Descriptive Statistics and Cluster Analysis for Extreme Rainfall in Java Island

K E Komalasari¹, H Pawitan² and A Faqih²

¹Postgraduate student at Bogor Agricultural University, Indonesia
²Department of Geophysics and Meteorology, Bogor Agricultural University, Indonesia

Email: kurnia.bmkg@gmail.com

Abstract. This study aims to describe regional pattern of extreme rainfall based on maximum daily rainfall for period 1983 to 2012 in Java Island. Descriptive statistics analysis was performed to obtained centralization, variation and distribution of maximum precipitation data. Mean and median are utilized to measure central tendency data while Inter Quartile Range (IQR) and standard deviation are utilized to measure variation of data. In addition, skewness and kurtosis used to obtain shape the distribution of rainfall data. Cluster analysis using squared euclidean distance and ward method is applied to perform regional grouping. Result of this study show that mean (average) of maximum daily rainfall in Java Region during period 1983-2012 is around 80-181 mm with median between 75-160 mm and standard deviation between 17 to 82. Cluster analysis produces four clusters and show that western area of Java tent to have a higher annual maxima of daily rainfall than northern area, and have more variety of annual maximum value.

Keywords: extreme rainfall, maximum daily rainfall, descriptive statistics, cluster analysis

1. Introduction
Rainfall is one element of climate has a major influence on human life, but in extreme conditions of rainfall often causes damages to human life and the natural environment. Extreme weather (including extreme rainfall) as a rare event based on the statistical distribution on a certain area, and will be specifics for different place [1].

There are two ways to identifying extreme rainfall from real data. The first approach considers the maximum or minimum of variable taking in successive periods. The second approach focuses in the appearance of value exceeding a given threshold [2]. In this study will using block maxima method for one day period to identify extreme rainfall and conducted 40 rain stations in Java Island corresponding daily rainfall observation data for period 1983-2012.

Descriptive statistics analysis was performed to analysis condition of centralization data, variation and distribution patterns of maximum daily rainfall data in Java Region. Regional clustering will generated to obtain geographical distribution of maximum of daily rainfall in Java Island.

2. Data and Methodology
Data in this study is daily rainfall data in Java Region (5°25'-8°48'S and 105°1'-114°4'E) which consists of 40 rain stations that spread across six provinces in Java Region during the period 1983-2012. Distribution map of the Station a can be seen in Figure 1.
2.1. Identification of Maximum Daily Rainfall
WMO has recommended 11 indices to identified extreme rainfall [3], and one of them is maximum daily rainfall that used in this study. Analysis are based on maximum daily rainfall from 40 stations over Java regions that identified from maximum daily rainfall amount in a year during period 1983-2012.

2.2. Descriptive Statistics Analysis
Descriptive statistics analysis performed for maximum daily rainfall during period 1983-2012. Statistical analysis used to describe the condition of centralization data, variation and distribution patterns of extreme rainfall data in Java Region. Measures of central tendency data that will used is mean (average) and median (middle value). Variation of data will measured by Inter Quartile Range (IQR) and obtained from the difference between third quartile to first quartile, standard deviation which indicates the deviation of actual data against the average formulated with:

$$Sd = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$  \hspace{1cm} (1)

The pattern of distribution data will measured by skewness and kurtosis, skewness value of normal distribution is zero, when data have extends distribution to right with a long tail to the higher value that mean the data have positive value skewness, in another hand if the data have distribution with a long tail to the lower value that mean the data has a value skewness negative, skewness value calculated by:

$$Sk = \frac{n}{(n-1)(n-2)} \sum \frac{(x_j - \bar{x})^3}{s}$$  \hspace{1cm} (2)

Kurtosis value indicates high or low data distribution curve with reference to the normal curve. Kurtosis value of normal distribution is 0, curve that higher than normal distribution has a positive
Kurtosis value, while the lower curve of the normal distribution have negative kurtosis value. Kurtosis value is calculated by the following formula.

\[ K_T = \left( \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{s} (x_j - \bar{x})^4 \right) - \frac{3(n-1)^2}{(n-2)(n-3)} \]  

(3)

**Figure 3.** Three type of distribution pattern with different kurtosis

2.3. **Principal Component Analysis**

Principal component analysis aims to avoid any multicollinearity between variables before cluster analysis. The principal component is combination of these variable that orthogonal (perpendicular) and independent each other. The principal component is formed by a sequence variance from the biggest to the smallest. If principal component \( y \) is a combination from \( x_1, x_2, x_3, \ldots, x_f \) so

\[ y = w_1x_1 + w_2x_2 + \ldots + w_px_p \]  

(4)

\( w_i \) is weights or coefficients for the variables i, \( x_i \) is the variable to i and \( y \) are linear combinations of the variables \( x \). Percentage of total diversity is considered sufficient if 75% or more is able to be explained by 4 or 5 first principal components [4].

2.4. **Cluster Analysis**

Cluster analysis is a technique for classifying objects into groups according to their particular characteristics. Each observation must have a high homogeneity in a group but high heterogeneity with other groups [5]. Cluster analysis carried out for seven first component from principal component analysis result. Squared euclidean distance is used to measure distance of each object in this study. The distance measure used in this study is squared euclidean distance is formulated as below[6]:

\[ d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + \ldots + (x_{ip} - x_{jp})^2} \]  

(5)

Cluster technique that used in this study is a ward method or also known as minimum variance method. Ward’s linkage was the most robust hierarchical clustering method in handling the wrong grouping due to noise in data [7, 8].

3. **Result**

3.1. **Descriptive Statistics**

Result of statistics descriptive analysis for 40 stations show that average of maximum daily rainfall in Java Region during period 1983-2012 is around 79.7 -181.1mm with median between 74.5-159.8mm and standard deviation between 17.2 to 81.8. Most of rainfall stations have positive kurtosis value that means distribution of data have heavy tails, and outliers. All of data have positive skewness that indicate skewness distribution curve have tail at value above the median, and mean is on the right of the peak value. List for detail can be seen in Table 1, box plots for 40 rainfall stations in Java in Figure 4.
Table 1. Station name and statistics of maximum daily rainfall for all the stations in Java Island.

| Name Station                | Mean  | Median | Kurtosis | Skewness | Standard Deviation | IQR   |
|-----------------------------|-------|--------|----------|----------|--------------------|-------|
| Stage of Bandung            | 79.7  | 76.0   | 1.9      | 1.2      | 17.6               | 19.8  |
| Cibukamanah                 | 115.6 | 112.5  | 1.3      | 0.8      | 31.5               | 37.5  |
| Sumurwatu                   | 86.7  | 75.0   | 4.7      | 1.9      | 31.2               | 26.0  |
| Staklim Bogor               | 127.9 | 119.1  | 2.6      | 1.4      | 34.4               | 38.8  |
| Cicurug                      | 97.9  | 91.0   | 11.8     | 2.8      | 45.4               | 33.8  |
| Karangkendal                | 85.3  | 82.5   | 0.5      | 0.8      | 24.4               | 38.3  |
| Bantardewa                  | 113.4 | 100.0  | 2.6      | 1.5      | 46.5               | 32.0  |
| Pacet                       | 90.2  | 92.2   | 0.3      | 0.5      | 17.6               | 23.5  |
| Leles                        | 93.6  | 79.0   | 20.8     | 4.3      | 66.6               | 34.0  |
| Nariewattie                 | 181.1 | 159.0  | 6.3      | 2.3      | 81.8               | 45.5  |
| Jatiwangi                   | 126.3 | 122.7  | 0.5      | 0.3      | 26.0               | 38.9  |
| Cimalaka                    | 80.3  | 76.5   | 2.2      | 1.3      | 18.4               | 21.5  |
| Bd.Bekasi                   | 91.3  | 89.5   | 0.3      | 0.6      | 22.0               | 31.0  |
| Linggarjati                 | 112.3 | 74.5   | 8.4      | 2.8      | 77.0               | 44.5  |
| Tng Priok                   | 122.3 | 106.7  | 0.1      | 1.0      | 40.0               | 47.1  |
| Halim                       | 130.2 | 108.8  | 4.6      | 2.1      | 52.0               | 42.8  |
| Stage of Tanggerang         | 112.5 | 97.9   | 5.8      | 2.2      | 46.8               | 40.3  |
| Pndk Betung                 | 128.3 | 109.1  | 6.3      | 2.5      | 63.7               | 33.2  |
| Budiarto Curug              | 102.2 | 89.2   | 6.5      | 2.4      | 34.6               | 25.1  |
| Serang                      | 86.7  | 84.0   | 3.6      | 1.4      | 21.9               | 24.8  |
| Semarang Klim               | 134.5 | 116.0  | 2.1      | 1.6      | 49.0               | 38.0  |
| Tegal                       | 112.8 | 99.4   | 1.5      | 1.4      | 38.8               | 34.8  |
| Cilacap                     | 173.3 | 159.8  | 0.7      | 1.0      | 64.4               | 59.4  |
| Adisucipto                  | 123.2 | 111.6  | 7.8      | 2.7      | 58.9               | 37.2  |
| Bojonegoro                  | 102.4 | 98.5   | -0.2     | 0.4      | 26.3               | 41.0  |
| Banyuwangi                  | 98.3  | 93.0   | -0.5     | 0.6      | 33.6               | 55.5  |
| Sangkapura                  | 138.5 | 121.7  | 3.0      | 1.7      | 53.4               | 54.8  |
| Kaliangat                   | 99.9  | 86.8   | 14.9     | 3.6      | 59.4               | 26.8  |
| Juanda                      | 103.7 | 100.9  | 1.8      | 1.2      | 19.8               | 25.2  |
| Stamt Perak I               | 94.5  | 92.0   | 3.4      | 1.5      | 30.1               | 30.2  |
| Stamt Perak II              | 88.2  | 81.1   | 0.2      | 0.8      | 22.2               | 27.5  |
| Stage of Tretes             | 136.3 | 128.5  | 0.5      | 0.6      | 26.3               | 29.5  |
| Tekung                      | 87.9  | 84.5   | 1.2      | 1.3      | 23.4               | 17.3  |
| Gandekan                    | 103.1 | 92.0   | 9.5      | 2.6      | 57.3               | 50.8  |
| Sukowono                    | 98.7  | 96.0   | -0.1     | 0.3      | 17.2               | 24.3  |
| Tapen                       | 98.2  | 92.0   | 0.6      | 1.0      | 21.4               | 20.0  |
| Saradan                     | 129.4 | 117.5  | -1.1     | 0.1      | 44.3               | 76.5  |
| bantur                      | 120.8 | 97.0   | 2.4      | 1.8      | 59.1               | 54.0  |
| Telebek                     | 102.4 | 93.5   | -0.6     | 0.7      | 23.0               | 31.5  |
| Botogardu                   | 103.2 | 97.0   | 6.0      | 2.0      | 28.6               | 33.0  |
3.2. Cluster Analysis for Maximum of Daily Rainfall

In this study we use 30 years maximum daily rainfall of 40 rainfall station period 1983-2012 to obtain clustering of rainfall station. Grouping of station was started with Principal Component Analysis (PCA). Principal component analysis aims to avoid multicollinearity between variables. Scree plot components based on variations can be seen in Figure 5 and a summary of each component are presented in Table 2.

![Figure 5. Scree plot of principal component based on variance.](image)

| PC.1   | PC.2   | PC.3   | PC.4   | PC.5   | PC.6   | PC.7   | PC.8   | PC.9   | PC.10  |
|--------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| Standard deviation | 107.72 | 67.12  | 63.55  | 61.32  | 56.28  | 55.31  | 49.05  | 45.23  | 42.95  | 38.10  |
| Proportion of Variance | 0.27   | 0.10   | 0.09   | 0.09   | 0.07   | 0.07   | 0.06   | 0.05   | 0.04   | 0.03   |
| Cumulative Proportion  | 0.27   | 0.37   | 0.47   | 0.56   | 0.63   | 0.70   | 0.76   | 0.80   | 0.85   | 0.88   |

According to the table 2, the first component can explain 27% variance of data, and then by adding a second component can explain until 37% variance of data. Following Morrison [4] that percentage of total diversity is considered sufficient if first 4 or 5 principal components account 75% or more of the total variance, thus in this study will using first 7 principal component that account 76% of total variance.
These seven components are formed from a combination of maximum daily rainfall for period 1983-2012. The equations for seven principal components are presented below:

\[
\text{PC1} = -0.13 \times t_{1984} - 0.13 \times t_{1985} - 0.32 \times t_{1986} - 0.19 \times t_{1987} - 0.13 \times t_{1988} - 0.15 \times t_{1989} - 0.18 \times t_{1990} \\
-0.20 \times t_{1991} + 0.3 \times t_{1992} - 0.31 \times t_{1993} + 0.16 \times t_{1995} - 0.14 \times t_{1996} - 0.14 \times t_{1997} - 0.11 \times t_{1998} \\
+ 0.24 \times t_{1999} - 0.15 \times t_{2000} - 0.24 \times t_{2001} - 0.15 \times t_{2003} - 0.18 \times t_{2005} - 0.33 \times t_{2007} - 0.18 \times t_{2008} \\
-0.17 \times t_{2009} - 0.18 \times t_{2011}
\]

\[
\text{PC2} = 0.13 \times t_{1983} + 0.14 \times t_{1985} - 0.29 \times t_{1986} - 0.23 \times t_{1987} + 0.14 \times t_{1990} + 0.13 \times t_{1991} - 0.37 \times t_{1993} + 0.16 \times t_{1997} - 0.50 \times t_{1999} + 0.14 \times t_{2000} - 0.34 \times t_{2007} - 0.16 \times t_{2008}
\]

\[
\text{PC3} = 0.13 \times t_{1983} + 0.27 \times t_{1985} + 0.33 \times t_{1986} + 0.4 \times t_{1987} - 0.15 \times t_{1991} - 0.31 \times t_{1992} + 0.29 \times t_{1993} \\
-0.2 \times t_{1995} - 0.34 \times t_{1996} + 0.23 \times t_{1999} - 0.13 \times t_{2001} + 0.16 \times t_{2002} - 0.12 \times t_{2003} - 0.18 \times t_{2005} \\
+ 0.27 \times t_{2006} - 0.133 \times t_{2007} + 0.1 \times t_{2011} + 0.11 \times t_{2012}
\]

\[
\text{PC4} = 0.22 \times t_{1987} + 0.54 \times t_{1988} - 0.22 \times t_{1992} + 0.13 \times t_{1993} + 0.38 \times t_{1995} - 0.19 \times t_{1996} + 0.32 \times t_{1999} - 0.24 \times t_{2000} - 0.25 \times t_{2002} - 0.22 \times t_{2004} - 0.12 \times t_{2007} + 0.14 \times t_{2008} - 0.12 \times t_{2011}
\]

\[
\text{PC5} = 0.12 \times t_{1984} + 0.2 \times t_{1986} - 0.33 \times t_{1988} + 0.38 \times t_{1990} + 0.46 \times t_{1995} + 0.21 \times t_{1997} - 0.12 \times t_{2000} \\
+ 0.14 \times t_{2001} + 0.17 \times t_{2006} - 0.51 \times t_{2007} + 0.2 \times t_{2008}
\]

\[
\text{PC6} = -0.15 \times t_{1985} - 0.16 \times t_{1986} - 0.12 \times t_{1987} - 0.31 \times t_{1988} - 0.29 \times t_{1990} + 0.12 \times t_{1991} + 0.46 \times t_{1995} + 0.10 \times t_{1996} - 0.14 \times t_{1997} - 0.32 \times t_{1998} - 0.16 \times t_{1999} - 0.28 \times t_{2000} + 0.17 \times t_{2001} \\
- 0.21 \times t_{2006} - 0.15 \times t_{2008} - 0.26 \times t_{2009}
\]

\[
\text{PC7} = -0.27 \times t_{1986} + 0.19 \times t_{1987} - 0.64 \times t_{1991} - 0.18 \times t_{1992} - 0.16 \times t_{1993} - 0.10 \times t_{2000} + 0.24 \times t_{2001} + 0.3 \times t_{2002} - 0.27 \times t_{2003} - 0.34 \times t_{2004} + 0.12 \times t_{2006} - 0.14 \times t_{2008}
\]

Cluster analysis by ward method with squared euclidean distance against these 7 first principal components of these stations. Dendrogram shows the process of grouping of each object separated into larger groups based on their distance. Dendrogram of this cluster analysis can be seen in Figure 6, list of cluster member and location of cluster member can be seen in Figure 7 and Table 3.

![Figure 6. Cluster Dendogram with 4 groups.](image-url)
Table 3. Member of each group.

| Group A          | Group B         | Group C         | Klaster D       |
|------------------|-----------------|-----------------|-----------------|
| Stageof Bandung  | Bojonegoro      | Cibukamanah     | Staklim Bogor   |
| Sumurwatu        | Banyuwangi      | Stageof Tanggerang | Nariewattie |
| Cicurug          | Juanda          | Budiarto Curug  | Jatiwangi       |
| Karangkendal     | Stamat Perak I  | Adisucipto      | Tnjg Priok      |
| Bantardewa       | Stamat Perak II | Kalianget       | Halim           |
| Pacet            | Tekung          | Gandekan        | Semarang Klim   |
| Leles            | Sukowono        | bantur          | Tegal           |
| Cimalaka         | Tapen           | Sangkapura      |                 |
| Bd.Bekasi        | Telebuk         | Stageof Tretes  |                 |
| Serang           | Botogardu       |                 | Saradan         |

From Figure 7 can be seen that cluster A, and B was spread all over Java Island except cluster C and D. Cluster C and D tent to spread in western area of Java Island. It’s mean that in western area of Java Island have more complex type of annual maximum of daily rainfall group.

Figure 7. Location of the cluster members, cluster A (red), cluster B (green), cluster C (blue), cluster D (purple).

Figure 8. Box plot for annual maximum data of each station based on his group.
Picture 8 and 9 show that cluster A has a lowest annual maximum value than the other clusters with small variations over period 1983-2012. Cluster B and C tend to have a higher value than Group A and have a larger variations, cluster C have tent to have a higher value than Group B. Cluster D just spread in western area in Java Island that have a highest variation value over period 1983-2012.

4. Conclusion
This study investigated descriptive statistics and grouping of rainfall station in Java Island for period 1983-2012. Most of rainfall stations have positive kurtosis value means the distribution of data have heavy tails, and outliers. All of data have positive skewness that indicate skewness distribution curve have tail at value above of the median, and mean is on the right of the peak value. Statistics descriptive analysis for 40 stations show that average of daily maximum rainfall in Java Region during period 1983-2012 were around 80 -181mm with median between 75-160mm and standard deviation between 17 to 82.

Cluster analysis produce 4 cluster that cluster A has a lowest annual maximum value than the other clusters with small variations over period 1983-2012. Cluster B and C tend to have a higher value than Group A and have a larger variations, cluster C have tent to have a higher value than Group B. Cluster D just spread in western area in Java Island that have a highest variation value over period 1983-2012. It is mean that western area of Java tent to have a higher annual maxima of daily rainfall than northern area, and have more variety of annual maximum value.

References
[1] Intergovernmental Panel on Climate Change (IPCC) 4th Assessment Report 2007 Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change. http://www.ipcc.ch/pdf/assessment-report/ar4/wg1/ar4-wg1-annexes.pdf
[2] Gilli M and Kellezi E 2006 An application of extreme value theory for measuring financial risk. Comput. Econ. 27 1-23
[3] Klein Tank A M G, Zwiers F W and Zhang X 2009 Guidelines on analysis of extremes in a changing climate in support of informed decisions for adaptation WCDMP No.72 WMO-TD No. 1500 http://eca.knmi.nl/documents/WCDMP_72_TD_1500_en_1.pdf
[4] Morisson D F 1990 Multivariate Statistical Method 3rd Edition (New York: McGrawHill Publishing Company)
[5] Sharma S 1996 Applied Multivariate Techniques (New York: John Wiley & Sons, Inc)
[6] Johnson R A dan Wichern D W 2007 *Applied Multivariate Statistical Analysis* (New Jersey: Pearson Prentice Hall, Inc)

[7] Singh W 2008 *Robustness of three hierarchical agglomerative clustering techniques for ecological data* (Iceland: University Of Iceland) [http://www.unuftp.is/static/files/rannsoknarritegdrdir/WarshaSingh_Masters_Thesis.pdf](http://www.unuftp.is/static/files/rannsoknarritegdrdir/WarshaSingh_Masters_Thesis.pdf)

[8] Balcan M F, Liang Y and Gupta P 2014 Robust hierarchical clustering *J. Machine Learning Research* **15** 1401-4051