Mathematics and Statistics 7(4A): 58-64, 2019
DOI: 10.13189/ms.2019.070708 http://www.hrpub.org

Tree-based Threshold Model for Non-stationary Extremes with Application to the Air Pollution Index Data

Afif Shihabuddin[1], **Norhaslinda Ali**[1,2,*], **Mohd Bakri Adam**[1,2]

[1]Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

[2]Department of Mathematics, Faculty of Science,Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia

Received July 1, 2019; Revised August 22, 2019; Accepted August 30, 2019

Abstract Air pollution index (API) is a common tool used to describe the air quality in the environment. High level of API indicates the greater level of air pollution which will gives bad impact on human health. Statistical model for high level of API is important for the purpose of forecasting the level of API so that the public can be warned. In this study, extremes of API are modelled using Generalized Pareto Distribution (GPD). Since the values of API are determined by the value of five pollutants namely sulphur dioxide, nitrogen dioxide, carbon monoxide, ozone and suspended particulate matter, data on API exhibit non-stationarity. Standard method for modelling the non-stationary extremes using GPD is by fixing the high constant threshold and incorporating the covariate model in the GPD parameters for data above the threshold to account for the non-stationarity. However, high constant threshold value might be high enough on certain covariate for GPD approximation to be a valid model for extreme values, but not on the other covariates which leads to the violation of the asymptotic basis of GPD model. New method for the threshold selection in non-stationary extremes modelling using regression tree is proposed to the API data. Regression tree is used to partition the API data into a stationary group with similar covariate condition. Then, a high threshold value can be applied within a group. Study shows that model for extremes of API using tree-based threshold gives a good fit and provides an alternative to the model based on standard method.

Keywords Air Pollution Index, Threshold Exceedances, Generalized Pareto Distribution, Non-stationary, Regression Tree, Tree-based Threshold

1 Introduction

Air quality is an important aspect of human life. In Malaysia, it is monitored and enforced by the Department of Environment (DOE). Air quality is determined by the Air Pollution Index (API) which is updated hourly by DOE. API value is derived from five main pollutants which are ground level ozone (O_3), nitrogen dioxide (NO_2), particulate matter (PM_{10}), carbon monoxide (CO) and sulphur dioxide (SO_2). High concentration of these pollutants in the air is harmful for everyone and also causing serious health problem. High concentration of SO_2, NO_2 and CO in the air can cause a heart and lungs problem [1]. High level of PM_{10} is associated with haze days which can limit the eyesight and cause the respiratory problem [2]. According to Malaysian Ambient Air Quality Guidelines (MAAQG), API values which above 100 are considered unhealthy and could threaten public health. Therefore it is important to understand the behavior of high level of API particularly to give health warnings for the public. In order to describe the behavior of high level API at a particular area, it is important to identify the distributions which best fit the data [3]. Extreme value distribution from extreme value theory are suitable in modelling such high values.

Extreme value theory (EVT) is a branch of statistics which study the stochastic behavior of a process at unusually large or small values [3]. Particularly, EVT provides procedures for tail estimation which are scientifically and statistically rational. Two significant results from EVT are first, the asymptotic distribution of the standardized series of maxima (minima) is shown to converge to the Gumbel, Frechet, or Weibull distributions. A standard form of these three distributions is called the generalized extreme value (GEV) distribution. The second significant result concerns the distribution of excess over a given threshold, where the limiting distribution is a generalized Pareto distribution (GPD). The application of EVT is important in various disciplines such as hydrology, meteorology, finance, engineering, insurance and environment. Common applications of EVT include modeling the extremes of heavy rainfall, sea-levels, pollution concentrations and many more. There are many application of extreme value

analysis (EVA) in the context of air pollution data set in many parts of the world. An early comprehensive review of the application of EVA to the air quality data (SO_2 and NO_2) can be found in [4] and [5]. [6] compares the performance of GEV and GPD fitted on PM_{10} data based on estimated parameters and return levels while [7] modelled API values which are above 100 using GPD. [8] use MRL plot to select thresholds for API, O_3 and PM_{10} data. Using Pickands dependence function plots, [8] shows that the PM_{10} and O_3 are the dominant pollutants which could affect the API at high level.

As the API value varies according to the variation of five main pollutants which are O_3, NO_2, PM_{10}, CO and SO_2, considering these pollutants into model for API seems reasonable. Modelling the high API in the presence of these covariates also known as model for non-stationary, requires a specific treatment to account for these non-stationarity. Standard extreme non-stationary model using GPD is to fixed a high threshold u while the effect of non-stationarity is accounted by the inclusion of covariate model in the GPD parameters. However, the pre-selected high u does not guarantee that it is high enough for all the covariate to produce those API values, hence, imposing an inaccurate approximation of GPD as a model for threshold exceedances [9]. One of the possible solution for these problem is to use high u for API values produced by similar pollutants condition. This can be achieved by grouping the pollutants into the same group which produce most, if not all, stationary API values. In this paper we propose a new threshold selection for non-stationary GPD model using a regression tree. This paper is organized as follows. In Section 2, we explain the data that will be used in this study followed by an explanation on a methodology in Section 3. Section 4 evaluate the performance of the proposed method by a simulation study. Section 5 will discuss the finding of the research and some concluding remarks in Section 6.

2 Data and Study Area

The hourly air quality data consist of air pollution index (API) and hourly average of ground level ozone (O_3), nitrogen dioxide (NO_2), particulate matter (PM_{10}), carbon monoxide (CO) and sulphur dioxide (SO_2) obtained from Department of Environment Malaysia for the period of 1st January 2008 until 31st December 2017. The data is from a Continuous Air Quality Monitoring Station located in Klang Valley which is Sekolah Kebangsaan Bandar Damansara Utama, Petaling Jaya station. This station are considered as residential and industrial areas where air pollutants are produced at a higher rate [10]. Besides, Petaling Jaya also located at the edge of Kuala Lumpur, the capital city of Malaysia which makes it a highly populated area. Figure 1 shows the location of Petaling Jaya station in the state of Selangor. The minimum, maximum and median values for API observations are 18, 257 and 54 respectively. According to Malaysian Ambient Air Quality Guidelines (MAAQG), API values which above 100 are considered unhealthy and could threaten public health. In this data set, there is 97 API observations which exceed 100.

Figure 1. Map of Selangor and Petaling Jaya station.

3 Methodology

3.1 Model Formulation for the Tree-based Threshold

Let $Y_1, Y_2, \ldots$ be a sequence of independent random variables with common continuous distribution function $F(y)$ and denote the upper end point as y^F. The extreme observations refer to those of the y that exceed some pre-determined high threshold u with $u < y^F$. According to [11], as $u \to y^F$, the distribution of the threshold exceedances, $z = y - u | y > u$ can be modeled by distribution function of the form

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

defined on the set $\{z : z > 0 \text{ and } (1 + \xi z/\sigma_u) > 0\}$. The distribution function defined by Eq. (1) is called the Generalized Pareto distribution (GPD). The parameters of the GPD are determined by the scale $\sigma_u > 0$ and shape $-\infty < \xi < \infty$. The density function of GPD is

$$g(z) = \frac{1}{\sigma_u}\left[1 + \xi\frac{z}{\sigma_u}\right]^{-1/\xi-1}. \tag{2}$$

The parameters of the GPD can be estimated by maximum likelihood estimation method. Suppose that the values $z_1, \ldots, z_k$ are the k threshold exceedances. The likelihood function derived from Eq. (2) is

$$L(\theta) = \prod_{i=1}^{k} \frac{1}{\sigma_u}\left[1 + \xi\frac{z_i}{\sigma_u}\right]^{-1/\xi-1} \tag{3}$$

with $\theta = (\sigma_u, \xi)$. By taking a logarithm, the likelihood function given by Eq. (3) becomes

$$\ell(\theta) = -k \log \sigma_u - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \log \left(1 + \xi \frac{z_i}{\sigma_u}\right). \quad (4)$$

We use numerical optimization method to optimize the log-likelihood function in Eq. (4) since the analytical maximization is not possible.

In real-life applications, the distribution of the data sets cannot always be assumed to be identically distributed. This situation which is known as non-stationary is often apparent because of seasonal effects, trends or because the variable of interest is related to covariate. The usually adopted approach is by using the standard extreme value models as a basic templates that can be enhanced by statistical modeling.

Let $Y_1, Y_2, \ldots$ be a non-stationary series and information about some covariates $\{\mathbf{X}\}$ are available. Suppose that the random variable Y are related to the random variables $\mathbf{X}$. The standard method for modeling the extremes of non-stationary series is focuses on retaining a constant high threshold u and incorporating the covariate models in Generalized Pareto (GP) parameters to account for the non-stationarity [12]. The distribution of the threshold exceedances from a non-stationary series can be model by

$$(y - u | y > u) \sim \text{GP}(\sigma_u(\mathbf{x}), \xi(\mathbf{x})) \quad (5)$$

where $\sigma_u(\mathbf{x})$ and $\xi(\mathbf{x})$ are the covariate models. The distribution of the threshold excesses in Eq. (5) can be approximate by the GP if each covariate $\mathbf{x}$ have a high enough threshold. However, high enough threshold for one covariate might not be high enough for the other covariate to produce those high y values leads to the invalidity of the GP to approximate the distribution of threshold exceedances. To remedy the problem, we propose a tree-based threshold to model the threshold exceedances. The regression tree is use to partition the y sequences into m homogeneous stationary clusters. In order to obtain a stationary cluster, we apply a stationary test and use a stopping criteria for growing the tree. We defer the discussion of these to Section 3.2.

Suppose that the observations within each clusters produced by regression tree are stationary or approximately stationary. Then, a constant high threshold can be set within each clusters producing a different threshold in each of the clusters known as tree-based threshold. Then, the distribution of the tree-based threshold exceedances can be model by

$$(y_k - u_k | y_k > u_k) \sim \text{GP}(\sigma_{u_k}, \xi) \quad (6)$$

where y_k is the observations within the cluster k ($k = 1, 2, \ldots, m$) and u_k is a threshold set in a cluster k. Denote the tree-based threshold exceedances $z_k = y_k - u_k$, the distribution of z_k has a form of

$$G(z_k) = 1 - \left[1 + \frac{\xi z_k}{\sigma_{u_k}}\right]_+^{-1/\xi}. \quad (7)$$

The density function of distribution function of Eq. (7) is similar as in Eq. (3). The estimation of the parameters is simply done using maximum likelihood estimation by maximizing the likelihood function as given in Eq. (3).

If the covariate model is still needed to model the distribution of the tree-based threshold exceedances due to the inability of the regression tree produced the stationary observations in each cluster, GP model with covariate function in the parameters can also be estimated using maximum likelihood estimation method. Let $z_{i1}, z_{i2}, \ldots, z_{ik}$ for $i = 1, 2, \ldots, n_{uk}$ where n_{uk} is a number of exceedances in cluster k, be a tree-based threshold exceedances follow the $\text{GP}(\sigma_{u_k}(\mathbf{x}), \xi(\mathbf{x}))$ where each of $\sigma_{u_k}(\mathbf{x})$ and $\xi(\mathbf{x})$ have an expression in terms of parameters vector and covariates. Denoting β as a vector of parameters of a covariate model, the numerical technique is required to optimize the likelihood function derived from Eq. (7) to estimate β.

3.2 Stopping Criteria in Regression Tree

Regression tree is a supervised learning method that construct a flowchart-like tree from the data as a prediction tree model and uses the model to classify the future data [13]. Regression tree consists of one parent node, internal nodes and terminal nodes. In our study, we refer to the terminal nodes as a cluster which consist of stationary observations. To determine which cluster an observation is belongs to, all observations are placed at the root (parent node) of the tree. We follow a path from the root and proceed to one of the internal node called leaf by following a question that split the parent node. The observations with *yes* answer will be placed at the left leaf (internal node) while observations with *no* answer will be placed at the right leaf (internal node). The tree model is fitted using binary recursive partitioning where a parent node in a decision tree is split into two internal nodes based on the splitting criterion. The split is chosen such that the impurity level of the tree is reduced the most by the split. The tree impurity level is measured by sum squared errors of the tree which given by

$$S = \sum_{c \in leaves(T)} \sum_{i \in c} (y_i - m_c)^2$$

where $m_c = \frac{1}{n_c} \sum_{i \in c} y_i$, is the mean of observations within leaf c.

The binary partitioning process will be applied over and over again until it meets some stopping criteria. The stopping criteria is set to control the size of the tree so that the tree will stop to grow when the observations within the clusters are stationary. In this study, we set a value δ, such that the reduction of sum squared errors of the tree after each split is not less than δ. We consider values of δ between 0.000001 and 0.01 with 0.000001 interval. The maximum value of δ which grow the tree with stationary observations within a cluster will be chosen as a stopping criteria. The stationarity of the observations within the cluster is tested using Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test at significance level $\alpha = 0.05$. Then, a constant high threshold can be set within each resulted clusters.

The threshold is set at q^{th} percentile where q kept similar for all clusters so that the rate of exceedances remain constant throughout the data set. Since each cluster has different number of observations, the threshold value might differ for each clusters. In other words, each observation will have their own threshold value. These threshold values are arranged according to the index of observations producing a varying threshold. In this study, the 95^{th} percentile value for threshold are chosen for each cluster. This percentile value is reasonably sufficient for GPD approximation be a valid limiting distribution for threshold exceedances while still keep the number of exceedances large enough for the model estimation [14].

4 Simulation Study

In this section we will illustrate the efficiency of the tree-based threshold method over the standard method for modelling the non-stationary extremes by a simulation study. We simulate random numbers from generalized extreme value distribution using inverse sampling method. Our argument for the choice of GEV distribution is as follows. If $Y_1, \ldots, Y_n$ is distributed as GEV(μ, σ, ξ), then, it can be shown that the block maxima $M_n = \max(Y_1, \ldots, Y_n)$ will also GEV(μ^*, σ^*, ξ) with

$$\mu^* = \mu + \frac{\sigma(n^{\xi} - 1)}{\xi} \text{ and } \sigma^* = n^{\xi}\sigma.$$

According to [3], if the distribution of a block maxima is GEV, then the excesses of high enough threshold, u can be approximated by GPD with parameter $\tilde{\sigma}$ and ξ where

$$\tilde{\sigma} = \sigma^* + \xi(u - \mu^*).$$

Here, parameter ξ is equal to that of the corresponding parameter ξ in GEV distribution.

Covariates model is incorporated in the GEV location parameter, μ to induce non-stationarity in the simulated random numbers. Two covariate models are used which are:

1. $\mu = \mu_0 + \mu_1 \left(\frac{t}{n+1}\right) + \mu_2 x$ for linear trend,
2. $\mu = \mu_0 + \mu_1 \cos(\frac{2\pi t}{n}) - \mu_2 \sin(\frac{2\pi t}{n}) + \mu_3 x$ for cyclic trend

where t and n represent time covariate and number of observations respectively. Another covariate x is generated from standard normal distribution. Time covariate, t is included to create trends in the data sets, while the covariate x represents a random variable which might affect the variable y. The time covariate is simply an increasing index from 1 until 3653 which corresponds to number of days in 10 years. The covariate x is simulated using function `rnorm` in R statistical software. We consider four non-stationary GEV data sets of size $n = 3653$, each containing either a linear or a cyclic trend in parameter μ with shape parameter $\xi = 0.4$ and $\xi = -0.4$. The abbreviation of non-stationary GEV data sets are given in Table 1. The scale parameter, σ is fixed at 1 for all data sets. The location parameter μ_0, μ_1, μ_2, μ_3 are chosen arbitrarily and given in Table 2.

Table 1. The Abbreviation for simulated non-stationary GEV data sets.

Data set	Abbreviation
GEV with Linear trend and Positive Shape Parameter	GEVLP
GEV with Linear trend and Negative Shape Parameter	GEVLN
GEV with Cyclic trend and Positive Shape Parameter	GEVCP
GEV with Cyclic trend and Negative Shape Parameter	GEVCN

Table 2. Location parameter for simulated data sets.

Data set	μ_0	μ_1	μ_2	μ_3
GEVLP	1	10	1	-
GEVLN	1	10	1	-
GEVCP	1	5	5	1
GEVCN	1	10	10	1

The tree-based threshold selection method is applied to the simulated data sets. The excesses of tree-based threshold are modelled by both stationary and non-stationary GP model. Covariate models are incorporated in the scale parameter of non-stationary GP model such that the scale parameter is either

$$\sigma = \exp\left\{\sigma_0 + \sigma_1 \left(\frac{t}{n+1}\right) + \sigma_2 x\right\}, \tag{8}$$

or

$$\sigma = \exp\left\{\sigma_0 + \sigma_1 \cos\left(\frac{2\pi t}{n}\right) - \sigma_2 \sin\left(\frac{2\pi t}{n}\right) + \sigma_3 x\right\} \tag{9}$$

where Eq. (8) is for data set with linear trend while Eq. (9) is for data set with a cyclic trend. The performance of the fitted stationary GP and non-stationary GP models are compared using Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The AIC and BIC values are shown in Table 3. Table 3 shows that the AIC and BIC values of stationary model fitted to the tree-based threshold exceedances for simulated data with positive shape parameters are smaller compared to the non-stationary model. For simulated data with negative shape parameter, both AIC and BIC show a negative values indicates less information loss than a positive values. Comparison between stationary and non-stationary models shows that, in overall, the AIC and BIC values are smaller for stationary than non-stationary therefore favor the stationary model in modelling the tree-based threshold exceedances. This conclude that regression tree method are able to produce most stationary data within the cluster, hence simpler model can be fit into the threshold exceedances.

Now we compare the performance of tree-based threshold method using stationary model with the standard method which is based on a constant high threshold u. For a standard method, value of threshold is fixed at 95^{th} percentile of the simulated data. Threshold exceedances for standard method are modelled using GPD with the covariate models incorporated into the scale parameter as given in Eq. (8) and Eq. (9). The Root

Table 3. The AIC and BIC values for stationary GP and non-stationary GP models.

Data Set	Exceedances Model	AIC	BIC
GEVLP	Stationary	909.778	916.293
	Non-stationary	912.913	925.943
GEVLN	Stationary	-147.016	-140.341
	Non-stationary	-145.123	-131.773
GEVCP	Stationary	849.858	856.495
	Non-stationary	852.049	868.640
GEVCN	Stationary	-102.817	-95.923
	Non-stationary	-110.466	-93.232

Mean Squared Error (RMSE) and coefficient of determination (R^2) values are used to compare the performance of both methods. Result in Table 4 shows that the RMSE values for tree-based threshold method are smaller compared to the standard method except for GEVLP. However, the value of RMSE for tree-based threshold method is quite close to the RMSE for standard method with covariates model indicating that these two methods are comparable with the advantage to the tree-based threshold method because of less parameter has to be estimated. Moreover, the R^2 values for tree-based threshold selection method is closer to 1 compared to standard method.

Table 4. The RMSE and R^2 values for threshold exceedances model.

Data set	Exceedances Model	RMSE	R^2
GEVLP	Tree-based	3.520	0.960
	Standard	2.171	0.971
GEVLN	Tree-based	0.287	0.998
	Standard	0.647	0.953
GEVCP	Tree-based	3.700	0.973
	Standard	4.890	0.910
GEVCN	Tree-based	0.370	0.999
	Standard	0.516	0.933

5 API Data Analysis

In this section, the proposed tree-based threshold selection method is applied to daily maxima of API and covariates PM_{10}, O_3, SO_2, NO_2 and CO data as described in Section 2. Table 5 shows the percentage of missing values for the data sets. As the percentage of missing values in several covariate are quite high, we use the Full Conditional Specification technique to impute the missing data. In this technique, each incomplete variable is imputed by a separate model. The imputation is done using `mice` package in `R` software and the algorithm is completely discussed in [15]. The descriptive statistics for API data and the covariates after the imputation is given in Table 6. From Table 6, the highest API recorded during the study period is 257 which falls in very unhealthy level.

We develop the regression tree for API and covariates data using procedure discussed in Section 3.2 with $\delta = 0.000083$.

Table 5. Percentage of missing values for API and covariates data.

Variable	API	PM_{10}	O_3	SO_2	NO_2	CO
Percentage	0.16	0.68	6.04	20.53	20.58	24.91

Table 6. Descriptive statistics for API and covariates data.

Variable	Mean	Median	Minimum	Maximum	Variance
API	56.739	54	18	257	39.244
PM_{10}	51.084	45.306	9.129	389.77	842.942
O_3	0.032	0.024	0	0.131	0.0009
SO_2	0.004	0.003	0	0.358	5.628×10^{-5}
NO_2	0.028	0.027	0	0.202	0.0001
CO	1.289	1.226	0.025	7.412	0.243

The regression tree shown in Figure 2 produce 71 clusters with O_3 and PM_{10} become a dominant covariates that split the API data. For each resulted clusters, we set 95^{th} percentile threshold producing a covariate-varying threshold known as tree-based threshold as shown in Figure 3. The exceedances of tree-based threshold is then modelled by both stationary and non-stationary GP model. Since PM_{10} and O_3 are the dominant pollutants that split the API data, we consider these pollutants in modelling the tree-based threshold excesses of API data for non-stationary GP model. Study by [8] also shows that PM_{10} and O_3 are the dominants pollutants that affect the variation in API data. The covariates model is incorporated in the scale parameter of GP model such that $\sigma = \exp(\sigma_0 + \sigma_1 O_3 + \sigma_2 PM_{10})$ where the exponential function is used to ensure that the positivity of σ is respected for all values of PM_{10} and O_3. Table 7 shows the parameter estimates of stationary and non-stationary GP model fitted to the tree-based threshold excesses API data. From Table 7, both models have positive shape parameter indicates that the distribution of tree-based threshold exceedances for API data is unbounded. The performance of the stationary and non-stationary GP model fitted to the tree-based threshold exceedances are evaluated using AIC and BIC. The AIC and BIC values are shown in Table 8. Based on Table 8, the AIC and BIC values are smaller for stationary GP model compared to non-stationary model, which conclude that modelling tree-based threshold excess with stationary GP model produce less information loss. Hence, this method provide much simpler model to explain the variation in API data. The goodness-of-fit (GoF) of the stationary GP model is tested using Anderson-Darling (AD) test and Cramer von Misses (CVM) test. The p-values of the GoF tests shown in Table 9 indicates that the stationary GP model fit the tree-based threshold exceedances of API data well.

Table 7. Parameter estimates of stationary and non-stationary GP model fitted to the tree-based threshold exceedances of API data.

Exceedances Model	σ_0	σ_1	σ_2	ξ
Stationary	3.313380	-	-	0.276766
Non-stationary	1.109091	-0.129322	0.002481	0.224365

We also compare the tree-based threshold selection method with the standard method using RMSE and R^2. A constant threshold is set at 95^{th} percentile for standard method. Results in Table 10 shows that RMSE value for tree-based threshold method are smaller than the RMSE value using standard method indicating that the tree-based threshold method pro-

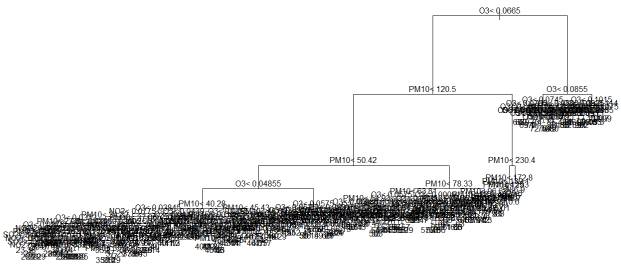

Figure 2. Regression tree for API data.

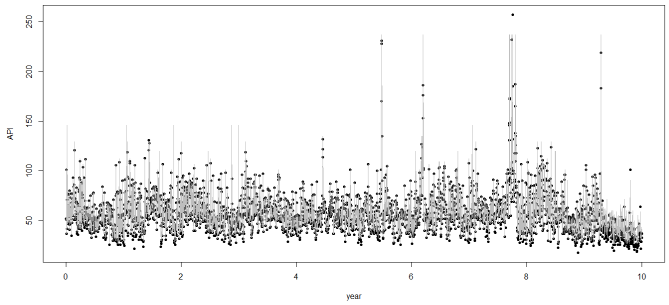

Figure 3. API data and tree-based threshold line obtained from regression tree.

Table 8. AIC and BIC values fitted to the tree-based threshold exceedances of API data.

Exceedances Model	AIC	BIC
Stationary	325.732	330.0808
Non-stationary	328.552	337.2496

Table 9. The p-values of AD and CVM tests for stationary GP model.

AD test	CVM test
0.8686	0.8513

duces less error and better at forecasting predicted values. This result also supported by value of R^2 which is much closer to 1 compared to the R^2 value for standard method.

Table 10. RMSE and R^2 values applied on API data.

Method	RMSE	R^2
Tree-based	5.262	0.995
Standard	163.258	0.919

6 Conclusion

In this paper, a new and simple method for threshold selection for the GPD in the presence of covariate is presented. The method uses regression tree to partition the data sets into approximately stationary series. The excesses of the tree-based threshold is shown to be better fitted with stationary GP model compared to non-stationary GP model producing much simpler model to explain the variation in data sets. Comparison made with the standard method shows that the proposed tree-based threshold is much better in terms of producing less error and better at forecasting values. In modelling the API data, the tree-based threshold is sufficiently enough to produce a stationary threshold exceedances so that much simpler model could be fitted in order to explain the variation in API data. In practice, our method can be seen as an additional tool that complements existing threshold selection methods.

Acknowledgment

The authors would like to thank the Department of Environment, Malaysia for providing the air quality data. This work was funded by the Geran Putra - Inisiatif Pensyarah Muda, Universiti Putra Malaysia (GP-IPM/2016/9513100)

REFERENCES

[1] Abdullah, M. Z. & Alias, N. A. (2018). Variation of PM_{10} and heavy metals concentration of suburban area caused by haze episode. *Malaysian Journal of Analytical Sciences*, 22(3):508–513.

[2] Al-Dhurafi, N. A., Masseran, N., Zamzuri, Z. H., & Razali, A. M. (2018). Modeling unhealthy air pollution index using a peaks-over-threshold method. *Environmental Engineering Science*, 35(2), 101-110.

[3] Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). *An introduction to statistical modeling of extreme values* (Vol. 208). London: Springer.

[4] Roberts, E. M. (1979a). Review of statistics of extreme values with applications to air quality data, Part I, Review. *Journal of the Air Pollution Control Association*, 29: 632-637.

[5] Roberts, E. M. (1979b). Review of statistics of extreme values with applications to air quality data, Part II, Applications. *Journal of the Air Pollution Control Association*, 29: 733-740.

[6] Amin, N. A. M., Adam, M. B., & Aris, A. Z. (2015). Extreme value analysis for modeling high PM_10 level in Johor Bahru. *Jurnal Teknologi*, 76(1).

[7] Masseran, N., Razali, A. M., Ibrahim, K., & Latif, M. T. (2016). Modeling air quality in main cities of peninsular malaysia by using a generalized pareto model. *Environmental monitoring and assessment*, 188(1):65.

[8] Al-Dhurafi, N. A., Masseran, N., Zamzuri, Z. H., & Razali, A. M. (2018). Modeling unhealthy air pollution index using a peaks-over-threshold method. *Environmental Engineering Science*, 35(2):101–110.

[9] Northrop, P. J., & Jonathan, P. (2011). Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights. *Environmetrics*, 22(7), 799-809.

[10] Ling, H. L. O., Musthafa, S. N. A. M., & Rasam, A. R. A. (2014). Urban environmental heath: Respiratory infection and urban factors in urban growth corridor of Petaling Jaya, Shah Alam and Klang, Malaysia. *Sains Malaysiana*, 43(9), 1405-1414.

[11] Pickands III, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, 3(1), 119-131.

[12] Davison, A. C., & Smith, R. L. (1990). Models for exceedances over high thresholds. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3), 393-425.

[13] Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees*. London: Chapman and Hall.

[14] Eastoe, E. F., & Tawn, J. A. (2009). Modelling non-stationary extremes with application to surface level ozone. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 58(1), 25-45.

[15] Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of statistical software*, 1-68.