 10, 50 and 100 years return levels

Stations	Maximum likelihood estimates			Posterior means		
	10	50	100	10	50	100
Johor Bahru						

	98.19 (90.18, 106.20)	132.77 (113.29, 152.25)	149.47 (122.22, 176.73)	99.79 (92.01, 109.97)	136.94 (119.54, 165.08)	154.34 (131.69, 196.15)
Pasir Gudang	118.69 (108.16, 129.22)	163.49 (136.14, 190.84)	185.36 (146.33, 224.39)	120.48 (110.74, 133.68)	168.46 (144.65, 207.13)	192.56 (159.17, 250.01)
Muar	142.12 (124.57, 159.67)	227.43 (171.82, 283.05)	277.25 (191.51, 362.99)	145.11 (129.26, 167.86)	238.22 (190.39, 320.26)	294.51 (221.72, 426.17)

5. Conclusion

EV theory affords some understanding to the tails of a distribution where standard models have proved unreliable. Since extreme environmental events may cause huge loss of properties and affect human lives, it is important to understand the behavior of such uncommon events and predict the upcoming occurrence. EV theory provides a sufficient model that could be used as a predictive tool for presentations on future air pollution scenarios and to help manage air pollution problems. The return level estimates verify the high level of extreme PM_{10} in upcoming occurrences. Bayesian approach is an alternative statistical analysis that could improve the model with the availability of prior knowledge. Thus, the application of prior information on extreme PM_{10} occurrences from an expert is recommended for a more reliable analysis and prediction.

Acknowledgements

The author would like to thank the Department of Environment, Malaysia for the data used and the Ministry of Education Malaysia for the SLAI scholarship and research grant (ERGS/1/2013/STG06/UPM/02/3).

References

1. Coles SG. *An Introduction to Statistical Modelling of Extreme Values*. Springer–Verlag, London; 2001.
2. Behrens CN., Lopes HF and Gamerman D. Bayesian analysis of extreme events with threshold estimation. *Statistical Modelling* 2004; **4**: 227-244.
3. Haan D and Ferreira L. *Extreme Value Theory: An Introduction*. Springer, New York; 2006.
4. Payus C, Abdullah N, and Sulaiman N. Airborne Particulate Matter and Meteorological Interactions During the Haze Period in Malaysia. *International Journal of Environmental Science and Development* 2013; **4** (4): 398-402.
5. Dominick D, Juahir H, Latif MT, Zain SM and Aris AZ. Spatial assessment of air quality patterns in Malaysia using multivariate analysis. *Atmospheric Environment* 2012; 172-181.
6. Yusof NFFM, Ramli NA and Yahaya AS. Extreme Value Distribution for Prediction of Future PM $_{10}$ Exceedences. *International Journal of Environmental Protection* 2011; **1**(4): 2836.
7. Sharma P, Avinash C, Kaushik SC, Sharma P and Suresh J. Predicting Violations of National Ambient Air Quality Standards Using Extreme Value Theory for Delhi City. *Atmospheric Pollution Research* 2012; **3**: 170-179.
8. Lu H S. Estimating the Emission Source Reduction of PM_{10} in Central Taiwan. *Chemosphere* 2004; **54** (7): 805-814.
9. Coles SG and Tawn JA. A Bayesian Analysis of Extreme Rainfall Data. *Appl Statistics* 1996; **45**:463-478.
10. Carlin BP and Louis TA. Bayesian Method for Data Analysis. Third Edition. Chapman and Hall/CRC, United States of America; 2011.
11. Chib S and Greenberg E. Understanding the Metropolis-Hastings algorithm. *The American Statistician* 1995; **49**(4): 327-335.
12. Gamerman D and Lopes HF. Markov chain Monte Carlo: stochastic simulation for Bayesian inference. CRC Press USA; 2006.
13. Brooks SP and Roberts GO. Assessing convergence of Markov chain Monte Carlo algorithms. *Statistics and Computing* 1998; **8**(4): 319-335.
14. Coles SG and Powell EA. Bayesian Methods in Extreme Value Modelling: A Review and New Developments. *International Statistical Review* 1996; **64:** 119-136.
15. Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH and Teller E. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics* 1953; **21**(6): 1087-1092.
16. Hastings W K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 1970; **57**(1): 97-109.
17. Juneng L, Latif MT, Tangang F T and Mansor H. Spatio-temporal characteristics of PM_{10} concentration across Malaysia. *Atmospheric Environment* 2009; **43:** 4584-4594.