Journal of Mathematics and Statistics 9 (4): 357-366, 2013
ISSN: 1549-3644

doi:10.3844/jmssp.2013.357.366 Published Online 9 (4) 2013 (http://www.thescipub.com/jmss.toc)

IDENTIFICATION AND CHARACTERIZATION OF EXTREME RAINFALLS DISTRIBUTION IN MALANG RESIDENCE

[1]Amran, [2]Iriawan Nur, [3]Subiono and [2]Irhamah

[1]Department of Statistics, Faculty of Mathematics and Natural Science,
Institute Teknologi Sepuluh Nopember, Surabaya, Indonesia and
Department of Mathematics, Faculty of Mathematics and Natural Science,
Hasanuddin University, Makassar, Indonesia
[2]Department of Statistics, Faculty of Mathematics and Natural Science,
Institute Teknologi Sepuluh Nopember, Surabaya, Indonesia
[3]Department of Mathematics, Faculty of Mathematics and Natural Science,
Institute Teknologi Sepuluh Nopember, Surabaya, Indonesia

Received 2013-04-25; Revised 2013-10-21; Accepted 2013-11-30

ABSTRACT

Extreme rainfalls often occur everywhere just in a moment, very difficult to be anticipated and produce very detrimental impact to the environment and human society. Floods and landslides are influenced by high variability of extreme rainfalls, especially in the watershed area for floods and the hills as well as mountains for landslides, such as in Malang Residence, East Java, Indonesia as a case study in this study. The prediction tools for determining location and time of the next extreme rainfalls event will occur are required. The behavior of extreme rainfalls measured on one or several stations rain gauge could be approximated by Generalized Pareto (GP) Distribution. The prediction tools must be able to identify and characterize parameters of the GP Distribution such as shape and scale parameters over the entire area. Shape parameter of GP distribution has associated with characteristics of extreme rainfalls distributions. To identify characteristics of shape parameter on each station and their similarity, an algorithm to make a partition of shape parameters into several spatial clusters and investigate the type of distribution was proposed. In order to determine threshold value, mean residual life plot and stability of modified scale and shape parameters at a range of thresholds were used, Maximum Likelihood method was utilized to estimate parameter value and k-means method combined by Silhouette values to make the cluster of extreme rainfalls distribution. By using rainfalls data on twenty eight different stations rain gauge, the results showed that the proposed algorithm well performed and extreme rainfalls were heterogeneous with three type of GP distribution. In general, shape parameter values were negative and positive except on nine stations which were close to zero and were well partitioned by six clusters.

Keywords: Extreme Rainfalls, Generalized Pareto Distribution, Shape Parameter, k-Means Algorithm, Silhouette Value

1. INTRODUCTION

Extreme rainfalls often occur everywhere just in a moment, very difficult to be anticipated and produce very detrimental impact to the environment and human society. The amount of negative impacts of floods and landslides caused by extreme rainfalls requires expert forecasting tool in local-scale

Corresponding Author: Amran, Department of Statistics, Faculty of Mathematics and Natural Science, Institute Teknologi Sepuluh Nopember, Surabaya, Indonesia and Department of Mathematics, Faculty of Mathematics and Natural Science, Hasanuddin University, Makassar, Indonesia

precipitation in order to anticipate or mitigate losses that may occur (Bermudez and Kotz, 2010; Buishand *et al*., 2008; Muller *et al*., 2009). Several researches in last decade years have been conducted their research on characterizing and modeling the large-scale temporal (annual and seasonal) of extreme rainfalls. They use variety of statistical methods for analyzing extreme rainfalls data, such as linear regression analysis, non-stationary frequency analysis (Tramblay *et al*., 2013), Bayesian approach, time series analysis (Tularam and Ilahee, 2010) and semi-parametric and parametric method (AghaKouchak and Nasrollahi, 2010), Peak over threshold method using GP distribution (Li *et al*., 2005).

Numerous studies have published. Commonly they use probability density function to describe the frequency distribution of extreme rainfalls. Among the conclusions were obtained, generally employing the GP distribution (AghaKouchak and Nasrollahi, 2010; Diebolt *et al*., 2008; Fawcett and Walshaw, 2007; Li *et al*., 2005) that can describe well the annual extreme rainfalls in some dependent locations (Coles, 2001). However, the GP distribution well performed only in homogeneous locations of observations.

Extreme rainfalls were defined by the magnitude of rainfalls which greater than a certain threshold value (AghaKouchak and Nasrollahi, 2010; Behrens *et al*., 2004; Li *et al*., 2005). These extreme rainfalls were might caused the volume of river water abundance and rapid soil saturation point will be fast achieved. These conditions make rivers and ground into a dangerous and vulnerable to floods and landslides.

In Malang residence, East Java, Indonesia, where the average rainfalls are relatively high at each year and its geographical conditions of the hilly produce fertile soil for farming and agriculture. This area is famous for its abundant agricultural products which attracts tourism. However, conditions of vulnerability to floods and landslides are also a serious threat. Hilly areas with cliffs that are cheated by high rainfalls and watershed conditions were often experienced deposition, contributing significantly to the occurrence of floods or landslides.

In order to anticipate possible losses caused by floods and landslides, the first step should be done is by developing tools or methods of predicting extreme rainfalls locally. First of all, identification and characterization of extreme rainfalls distribution, especially their hidden spatial patterns over the entire area, are required.

Hidden spatial patterns of extreme rainfalls are important to explore behavior of extreme rainfalls in the area of observations. If there are more than one pattern in such area of investigation (heterogeneous area), then the identification and characterization of extreme rainfalls distribution procedure to find homogeneous sub-areas with their own spatial pattern or characteristic are needed. These homogeneous sub-areas would help practitioner easy to use statistical tools appropriated with pattern/characteristic of each sub-area as well as to provide more robust estimator of GP distribution.

In this research, we identify rainfalls extreme distribution for each station rain gauge. Furthermore, this study will emphasize dissimilarity of distribution especially shape parameter of distribution which representative the quantity of extreme rainfalls. Parameters with dissimilarity values will be put in to different cluster and the others in same cluster.

We use twenty eight stations rain gauge measurement at different locations and daily observations for ten years. Extreme rainfalls data are determined based on an adequate threshold values on each station. These extreme rainfalls on each station are fitted to GP distribution. Due to different variability of rainfalls in each station, shape parameter values of GP distribution would be vary and could be clustered. Using k-means algorithm coupled with silhouette values, the most representative number of clusters representing the heterogeneous of shape parameters and the characteristic of shape parameters on each cluster could be established. The proposed method has two advantages. First, the distribution of rainfalls data based on the location of the observation can be modeled without considering extreme data in other years because of the independence between extreme rainfalls in each year. Second, this method is parametric because it is based on the hypothesis that distribution of extreme rainfalls data is GP Distribution.

Heterogeneous area of observation often found in the real data. The prior analysis of such area is required to employ many statistics methods further especially partition over the entire area become some sub areas or clusters with homogeneous characteristics. The proposed method will identify well and could capture the characteristics of heterogeneous extreme rainfalls through different characteristics on each cluster that arise in the area of observations.

2. MATERIALS AND METHODS

2.1. Data Sets

Extreme rainfalls data were measured at twenty eight different locations in the Malang Residences. These locations have different heights from sea level. Rainfalls data were measured daily and extreme rainfalls were determined by selecting threshold value. The number of

5479 data was used in this research which was measured during 1996 to 2010. The data therefore would be identified and characterized their distribution.

The description of the data set which is provided in **Table 1** shows that skewness values greater than zero on each station. It leads to asymmetric distribution. The assumption of Gaussian distribution, therefore, is not appropriate.

2.2. Probability Density Function

In order to construct a forecasting tool, identification and characterization of rainfalls distribution are needed. Identification and characterization distribution can be done by probability plot, quantile plot, return level plot and density plot.

The sample distribution of extreme rainfalls can classify different cluster by identifying their shape parameter. There are many kind of shape parameter which suitable with the sample data and then a classification algorithm is used on selected samples to obtain the cluster distribution of extreme rainfalls. Variability of shape parameter in each cluster can be treated as random variable distribution.

2.3. Generalized Pareto Distribution

GP distribution is a right skewed distribution and has three parameters, called shape parameter (ξ), location parameter (μ) and scale parameter (σ). Shape parameter represents the tail index that could be positive, zero, or negative.

Let X be a random variable, which follow GP distribution which has Cumulative Distribution Function (Bermudez and Kotz, 2010):

$$f(x:\xi,\mu,\sigma)=\begin{cases}1-\left(1-\xi\dfrac{x-\mu}{\sigma}\right)^{1/\xi}, & \xi\neq 0\\ 1-\exp\left(-\dfrac{x-\mu}{\sigma}\right), & \xi=0\end{cases} \quad (1)$$

The range is $\mu\leq x<\infty$, if $\xi\leq 0$ and $\mu\leq x\leq \mu+\sigma/\xi$, if $\xi>0$. Where the distribution has a finite upper bound values, $\mu+\sigma/\xi$.

The probability distribution function can be derived from Equation 1 as follows:

$$f(x:\xi,\mu,\sigma)=\begin{cases}\dfrac{1}{\sigma}\left(1-\xi\dfrac{x-\mu}{\sigma}\right)^{\frac{1}{\xi}-1}, & \xi\neq 0\\ \dfrac{1}{\sigma}\exp\left(-\dfrac{x-\mu}{\sigma}\right), & \xi=0\end{cases} \quad (2)$$

Based on threshold value which is obtained from mean residual life plot and stability parameters plot on each station, therefore we can estimate ξ and σ values in Equation 2. Maximum Likelihood Estimation (MLE) method was choosen to estimate ξ and σ due to the number of observations of rainfalls in Malang residence are large enough. MLE would be efficient, if the sample size is large (Hosking and Wallis, 1987; Li *et al*., 2005). Suppose that y is a corrected x by threshold value (μ), the logarithm natural likelihood would be:

$$\text{In } L(\xi,\sigma,y)=\begin{cases}-n\,\text{In}\sigma+\left(\dfrac{1}{\xi}-1\right)\sum_{i=1}^{n}\text{In}\left(1-\dfrac{\xi y_i}{\sigma}\right), & \xi\neq 0\\ -n\,\text{In}\,\sigma=\dfrac{1}{\sigma}\sum_{i=1}^{n}yi, & \xi=0\end{cases} \quad (3)$$

where, $y_i = x_i-\mu$, known as excess.

The characteristics of family of GP can be explained as follows: if $\xi<0$ then the GP distributions has a point until the end that is known as short-tailed, especially when $\xi = -1$, GP behave as uniform distribution. While, if $\xi = 0$, GP behave as exponential distribution. On the other hand if $\xi>0$ GP distribution known as Pareto distribution classified as heavy-tailed distribution. Different of ξ would represent heavy-tailness of GP adaptively to the data, which applicable for rainfalls modeling. It means that extreme rainfalls can be captured by heavy-tailness of GP distributions.

The mean and variance of the GP distributions respectively given as:

$$E(x)=\frac{\sigma}{1+\xi};\xi>-1 \quad (4)$$

And:

$$\text{Var}(x)=\frac{\sigma^2}{(1+\xi)^2(1+2\xi)};\xi>-\frac{1}{2} \quad (5)$$

In Equation 4 and 5, expectation and variance are not stable since depend on ξ and σ values. In contrast, GP distribution is stable on threshold. This property will guarantees that if the observational data follow GP distribution, then the data of which exceeds the threshold, are still GP distributed (Jockovic, 2012). Excess values (y_i) in Equation 3, therefore, would be GP distributed as well (Falk and Guillou, 2008).

Table 1. Description of daily rainfalls at 28 stations rain gauge, in Malang residence

Stations\statistics	Mean	Variance	Skewness	Kurtosis
Bantur	5,153	262,007	6,05	60,69
Blimbing	5,829	194,141	4,00	24,41
Bululawang	5,596	178,414	3,92	22,07
Dampit	5,500	218,801	4,55	30,18
Dau	4,730	145,895	3,89	18,51
Jabung	5,931	174,817	3,44	16,93
Jombok	7,275	267,080	3,93	22,87
Kantor CD	6,256	257,044	4,75	36,93
Karang suko	4,987	169,861	4,22	26,53
Kedungkandang	5,638	181,485	3,94	22,95
Kedungrejo	5,829	163,577	3,77	23,33
Ngaglik	4,411	116,799	3,98	22,06
Ngajum	5,731	196,941	3,86	19,39
Ngantang	9,102	386,717	3,54	17,94
Ngujung	4,653	122,430	3,62	17,05
Penarukan	5,894	216,747	4,69	35,12
Pendem	4,713	138,048	3,71	16,58
Pohgajih	6,189	200,063	3,86	22,45
Poncokusumo	6,788	173,194	2,66	10,26
Pujon	6,044	177,691	3,67	21,14
Sekar	7,716	256,563	3,15	12,67
Sumber pucung	4,654	141,936	4,29	27,06
Tajinan	5,710	186,611	3,37	13,87
Temas	4,483	111,964	3,60	16,84
Tinjumoyo	4,887	136,951	3,70	19,25
Tlekung	4,174	115,521	4,00	20,50
Tumpukrenteng	5,888	219,137	3,83	18,03
Turen	5,897	217,237	4,16	23,98

2.4. The Proposed Algorithm

In this research, we offer an algorithm to identify and characterize the distribution of extreme rainfalls over the entire observation area, given rainfalls data. The algorithm has three main stages that perform sequentially, that are threshold selection, fitting GP distribution and then partitioning the shape parameter of GP.

In the threshold selection stage, we use mean residual life plot as well as stability of modified scale and shape parameters across a range of different thresholds to determine the threshold value. An adequate threshold would distinguish rainfalls as extreme or not. Extreme rainfalls are rainfalls with its quantity greater than threshold value.

Furthermore, in the fitting distribution stage, the extreme rainfalls data would be fitted by GP distribution (Mackay *et al*., 2011). In this stage the ability to identify ξ parameter in each location is most important. The more vary and extreme rainfalls with high quantity, the higher ξ value that represent heavy tail GP distribution.

Finally, in the end stage, collection of ξ values from different locations which were obtained in the second stages has to be partitioned in to several clusters according to their own characteristics. Moreover, k-means method coupled with silhouette algorithm will be employed to identify the most appropriate number of clusters.

The k-means method, will classify ξ values into k clusters. Dissimilarity of ξ values, represented as distance among ξ values, will be used to put its ξ values into different clusters. The closer among ξ values, the more similar the ξ values which tend to put into the same partition. The k-means algorithm is run iteratively to minimizes the sum of distances among its ξ to cluster centroid. It has to be done for all clusters. The smaller of total sums of distances, the better partition would be found.

One crucial step in k-means is to determine how many clusters which fit to the ξ data. Silhouette algorithm which has ability to determine the number of clusters precisely would be used here to overcome the crucial step above. It is due to Silhouette algorithm employ such measurement other than total sums of distances which more sensitive to the characteristics of members of clusters. The Silhouette algorithm, S(i), is defined by Equation 6 as:

$$S(i) = \frac{\min\left(b(i,:),2\right) - a(i)}{\max(a(i), \min\left(b(i,:)\right))} \quad (6)$$

where, a(i) is the average distance from i-th member to others member in one cluster and b(i, k) is the average distance from the i-th member to members in another cluster k. Range of silhouette values lay between -1 to +1. If distances both win member of clusters to other clusters is close to +1 then they are categorize distinctly. On the other hand, it could not be distinctly categorized when their distances is close to zero. Meanwhile, if silhouette values close to -1, then there is probably assigned to the wrong clusters.

The silhouette plot of S(i) value describes similarity inter cluster and dissimilarity between clusters visually. The plot represented by collected of bars, provides information about how well-separated the resulting of k-means algorithm. If most of bars plot in each cluster close to one, then it shows that its clusters are well-separated. Otherwise, members of its clusters are not separated distinctly when bars plot close to zero. There are probably wrong assigned, on the other hand, when bars plot close to -1.

The proposed algorithm is described as follows:

- Collect daily rainfalls data on each station
- Select the threshold value of serial rainfalls data based on Mean residual life plot and stability of parameters plot.
- Identify which rainfalls that greather than threshold value, categorize all extreme rainfalls on each station and construct as an extreme rainfalls series
- Fit GP distribution to the extreme rainfalls series by estimating ξ and σ parameters using maximum likelihood method on each station
- Determine certain number of cluster and partition ξ values into those clusters using k-means method coupled with silhouette algorithm
- Calculate mean silhouette values of optimal clusters and the total sums of distances
- Repeat step 5 and step 6 for different number of clusters to find optimal number of cluster based on maximum mean silhouette values of optimal clusters and the total sums of distances. Otherwise, use silhouette when total sums of distances fail to characterize the optimal number of clusters
- Characterize the distribution of each cluster by investigating the shape of GP distribution (bounded tail, light tail, or heavy tail)

3. RESULTS

We perform maximum likelihood method to fit data with GP distribution on each station rain gauge. Especially, in estimating ξ and σ parameters, the threshold value is choosen based on mean residual life plot and stability of parameters plot. Nelder-mead method was used to calculate maximum likelihood estimator iteratively to reach maximum values of the likelihood function. Moreover, the diagnostic plots consist of probability plot, quantile plot, return level plot and density plot were employed to asses efficiency or quality of MLE (**Table 2**).

Generally, plots on each data series at twenty eight stations rain gauge illustrate that model which was obtained by MLE fitted the data well. Example of diagnostic plots at Jabung station as shown in **Fig. 1 and Fig. 2.** In **Fig. 1a**, probability plot consist of model of GP distribution in the vertical axes and extreme rainfall data in horizontal axes, the plot showed that model and data tend to be putted into linear line, the similar result was obtained for quantile plot in **Fig. 1b**. Return level plot in **Fig. 2a** display return level on vertical axes and return period on horizontal axes, for negative shape parameter value, return level plot would be convex, return level plot would be linear when shape parameter values close to zero and concave for positive shape parameter values. Shape parameter value at Jabung station is 0.045 (close to zero) and the return level plot is linear. Density plot in **Fig. 2b** consist of density model (blue line) and histogram of data, represent compatibility between model and histogram of data.

The results show that all shape parameters on each station were convergen and vary on each station in range of (-0.237, 0.174). Normality test for shape parameter values were performed using Kolmogorov-Smirnov test. The result showed that distribution of shape parameter values can be fitted by Gaussian distribution (P-value >0.150). The k-means algorithm coupled with silhouette values, therefore, can be employed for clasifying shape parameter values.

Table 3 showed total sums of distances and mean silhouette values for two to ten number of clusters. The total sums of distances decreased, while the number of clusters increased and maximum mean silhouette value was found for six number of clusters.

Based on the number of clusters which was obtained by silhouette values in **Table 3**, then we investigate members of clusters. Members on each cluster in Malang Residence therefore, can be determined using k-means algorithm. The results are shown in **Table 4**. All stations were divided into six different clusters which indicate heterogeneous characteristics of extreme rainfalls distribution. In general, extreme rainfalls distribution in Malang residence has negative and positive shape parameter values, except on nine stations in the fifth cluster which are close to zero.

Figure 3 showed the results obtained by Silhouette plots for ξ values in six clusters. As it can be observed, most of the silhouette values are greater than 0.6, although there are two members with low silhouette values.

The detail of spatial characteristics in Malang residence on each cluster is described as follows.

Cluster 1

The first cluster consists of Sumber pucung, Pujon, Kedungkandang, Pendem and Ngaglik stations. All of the shape parameter values on each member of cluster are negative.

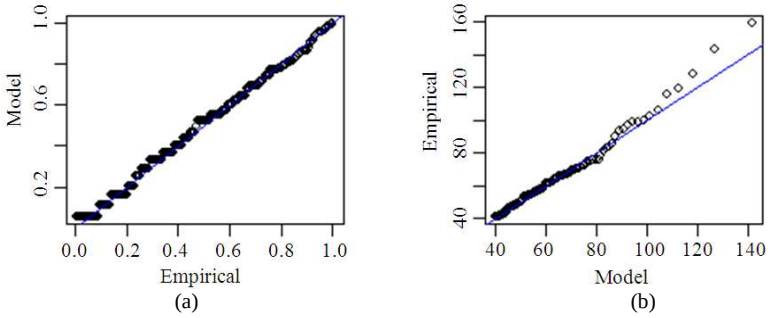

Fig. 1. Diagnostic plots at Jabung station consist of: (a) robability plot, (b). Quantile plot

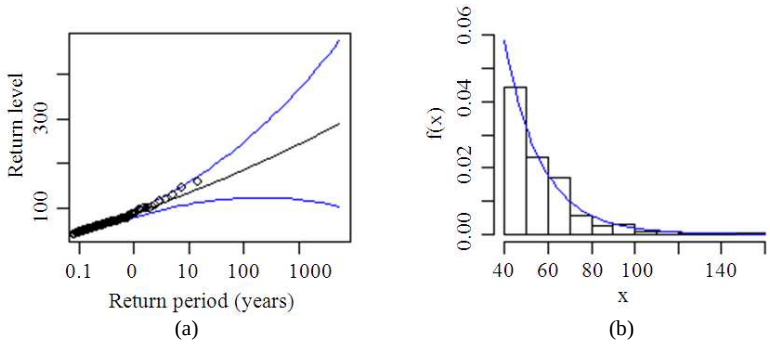

Fig. 2. Diagnostic plots at Jabung station consist of: (a) Return level plot, (b) Density plot

Table 2. The results of parameters estimation by MLE method

Stations	ξ	σ
Bantur	0.174	25.070
Blimbing	0.007	16.425
Bululawang	0.027	19.386
Dampit	0.014	25.565
Dau	-0.237	20.971
Jabung	0.045	17.145
Jombok	0.101	22.042
Kantor dinas	0.132	24.184
Karang suko	0.020	20.601
Kedungkandang	-0.104	27.198
Kedungrejo	0.169	17.856
Ngaglik	-0.094	18.604
Ngajum	-0.011	21.475
Ngantang	0.004	27.557
Ngujung	-0.200	22.903
Penarukan	0.107	23.569
Pendem	-0.117	17.951
Pohgajih	0.099	18.269
Poncokusumo	0.133	11.460
Pujon	-0.076	20.510
Sekar	-0.218	37.591
Sumber pucung	-0.063	19.119
Tajinan	0.144	15.036
Temas	0.021	15.387
Tinjumoyo	-0.211	19.205
Tlekung	0.104	17.566
Tumpukrenteng	-0.166	20.946
Turen	-0.012	24.075

Table 3. Total sums of distances and Silhouette mean values

Number of Cluster	Total sums of distances	Mean Silhouette Values
2	1.616	0.5552
3	0.874	0.6116
4	0.505	0.7189
5	0.385	0.6870
6	0.324	0.7409
7	0.298	0.6450
8	0.241	0.6235
9	0.217	0.6235
10	0.185	0.6514

Cluster 2

Tumpukrenteng, Sekar, Tinjumoyo, Ngujung and Dau stations which are grouped in the second cluster, have negative shape parameter same with cluster one, but shape parameter values in this cluster less than shape parameter values in cluster one. Characteristics of extreme rainfalls of the second cluster are same with characteristis of the first cluster.

Cluster 3

Members of the third cluster which are Tlekung, Pohgajih, Penarukan and Jombok, have positive shape parameter values around 0.1, thus they have heavy tail characteristics.

Table 4. The results of k-means algorithm

Cluster	Number of Members	Stations member (Stations name and shape parameter values)			
1	5	Sumberpucung	(-0.063)	Pendem	(-0.117)
		Pujon	(-0.076)	Ngaglik	(-0.094)
		Kedungkandang	(-0.104)		
2	5	Tumpukrenteng	(-0.166)	Ngujung	(-0.200)
		Sekar	(-0.218)	Dau	(-0.237)
		Tinjumoyo	(-0.211)		
3	4	Tlekung	(0.104)	Penarukan	(0.107)
		Pohgajih	(0.099)	Jombok	(0.101)
4	3	Tajinan	(0.144)	Kantor CD	(0.132)
		Poncokusumo	(0.133)		
5	9	Turen	(-0.012)	Karangsuko	(0.020)
		Temas	(0.021)	Jabung	(0.045)
		Blimbing	(0.007)	Dampit	(0.014)
		Ngantang	(0.004)	Bululawang	(0.027)
		Ngajum	(-0.011)		
6	2	Kedungrejo	(0.169)	Bantur	(0.174)

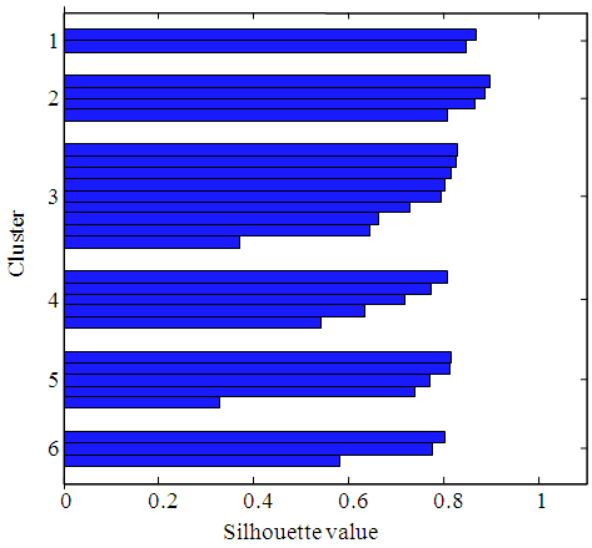

Fig. 3. Silhouette plots for six clusters

Cluster 4

Tajinan, Poncokusumo and Kantor CD are members of the fourth cluster. Although, shape parameter values in this cluster close to shape parameter values in third cluster but, its values are greater than shape parameter values of cluster three. However, their characteristics are similar.

Cluster 5

The fifth cluster was characterized by negative and positive shape parameter values which are close to zero. Turen, Temas, Blimbing, Ngantang, Ngajum, Karangsuko, Jabung, Dampit and Bululawang stations are members of the fifth cluster.

Cluster 6

Cluster sixth was identified by positive shape parameter value. Shape parameter values in this cluster are the greatest of all parameter values over entire area of observation. The characteristic of cluster sixth are similar with the third and the fourth clusters.

4. DISCUSSION

Identification and characterization of extreme rainfalls proposed in this study, is similar with regional frequency analysis by Yang *et al*. (2010) especially in the identification of homogeneous regions step by clustering analysis. This clustering procedure is used to utilize full advantage of information in different data series within homogeneous clusters, in order to obtain more robust estimator. Yang *et al*. (2010) employed average linkage method and ward's method to identify homogeneous regions. Those both methods, however, tend to bias toward globular cluster. To cluster shape parameter values of extreme rainfalls distribution in Malang residence, this study use k-means algorithm coupled with silhouette values. The advantages of k-means algorithm is computationally faster and can produce tighter clusters. After running average linkage and ward's method compare to k-means algorithm on six clusters, the results showed that on each cluster, variance of average linkage or ward's method were

greater than variance of k-means algorithm and total of variance of average linkage or ward's method = 9×10^{-3}, while k-means algorithm = 2×10^{-3}. It means that k-means algorithm produced tighter clusters than average linkage or ward's method.

This paper have succeded to clasify homogeneous characteristics clusters of extreme rainfalls distribution in Malang Residence, Indonesia, through a clustering algorithm of shape parameter of GP distribution. The proposed algorithm can be used to make partition of heterogeneous area of observations into some sub-areas with homogeneous characteristic. These findings are important to employ many statistical tools appropriate with their own characteristic.

Parameters estimation using MLE method have been used by some researchers. For example, Ahmed *et al*. (2010) employed MLE for parameters of weibull distribution, Al-Athari (2011) estimate parameters of double pareto by MLE, using MLE to estimate parameters of Logistic regression. Meanwhile, parameters estimation of GP distribution using MLE, have been used by Chaouche and Bacro (2006); Husler *et al*. (2011); Li *et al*. (2005); Mackay *et al*. (2011) and Zhao (2010).

The quality of MLE for the extreme rainfalls distribution in Malang residence, were measured by diagnostic plots. **Fig. 1 and 2** displayed model and data of extreme rainfall at Jabung station. In **Fig. 1a and 1b**, model and data were putted into linear line. It showed that model are fitted with data. **Figure 2a** showed that return level line is linear, it is appropriate with shape parameter estimator which close to zero. In **Fig. 2b**, model showed concistency with the histogram. Based on these plots, model fits data well at Jabung station. The similar results were obtained on other stations.

Table 2 shows the estimate values were obtained by MLE method. It can be seen that the shape parameter values were vary on each station, their values were less than 0.5 which are valid for using maximum likelihood method (Bermudez and Kotz, 2010; Castillo and Daoudi, 2009) and their values were greater than -0.5 which are concistency, asymptotic efficiency and asymptotic normality (Zhao, 2010). It also indicates that extreme rainfalls are heterogeneous in Malang Residence. Thus, it was recommended to make a partition over the entire area based on shape parameter values.

Furthermore, we only focus on the characteristics of shape parameters of extreme rainfalls distribution, while scale parameter cannot describe the extreme rainfalls characteristics. Shape parameter values which are greater than zero show that the distribution does not have upper bound. Negative shape parameter value denotes that the distribution has an upper bound, it can be interpreted that the magnitude of extreme rainfalls never beyond upper bound. The distribution is unbounded if shape parameter values is zero (Coles, 2001). According to the results in **Table 2**, we performed k-means algorithm and using two criteria that are silhouette values and total sums of distances for drawing an inference.

Mean silhouette values as shown in **Table 3** suggests that ξ values can be classified into three to ten clusters since silhouette mean values greater than 0.6. The greater mean silhouette values, the more similar members of inter clusters and the higher dissimilar among clusters. The maximum mean average silhouette value is 0.7409. We conclude that the characteristics of extreme rainfalls over the entire area can be partitioned in six clusters.

In General, characteristics of extreme rainfalls in Malang residence are heterogeneous and follow three type of GP distribution. These are short tail, light tail and heavy tail distributions.

Sumber Pucung, Pujon, Kedungkandang, Pendem, Ngaglik, Tumpukrenteng, Sekar, Tinjumoyo, Ngujung and Dau stations have short-tailed distribution, since $\xi<0$. It can be interpreted that extreme rainfalls are frequently occurred in these stations and never exceed an upper bound value. Shape parameter values at Turen, Temas, Blimbing, Ngantang, Ngajum, Karangsuko, Jabung, Dampit and Bululawang are close to zero. It means that distribution of extreme rainfalls in these stations can be approximately by exponential distribution, since $\xi\approx0$. The tail of this distribution is decreased exponentially and known as light tail distribution. The characteristics of extreme rainfalls with light tail distribution can be interpreted as rare event but could be appeared in very high quantity. Tlekung, Pohgajih, Penarukan, Jombok Tajinan, Poncokusumo, Kantor CD, Kedungrejo and Bantur, have similar characteristic. The distribution of extreme rainfalls in this cluster have heavy tail since $\xi>0$. It can be approximately as t distribution. In other words, extreme rainfalls frequently occurred and the quantity of the extreme rainfalls does not have upper bound.

The homogeneous clusters are important for a practicioner to select estimation method among several existing estimation approaches which are available. Information about small or large sample size and heavy or light tail of underlying distributions in addition, are required as prior analysis for practicioner with solve Bermudez and Kotz (2010) problem statement.

5. CONCLUSION

In this study, we proposed an algorithm to identify and characterize extreme rainfalls distribution over the whole area of observation. Implementation of the algorithm using extreme rainfalls data in Malang residence showed good performance. It looks perfectly differentiated among similar extreme rainfalls into a cluster and others dissimilar extreme rainfalls into different clusters as well as it could distinguish characteristics on each cluster. The results show that extreme rainfalls in Malang residence were well partitioned by six clusters and follow three type of GP distribution. Almost all stations rain gauge have negative and positive shape parameter values known as GP distribution with short and heavy tail, except on nine stations with light tail distribution.

The extreme rainfalls represented by GP distribution with short tail were found in cluster one and cluster two. It means that there are extreme rainfalls with an upper bound value. While, the light tail one has unbounded tail, that means the extreme rainfalls are rarely happen, as found in cluster fifth. It could be inferred that with small probability the very high quantity rainfalls would happen. Stations in cluster third and cluster fourth are representing GP distribution with positive shape parameter value. The distribution of extreme rainfalls has heavy or unbounded tail. It could be interpreted that there are extreme rainfalls with frequently occurred and could achieve high quantity.

6. ACKNOWLEDGEMENT

We thank anonymous reveiwers for their valuable advices which greatly enhanced the quality of this study. This research was supported by Hibah Program Guru Besar of Institut Teknologi Sepuluh Nopember Surabaya, under contract No. 0750.256/I2.7/PM/2011. We would like to gratefully acknowledge Institut Teknologi Sepuluh Nopember, Hasanuddin University and Badan Pengelola Waduk dan Sungai (BPWS) Kabupaten Malang, for providing us the data set used in this study.

7. REFERENCES

AghaKouchak, A. and N. Nasrollahi, 2010. Semi-parametric and parametric inference of extreme value models for rainfall data. Water Resour. Manage., 24: 1229-1249. DOI: 10.1007/s11269-009-9493-3

Ahmed, A.M., H.S. Al-Kutubi and N.A. Ibrahim, 2010. Comparison of the bayesian and maximum likelihood estimation for weibull distribution. J. Math. Stat., 6: 100-104. DOI: 10.3844/jmssp.2010.100.104

Al-Athari, F.M., 2011. Paramter estimation for double pareto distribution. J. Math. Stat., 7: 289-294. DOI: 10.3844/jmssp.2011.289.294

Behrens, C.N., H.F. Lopez and D. Gamerman, 2004. Bayesian analysis of extreme events with threshold estimation. Stat. Modell., 4: 227-244. DOI: 10.1191/1471082X04st075oa

Bermudez, P.D.Z. and S. Kotz, 2010. Parameter estimation of the generalized pareto distribution-Part I. J. Stat. Plann. Inference, 140: 1353-1373. DOI: 10.1016/j.jspi.2008.11.019

Buishand, T.A., L. de Haan and C. Zhou, 2008. On spatial extremes: With application to a rainfall problem. Ann. Applied Stat., 2: 624-642. DOI: 10.1214/08-AOAS159.

Castillo, J.D. and J. Daoudi, 2009. Estimation of generalized pareto distribution. Stat. Probability Lett., 79: 684-688. DOI: 10.1016/j.spl.2008.10.021

Chaouche, A. and J.N. Bacro, 2006. Statistical inference for the generalized pareto distribution: Maximum likelihood revisited. Commun. Stat. Theory Meth., 35: 785-802. DOI: 10.1080/03610920500501429

Coles, S., 2001. An Introduction to Statistical Modeling of Extreme Values. 1st Edn., Springer, London, ISBN-10: 1852334592, pp: 208.

Diebolt, J., A. Guillou, P. Naveau and P. Ribereau, 2008. Improving probability-weighted moment methods for the generalized extreme value distribution. REVSTAT-Stat. J., 6: 33-50.

Falk, M. and A. Guillou, 2008. Peaks-over-threshold stability of multivariate generalized Pareto distributions. J. Multivariate Anal., 99: 715-734. DOI: 10.1016/j.jmva.2007.03.009

Fawcett, L. and D. Walshaw, 2007. Improved estimation for temporally clustered extremes. Environmetrics, 18: 173-188. DOI: 10.1002/env.810

Hosking, J.R. and J.R. Wallis, 1987. Parameter and quantile estimation for the Generalized pareto distribution. Technometrics, 29: 339-349. DOI: 10.2307/1269343

Husler, D., D. Li and M. Raschke, 2011. Estimation for the generalized pareto distribution using maximum likelihood and goodness of fit. Commun. Stat. Theory Meth., 40: 2500-2510. DOI: 10.1080/03610920903324874

Jockovic, J., 2012. Quantile estimation for the generalized pareto distribution with application to finance. Yugoslav J. Operat. Res., 22: 297-311. DOI: 10.2298/YJOR110308013J

Li, Y., W. Cai and E.P. Campbell, 2005. Statistical modeling of extreme rainfall in Southwest Western Australia. J. Climate, 18: 852-863. DOI: 10.1175/JCLI-3296.1

Mackay, E.B.L., P.G. Challenor and A.S. Bahaj, 2011. A comparison of estimators for the generalised Pareto distribution. Ocean Eng., 38: 1338-1346. DOI: 10.1016/j.oceaneng.2011.06.005

Muller, A., P. Arnaud, M. Lang and J. Lavabre, 2009. Uncertainties of extreme rainfall quantiles estimated by a stochastic rainfall model and by a generalized Pareto distribution. Hydrol. Sci. J., 54: 417-429. DOI: 10.1623/hysj.54.3.417

Tramblay, Y., L. Neppel, J. Carreau and K. Najib, 2013. Non-stationary frequency analysis of heavy rainfall events in southern france. Hydrol. Sci. J., 58: 1-15. DOI: 10.1080/02626667.2012.754988

Tularam, G.A. and M. Ilahee, 2010. Time series analysis of rainfall and temperature interaction in coastal catchments. J. Math. Stat., 6: 372-380. DOI: 10.3844/jmssp.2010.372.380

Yang, T., Q. Shao, Z. Hao, X. Chen and Z. Zhang *et al*., 2010. Regional frequency analysis and spatio-temporal pattern characterization of rainfall extremes in the Pearl River Basin, China. J. Hydrol., 380: 386-405. DOI: 10.1016/j.jhydrol.2009.11.013

Zhao, X., 2010. Extreme value modelling with application in finance and neonatal research. PhD Thesis, University of Canterbury.