

Comparison on the Analysis on PM₁₀ Data based on Average and Extreme Series

Nor Azrita Mohd Amin^{1,*} and Siti Aisyah Zakaria¹

¹Institute of Engineering Mathematics, Universiti Malaysia Perlis, Kampus Pauh Putra, 02600 Arau, Perlis

Abstract. The main concern in environmental issue is on extreme phenomena (catastrophic) instead of common events. However, most statistical approaches are concerned primarily with the centre of a distribution or on the average value rather than the tail of the distribution which contains the extreme observations. The concept of extreme value theory affords attention to the tails of distribution where standard models are proved unreliable to analyse extreme series. High level of particulate matter (PM₁₀) is a common environmental problem which causes various impacts to human health and material damages. If the main concern is on extreme events, then extreme value analysis provides the best result with significant evidence. The monthly average and monthly maxima PM₁₀ data for Perlis from 2003 to 2014 were analysed. Forecasting for average data is made by Holt-Winters method while return level determine the predicted value of extreme events that occur on average once in a certain period. The forecasting from January 2015 to December 2016 for average data found that the highest forecasted value is 58.18 (standard deviation 18.45) on February 2016 while return level achieved 253.76 units for 24 months (2015-2016) return periods.

1 Introduction

Environmental quality management is more concerned about extreme situations due to its various hazardous impacts. However, most statistical methods are concerned primarily with what goes on in the centre of a statistical distribution, and do not pay particular attention to the tails of the distribution. Statistical modelling of extreme air quality has a very practical motivation since these events have major effects on human and ecosystem. Degradation of ambient air quality standards reduces visibility, impairs air, land and water transportation and seriously affects the economy.

The region's drier weather conditions has led to escalation in hotspot activities that are caused mainly by land clearing and "slash and burn" agricultural practices [1]. Improper management of open burning for commercial plantation sectors, heavy industries and business activities has made the situation worse. High PM₁₀ levels have been a common problem in Malaysia especially in the dry season. During the haze periods, PM₁₀ was found as the main pollutant while the other air quality parameters remained within the permissible healthy standards [2]. Study conducted by Dominick et al. [3] found that air pollution in eight selected air monitoring stations in Malaysia based on year 2008 to 2009 are predominantly influenced by PM₁₀. Yusof et al. [4] claimed that the hourly average PM₁₀ 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.

A time series is a collection of past values of the variable being predicted. Holt-Winters method was first proposed in the early 1960s. It uses a process known as exponential smoothing. Exponential smoothing in its simplest form should only be used for non-seasonal time series exhibiting a constant trend or also known as a stationary time series. The smoothed series depends on all previous values, with the most weight given to the most recent values. The Holt-Winters model uses a modified form of exponential smoothing. It applies three exponential smoothing formulae to the series which represent by α , β and γ hyper parameters. Firstly, the level (or mean) is smoothed to give a local average value for the series. Secondly, the trend is smoothed and lastly each seasonal sub-series (ie all the January values, all the February values and so on, for monthly data) is smoothed separately to give a seasonal estimate for each of the seasons. The Holt-Winters method set the hyper parameters γ to 0 which apply the exponential smoothing with trend and without seasonal component [5].

Extreme value (EV) theory is unlike other statistical approaches since its focus is on the tail of distribution either on maxima or minima values. The scope of EV theory has been widely explored in various fields. Recently, it has become a vigorously research area due to its implication in many applications. Literatures on EV theory among others are by Coles [6] and Haan [7] that provide from basic EV theory to the application of EVT in various fields. Basically, there are two main models for extreme series which is generalized extreme value distribution (GEV) and generalized Pareto distribution (GPD). This study focuses on extreme PM₁₀

* Corresponding author: norazrita@unimap.edu.my

concentrations based on GEV model. This paper aim to approximate the possible extreme PM₁₀ levels in the future.

2 Methodology

There are two main approaches on the analysis of PM₁₀ data in this research. The first approach is using the monthly average data where we use the time series approach to make a proper forecasting based on Holt's linear trend method. Second approach is based on EVT where the extreme data are extracted from the original data based on monthly maxima data.

2.1 Holt-Winters Method

Holt-Winters method consists in a simple yet effective forecasting procedure, based on exponential moving averages, covering both trend and seasonal models [5]. As mentioned earlier the Holt-Winters model assumes that the seasonal pattern is relatively constant over the time period. The exponential smoothing formulae applied to o a series with a trend and constant seasonal component using the Holt-Winters additive technique are:

$$\begin{aligned}\hat{a}_t &= \alpha(Y_t - \hat{s}_{t-p}) + (1-\alpha)(\hat{a}_{t-1} + \hat{b}_{t-1}) \\ \hat{b}_t &= \beta(\hat{a}_t - \hat{a}_{t-1}) + (1-\beta)\hat{b}_{t-1} \\ \hat{s}_t &= \gamma(Y_t - \hat{a}_t) + (1-\gamma)\hat{s}_{t-p}\end{aligned}$$

where:

- α, β and γ are the smoothing parameters
- \hat{a}_t is the smoothed level at time t
- \hat{b}_t is the change in the trend at time t
- \hat{s}_t is the seasonal smooth at time t
- p is the number of seasons per year

2.2 Extreme Value Theory

EVT provides a concrete theoretical groundwork on which statistical models for describing extreme events are properly set up. General discussions about EVT are about their conventional forms which are Gumbel, Frechet and Weibull that is unified into its general form generalized extreme value distribution (GEV) and the recent approach is based on generalized pareto distribution GPD which depends on the threshold value of the data. This paper analyse the data based on GEV model where the series of extreme data were extracted based on monthly maxima (block).

2.1.1 Generalized Extreme Value Distribution

The cumulative distribution function (cdf) and probability density function (pdf) of the GEV

distribution is given by (1) and (2) respectively. The GEV model has three parameters, $-\infty < \mu < \infty$, $\sigma > 0$ and $-\infty < \xi < \infty$ which refers to location, scale and shape parameter respectively. The ξ value determines the type of GEV distribution. $\xi > 0$ correspond to Frechet distribution, $\xi < 0$ correspond to Weibull distribution and $\xi = 0$ correspond to Gumbel distribution.

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

$$g(z) = \frac{1}{\sigma} \exp\left\{-\exp\left[-\left(\frac{z-\mu}{\sigma}\right)\right] - \left(\frac{z-\mu}{\sigma}\right)\right\} \quad (2)$$

Earlier works on EVT take on one of the three distributions and subsequently estimate the corresponding parameters. According to Coles [6], there are two weaknesses regarding this issue. First, a technique is required to choose the most appropriate distribution for the data analyzed and second, the inferences are made with the assumption that the choice is correct. Therefore, a better analysis could be done using GEV where the value of shape parameter, ξ itself will determine the most suitable tail behavior of the data.

2.1.2 Return Level

Application of EVT in air quality studies are concerned about how well the mathematical theory can be applied to further answer questions relating to the probability that pollutant concentration will exceed a certain level in a given period which refers to the return level. Awareness of the return levels of extreme air pollution events could benefit the development of air pollution risk management practices. This motivates the need to estimate the most terrible PM₁₀ level that will be occur over a certain period in the future.

Return levels in extremes explain the value of extreme events that occur on average once in a given period. For example, what is the PM₁₀ level that will be exceeded on average once in the next 100 years? It is convenient to interpret extreme value models in terms of quantiles or return levels rather than individual parameter values [8]. The return level for GEV with return period $1/p$ is defined by z_p .

$$z_p = \begin{cases} \mu - \frac{\sigma}{\xi} \left\{ -[-\log(1-p)]^{-\xi} \right\} & \xi \neq 0 \\ \mu - \sigma \log[-\log(1-p)] & \xi = 0 \end{cases} \quad (3)$$

3 Data Description

The air quality level in Malaysia is described in terms of Air Pollutant Index (API). API is an indicator of the air quality and is developed based on scientific assessment to indicate in a manner that can be easily understood, the presence of pollutants and its impact on health. The API scale and terms used in describing air quality levels are categorized as in **Table 1**. In extreme value analysis, most of the data considered range between moderate to

very unhealthy level for which necessary actions need to be taken. The CAQM measures concentrations of five major pollutants in the ambient air, namely, PM₁₀, sulphur dioxide (SO₂), nitrogen dioxide (NO₂), carbon monoxide (CO), and ozone (O₃). PM₁₀ is used to describe aerosol particles with diameter of less than 10µm for solids or liquids found suspended in the atmosphere.

The API is calculated based on the average daily concentrations of SO₂, NO₂, CO, O₃ and PM₁₀. Dominant air pollutant with the highest concentration considered as pollutants that will determine the IPU. Typically, concentration of PM₁₀ is the highest compared to the pollutant others, and this determines the API readings. PM₁₀ concentration is related to gases and particulates which are expected to originate mostly from industrial and vehicle emissions and also from some transboundary pollutions involving Malaysia. The three major sources of air pollutions especially in urban areas are mobile (motor vehicles), stationary (power stations, industrial fuel burning process and domestic fuel burning) and the burning of municipal and industrial waste.

Table 1. API status indicator

API scale	Air quality status
0-50	Good
51-100	Moderate
101-200	Unhealthy
201-300	Very unhealthy
301 and above	Hazardous

During haze periods, PM₁₀ was found as the main pollutant while the other air quality parameters remained within permissible healthy standards. The extreme level of PM₁₀ data is particularly due to the haze and biomass burning as well as industrial and vehicle emissions which usually contribute to high PM₁₀ levels.

3.1 Sampling Site

The data collected is for Perlis region. Perlis is located in the north peninsular Malaysia and bounded with Thailand in the north and Kedah in the south with total area around 795 km². Perlis is distributed to three parliament which is Kangar, Arau and Padang Besar and categorized as sub-urban area. Figure 1 shows the continuous air quality monitoring station (CAQMS) used to measure the air pollution located at Institut Latihan Perindustrian, Kangar. The device must be at 30 degree elevation from nearby building, which means that any taller building near it will interrupt the measurement to be recorded

In Perlis however, source of pollution is probably mainly from open burning of paddy fields, in most part of Perlis. Other than that is due to the haze from neighbouring country, Indonesia.

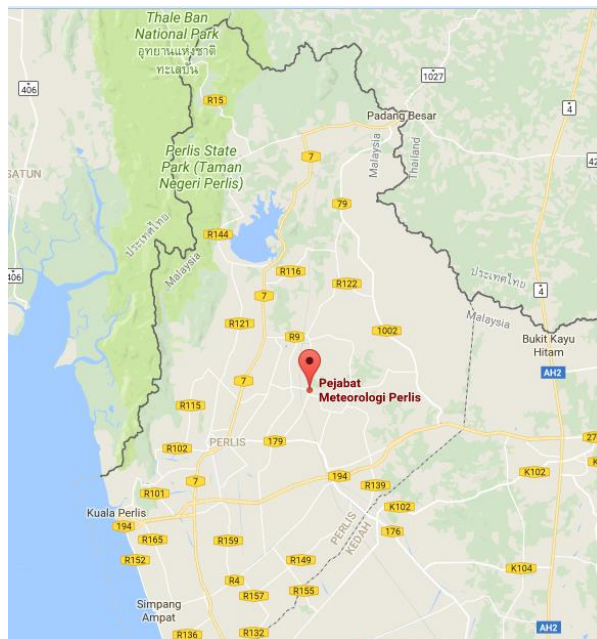


Fig. 1. Location of Continuous Air Quality Monitoring System Perlis.

3.2 PM₁₀ Data

PM₁₀ notation is used to describe aerosol particles with diameter less than 10µm in the form of solids or liquids found suspended in the atmosphere. Malaysia guideline for PM₁₀ concentrations for 24-hours average and 12-months average are 150µg /m³ and 50µg /m³ respectively. In air quality control, PM₁₀ is recognized as the most influencing atmospheric pollutant for air quality index in a majority of cities in Malaysia. The adverse effects of PM₁₀ to human health and material damages are the main reasons for extensive explorations on the behaviour of this pollutant. Prolonged exposure to high concentrations of PM₁₀ can be harmful to health especially on eye and throat irritations and respiratory problems among sensitive groups.

4 Results and Discussions

This section analyse the data based on the monthly average and monthly maxima PM₁₀ data for year 2003 to 2014. Results were compare for both series using descriptive statistics and forecast value for year 2015 and 2016 based on Holt-Winters method with hyper parameters γ to 0 which apply the exponential smoothing with trend and without seasonal component

4.1 Data Analysis

Fig. 2. and **Fig. 3.** show the plot of monthly average temperature data with corresponding summary statistics in **Table 2.** In general both figures doesn't show any seasonal trend. The average data show obvious different on the level between January 2004 to January 2006 to the rest of the period. Transboundary haze is expected contributed to the higher PM₁₀ recorded intermittently due to the land and forest fires in the Riau Province of

Central Sumatera, Indonesia. Additionally, quite persistent pattern from July 2006 to July 2013. Most of the data are lied between 30 and 50 units. The means for average data is 42.12 while maxima data is 144.53 units. From **Fig. 3.**, the obvious peaks are mostly happen in July especially for years 2005, 2008 and 2013. This is probably due to very hazy weather condition from forest fires in Sumatra. Haze is an annual problem during the monsoon season from May to September as winds blow the smoke across the Malacca Strait to Malaysia.

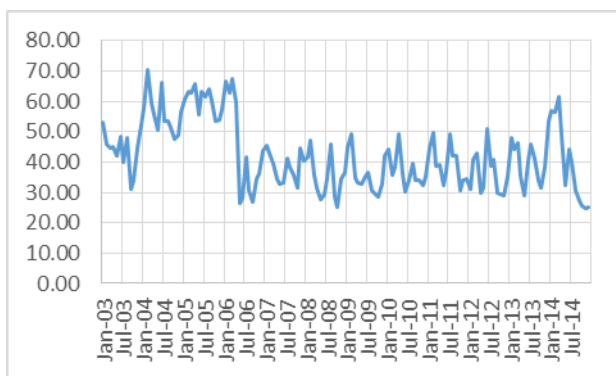


Fig. 2. Monthly Average of Temperature Data

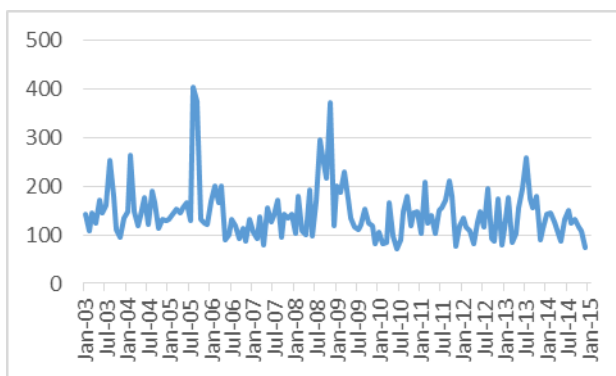


Fig. 3. Monthly Maxima of Temperature Data

Table 2. Descriptive Statistics for Monthly Average and Monthly Maxima Temperature Data for year 2003 to 2014

	Min	Max	Average	Std.
Monthly average	24.63	70.43	42.12	11.15
Monthly maxima	72.00	404.00	144.53	54.67

Inferences on the extremes of environmental events are essential as guidelines in designing structures in

Table 4. Forecast values of monthly average PM₁₀ for year 2013 to 2016 with observed values for year 2013 and 2014 (bold)

Months	Jan	Feb	Mac	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec
2013	47.81	44.38	46.26	35.53	28.97	40.54	45.75	41.38	34.00	31.32	38.17	53.59
	42.66	54.53	42.47	37.46	31.49	46.96	31.93	38.83	20.48	22.33	32.16	38.65
2014	56.72	56.45	61.39	48.99	32.26	44.12	40.09	30.77	27.31	25.76	24.63	25.27
	42.88	54.75	42.69	37.67	31.71	47.17	32.14	39.05	20.70	22.55	32.37	38.87
2015	43.09	54.96	42.90	37.89	31.92	47.39	32.36	39.26	20.91	22.76	32.59	39.08
2016	43.31	55.18	43.12	38.11	32.14	47.60	32.58	39.48	21.13	22.98	32.81	39.30

order to survive under the utmost extreme conditions. Extreme air pollutants caused various effects associated to human health and material damages. In many cases, the pollutants are responsible for huge impacts on economic performances. The EV theory is applied to model the extreme PM₁₀ pollutant for three air monitoring stations in Johor. This study analyse the extreme PM₁₀ data based on maximum likelihood estimation technique and hence estimate the return level. The estimated parameter for GEV model given in **Table 3.**

Table 3. Maximum likelihood estimates (with standard error) for GEV model

Parameter	μ	σ	ξ
MLE	119.67 (3.21)	34.01 (2.46)	0.14 (0.06)

4.2 Forecasting

The monthly average data is forecast by using Holt-Winters method. The observed, fitted and predicted monthly average PM₁₀ values are shown in **Fig. 4.** The confidence intervals for the forecast values are plotted in the same plot. The forecast values found that the highest forecast value from January 2015 to December 2016 is 58.18 unit which is on Feb 2016. The details forecasting values are given in **Table 4.** Compare to GEV model, the return level for 24 months return period gives the return level value is 253.76 unit. This means that the highest value predicted using GEV model will reach very unhealthy level in two years period.

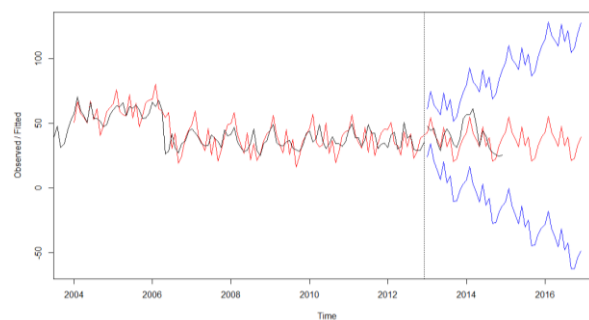


Fig. 4. Holt-Winters method of monthly average with 95% confidence limits in dashed blue lines. Original data are plotted in gray, fitted and predicted values in red.

5 Conclusions

From the monthly average data, the PM₁₀ data is fairly good except during periods of haze as indicated by its API. However, the average for monthly maxima series indicates that the pollutant in very unhealthy level. This obvious different is that the monthly average values explain the ordinary situation while the monthly maxima refer to the extreme cases only. Thus, the forecasting value for average data reflect the predicted values for the next two years period while the return level explain the maximum level that can be estimate in average for the next two years. Therefore, if our interest is on examining the extreme occurrence or catastrophic events, extreme value theory is the best methods that should be used.

The authors would like to thank Universiti Malaysia Perlis for the short term research grant (STG 9001-00532).

References

1. Amin N. A. M, Adam M. B and Aris A. Z. J. Teknologi **76(1)** 171–179 (2015). Payus C, Abdullah N, and Sulaiman N. Int. J. Env. Sci. Dev. **4** (4): 398-402 (2013).
2. Dominick D, Juahir H, Latif MT, Zain SM and Aris AZ. Atm. Env. 172-181 (2012).
3. Yusof NFFM, Ramli NA and Yahaya AS. Int. J. Env. Protect. **1(4)**: 28-36 (2011).
4. Meyer D. The Holt-Winter Method in Package ts. The Newsletter of the R Project. **2(2)** (2002)
5. Coles S. G. An Introduction to Statistical Modeling of Extreme Values, 3rd ed. London: Springer-Verlag. (2001)
6. Haan D. and Ferreira L. Extreme Value Theory: An Introduction. Springer, New York (2006).
7. Gilli M. and Kellezi E. Computational Economics. 207-228 (2006).
8. Juneng L, Latif MT, Tangang F T and Mansor H. Atmospheric Environment **43** 4584-4594 (2009).