THESIS – SS142501

STATISTICAL CLUSTERING OF HEAVY PRECIPITATION RADAR IMAGES IN SURABAYA USING GAUSSIAN MIXTURE MODEL

KIKI FERAWATI
NRP. 06211650010023

SUPERVISORS
Dr. rer. pol. Heri Kuswanto, M.Si.
Tomohiko Tomita, Ph.D.

MASTER PROGRAM
DEPARTMENT OF STATISTICS
FACULTY OF MATHEMATICS, COMPUTING, AND DATA SCIENCES
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2018

(this page is intentionally left blank)

STATISTICAL CLUSTERING OF HEAVY PRECIPITATION RADAR IMAGES IN SURABAYA USING GAUSSIAN MIXTURE MODEL

In partial fulfillment of the requirements for the degree of Magister Sains (M.Si.)
in
Institut Teknologi Sepuluh Nopember

By :
KIKI FERAWATI
NRP. 06211650010023

Date : July 23, 2018
Graduation Period : September 2018

Supervisors:
1. Dr. rer. pol. Heri Kuswanto, M.Si.
2. Tomohiko Tomita, Ph.D

Approved by :

Dr. rer. pol. Heri Kuswanto, M.Si. (Supervisor I)
NIP. 19820326 200312 1 004

Dean
Faculty of Mathematics, Computing, and Data Sciences
Institut Teknologi Sepuluh Nopember

Prof. Dr. Basuki Widodo, M.Sc
NIP. 19650605 198903 1 002

(this page is intentionally left blank)

STATISTICAL CLUSTERING OF HEAVY PRECIPITATION RADAR IMAGES IN SURABAYA USING GAUSSIAN MIXTURE MODEL

Name of Student	: Kiki Ferawati
NRP	: 06211650010023
Supervisors	: Dr. rer. pol. Heri Kuswanto, M.Si.
	Tomohiko Tomita, Ph.D.

ABSTRACT

Precipitation in Indonesia is affected by a wide range of weather variability. Understanding the characteristics of precipitation in the area is essential in order to predict heavy precipitation event. Characteristics of precipitation, e.g. its shape and pattern are important feature to predict extreme rainfall events obtained from radar images. This study applied the Gaussian Mixture Model (GMM) for high dimensional data clustering (hereafter denoted as HDDC) to cluster the shapes appearing in the radar images associated with heavy precipitation events in Surabaya. Another method used for this analysis is *K*-means clustering with principal component analysis (PCA). Using ITS precipitation data, the Hill Plot and Mean Residual Life Plot (MRLP) suggested that the extreme event is characterized with the precipitation above 1.5 mm per ten minutes. According to the Bayesian Information Criterion (BIC), the HDDC suggested 10 clusters to characterize the heavy precipitation patterns. Another clustering method, *K*-means with PCA is also applied to the data. However, out of the 10 clusters, several clusters show similar pattern, suggesting that 10 clusters are too many for the data. Reviewing the value of Pseudo-F and Silhouette of *K*-means and the BIC value of HDDC, 2 clusters are deemed best for radar images data. The analysis for both *K*-means and HDDC shows some inconsistency in terms of the cluster members, due to the small sample size. Hence, ensemble-based HDDC is proposed to overcome the problem. This method generated better results with robust cluster. It resulted in two clusters representing the pattern of precipitation system in Surabaya.

Keywords: Radar image, Heavy precipitation, Cluster, Gaussian Mixture Model, *K*-means

(this page is intentionally left blank)

PREFACE

This thesis is written as the graduation requirement for master program in Statistics, Institut Teknologi Sepuluh Nopember. Thanks to Allah SWT for the blessing, so that the writer can complete this thesis titled "Statistical Clustering of Heavy Precipitation Radar Images in Surabaya using Gaussian Mixture Model".

This thesis cannot be completed with support from people around me, be it academically or spiritually. Therefore, I would like to express my thanks to:

1. Dr. Heri Kuswanto, M.Si as my supervisor, for his support and guidance for me in writing this thesis.
2. Tomohiko Tomita, Ph.D as my co-supervisor, for his support and patience in teaching me about meteorological field during my exchange period and for the guidance in writing my thesis.
3. Dr. Dedy Dwi Prastyo, M.Si and Dr. Kartika Fithriasari, M.Si for the evaluation and suggestion for improving this thesis.
4. Dr. Suhartono, M.Sc as Head of Statistics Department.
5. All the lecturers in master program of Statistics in ITS.
6. My family, Father, Mother, Brother and Sister, for their support and encouragement, starting from my decision of taking master program, went to an exchange program until completing this thesis.
7. Epa Suryanto as a good discussion partner.
8. Mbak Shofi for helping me review this thesis.
9. Saidah as friend in the same department and during exchange program, we have gone through a lot together in our journey on writing our thesis.
10. Tete, Tri, Sella, Zakya and Arlene, as a good friend that I met in this master program.
11. Chiko, Nila, Yunita, Farida and other friends that I met during the exchange program.

Surabaya, July 2018

Kiki Ferawati

(this page is intentionally left blank)

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

(this page is intentionally left blank)

LIST OF TABLES

Page

(this page is intentionally left blank)

LIST OF ENCLOSURES

Page

(this page is intentionally left blank)

CHAPTER 1
INTRODUCTION

1.1 Background

Due to its location and landscape, precipitation in Indonesia is affected by a wide range of weather variability. Located at 11^{o}S – 6^{o}N and 95^{o} – 141^{o}E, Indonesia is a tropical country with two seasons, i.e., wet and dry season. Several climate variabilities such as El Nino Southern Oscillation (ENSO), Madden-Jullian Oscillation (MJO) and Indian Ocean Dipole (IOD) affect the precipitation in Indonesia (Hendon, 2003; D'Arrigo & Wilson, 2008; Hidayat & Kizu, 2010). Islam, Hayashi, Terao, Uyeda & Kikuchi (2005) argued that understanding the characteristics of precipitation such as shape, size, and direction of precipitation systems are very important. Weather radar systems and satellites have provided information on spatial patterns of precipitation which can be used to study the characteristics of the system (AghaKouchak, Nasrollahi, Li, Imam, & Sorooshian, 2010), which might provide useful information to predict heavy precipitation event. Research using radar data for precipitation prediction has been carried out since a long time ago. Harrison (2000) worked on improving precipitation estimates from weather radar using quality control and correction techniques. Wang et al. (2009) used rainfall radar imaging in a nowcasting system.

There are only few research involving radar images in Indonesia. Ilhamsyah (2013) utilized weather radar images to support marine and fisheries activities near Aceh area. By interpreting the weather radar images, Ilhamsyah was able to get information on potential hazardous areas, which was valuable information for fisherman to prepare for their activities in the sea. Paski (2017) studied about assimilating model from global forecast system output to radar and satellite image observation data. The result of the study was the rainfall predictions with the assimilation of satellite data shown to be the best results. With the development of the radar system in Indonesia, radar image now can be used as a tool to help on understanding weather system in Indonesia. Unfortunately, the weather radar data of the Meteorological Office Indonesia (BMKG) are available only on several spots

over Indonesia. As of 2016, there are 40 weather radars installed and 20 other radars were planned to be installed in 2017, 2018 and 2019 (BMKG, 2017). Most of radars are located in big cities in Indonesia, such as Tangerang, which is located near Jakarta, Semarang, and Surabaya.

Surabaya as the second biggest city in Indonesia is the center of economy of East Java. Its strategic location also made Surabaya the center for economic activity in Eastern Indonesia. As the city is currently growing and being home of offices and business centers, it rapidly transforms into trading center. Several major industries in Indonesia are based in Surabaya, and several area in Surabaya have become business center, with plenty of shopping mall, apartment and office building (Ostojic, Bose, Krambeck, Lim, & Zhang, 2013).

However, Surabaya has the risk of flooding. As the center of economic activity, flooding can disturb the flow of economy in the area. There are several causes of flood in Surabaya. One of them is sea level rise. Because Surabaya is located in the coastal area, such problem cannot be avoided. Imaduddina and Subagyo (2014) made flood risk zone map which identifies 5 risk levels according to the National Disaster Mitigation Guidance for coastal area in Surabaya. Another cause of flood in Surabaya is heavy rain. In the recent years, heavy rain has been the cause of flood in several areas in Surabaya. Five hours of rain caused flood in Surabaya in April 2016 (TEMPO, 2016). In May 2016, heavy rain of three hours straight caused flood in a number of regions in Surabaya (TEMPO, 2016; REPUBLIKA, 2016). Flood caused by heavy rain also happened in Surabaya on February 2017 (KOMPAS, 2017a). In November 2017, a flood with the level of 50 cm happened in Surabaya. There has been also a report of high precipitation intensity on the day it happened (KOMPAS, 2017b). Flood has become one of recurring problems in Surabaya every year. Due to the serious impacts of flood induced by heavy precipitation happened in Surabaya, thus predicting the heavy precipitation pattern in Surabaya is extremely important. Surabaya is located in region A or monsoonal regime. Region A has single peak of monthly rainfall around December-January, which is the peak of the wet season in Indonesia (Aldrian & Susanto, 2003).

As mentioned above, characteristics of precipitation system such as its shape are important factor to predict extreme rainfall events. In order to get a clear image of the shape of precipitation system at the time of heavy precipitation, the availability of radar images in Surabaya provided by BMKG can be a useful resource. However, radar image is updated every ten minutes, resulting in more than a hundred images available for a single day. Combining all the data for wet season of 2017/2018, from October 2017 to March 2018, there are more than ten thousand images. Furthermore, in a case where rain did not happen, the image will not contain any useful information. Analyzing all the images will be difficult and inefficient. Therefore, a step to filter the necessary data is important. Because the purpose of the analysis is to predict heavy precipitation in Surabaya through the characteristics of precipitation system, identifying the event of heavy precipitation in Surabaya is important for selecting the images for the analysis.

One of the methods used to analyze extreme event is Extreme Value Theory (EVT) introduced by Fisher and Tippett (1928). The aim of EVT is to predict the occurrence of rare events. The EVT has been widely applied in various field of research. Marimotou, Raggad & Trabelsi (2006) used EVT to manage energy price risks. Gilli and Kellezi (2006) applied EVT for measuring financial risk in major stock market indices. Among applications of EVT in climate are by Goldstein, Mirza, Etkin, & Milton (2003), by using EVT for constructing extreme climate scenarios. Cooley (2005) used EVT for developing models in several cases based on the issues in climate and weather studies. Rahayu (2013) used block maxima with Generalized Extreme Value (GEV) distribution approach to identify climate change in Indramayu. In the case of application of EVT to precipitation event, Montfort and Witter (1986) used GPD to fit rainfall series in Dutch. Langousis, Mamalakis, Puliga, & Deidda (2016) studied about estimating the threshold for Generalized Pareto Distribution for NOAA NCDC daily rainfall data. By applying EVT to the precipitation data, the heavy precipitation threshold as the criteria for selecting radar images can be determined.

The next step of analysis after identifying the event of heavy precipitation is clustering the shape of radar images when heavy precipitation occurs. Clustering the shape will be useful to understand the characteristics of precipitation in the same

clusters. Clustering methods are techniques of grouping based on similarities or distances between the objects (Johnson & Wichern, 2007). Generally, clustering method is divided into hierarchical and nonhierarchical cluster. Hierarchical clustering methods are either a series of successive mergers or a series of successive divisions. Single linkage, complete linkage, and average linkage are examples of hierarchical clustering method.

On the other hand, nonhierarchical clustering methods are clustering techniques that are designed to group items into collection of K clusters. *K*-means is a popular nonhierarchical clustering method. Aside from the popular hierarchical and nonhierarchical clustering methods, there is also clustering method based on statistical model, namely mixture model. The most common mixture model is Gaussian Mixture Model (GMM), with each mixture following normal distribution.

The GMM has been widely used for clustering problem. Ling and Zhu (2017) used GMM to predict precipitation events in Shanghai. In the case of image data, GMM is usually applied for the case of image segmentation. Kalti and Mahjoub (2014) used GMM to classify pixels based on weighted similarity distance. However, classical GMM has certain problems when faced with high dimensional data. Bouveyron (2007b) found that the classical GMM show a disappointing behavior when the size of the dataset is too small compared to the numbers of parameter to be estimated.

Image processing is a problem of high dimensional spaces. High dimensional data clustering (HDDC) is a method for clustering based on Gaussian Mixture Model designed for high dimensional data (Bergé, Bouveyron, & Girard, 2012). However, due to small number of images in heavy precipitation event, HDDC may not perform well, hence a new method of modified HDDC by using bootstrap resampling is proposed. This study adopted the ensemble concept applied to HDDC to obtain optimal cluster member for identifying the shape of precipitation system in Surabaya, East Java.

1.2 Research Questions

Based on the background described on the first section, the problem in this study is to find the cluster of heavy precipitation over Surabaya observed from radar images. The performance of several clustering approaches are evaluated.

1.3 Objective of The Study

The goal of this research are described below.

1. To examine the shape or pattern of radar image associated with heavy precipitation in Surabaya through cluster analysis.
2. To evaluate the performance of several clustering methods for identifying the radar image pattern of heavy precipitation in Surabaya.

1.4 Significance of The Study

This research is expected to be a significant material for:

1. For the student, to learn about application of GMM in radar image for heavy precipitation in Surabaya.
2. For BMKG, to help on optimizing the usage of radar image for heavy precipitation prediction.
3. For the next research, this research is expected to help on improving knowledge about application of statistics to climate data using extreme value analysis and clustering method.

1.5 Scope and Limitation

The limit of this study are described below.

1. Data used in this research spans from October 18, 2017 to March 31st, 2018, corresponding to rainy season in Indonesia.
2. The area of analysis is limited to 150×150 pixels in Surabaya area and only focused on R (red) component of the image.

(this page is intentionally left blank)

CHAPTER 2
LITERATURE REVIEW

2.1 Extreme Value Analysis

Extreme value analysis is the branch of statistics which attemps to characterize the tail of a distribution (Cooley D. , 2009). Extreme value theory is one of the most important statistical method. Extreme value analysis usually require estimation of the probability of events that are more extreme than any observed value (Coles, 2001). Throughout the year, extreme value analysis has been widely used in various disciplines such as insurance industry, risk assessment, assessment of meteorological change and more.

Peaks Over Threshold (POT) is a method for identifying extreme value by using a threshold. For the practical applications, the POT models are generally preferred because it is more efficient to use since all observations above the threshold are used to estimate parameters of the tail (Marimotou, Raggad, & Trabelsi, 2006). The data exceeding the threshold can be estimated well using Generalized Pareto Distribution (GPD) (Leadbetter, 1991; Beirlant, Goegebeur, Teugels, Waal, & Ferro, 2014). If Y is a random variable distributed as GPD with scale parameter σ and shape parameter ξ, the function is written in Equation 2.1.

$$G_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 + \frac{\xi y}{\beta}\right)^{-\frac{1}{\xi}} & \text{if } \xi \neq 0 \\ 1 - \exp\left(-\frac{y}{\beta}\right) & \text{if } \xi = 0 \end{cases} \tag{2.1}$$

where $\beta > 0$; $y \geq 0$ when $\xi \geq 0$ and $0 \leq y \leq -\beta/\xi$ when $\xi < 0$.

y = observation exceeding threshold u

β = scale parameter

ξ = shape parameter

For a Pareto distribution, the tail index α is the reciprocal value of ξ when $\xi > 0$. The special case happened when $\xi = 0$, as the GPD is the same as the exponential distribution with mean β (Ghosh & Resnick, 2010). The shape parameter ξ is important for determining the qualitative behavior of GPD. If $\xi < 0$, the excess distribution has an upper bound defined by $u - \beta/\xi$ and if $\xi > 0$ then

the GPD has no upper limit. Parameter of GPD can be estimated using maximum likelihood. Solving likelihood function of GPD is quite complicated so log-likelihood is used instead.

Determining threshold of extreme value is a bit tricky. There are several methods to decide the threshold value, two of them are visual assessment of threshold choice plot which require prior experience for interpretation, namely Hill plot and Mean Residual Life Plot.

2.1.1 Hill Plot

Hill plot is one of methods for determining threshold for extreme value theory. Identifying the threshold or tail index α of a dataset is really important for extreme value cases. Hill plot is very efficient when the dataset is from Pareto distribution (Drees, Haan, & Resnick, 2000).

The Hill estimator is the conditional maximum likelihood estimator for heavy-tailed distributions. If the data exceeding threshold u and follows Pareto distribution with index α, the distribution exceeding u is given in Equation 2.2.

$$F^{[u]}(x) = 1 - \left(\frac{u}{x}\right)^{\alpha}, x \geq u \qquad (2.2)$$

The data is denoted by $\{x_i\}_{i=1}^{N}$, with sample size of N whose k largest value exceeds the threshold u. The parameter of Hill estimator can be estimated by equation below.

$$\hat{\xi}_{k,N} = \left(\hat{\alpha}_{k,N}\right)^{-1} = \frac{1}{k}\sum_{i=1}^{k}[ln\, x_{N-i+1} - ln\, x_{N-k}] \qquad (2.3)$$

with $x_{(i)}$ as the order statistics of the series x, with $x_{(N)} > x_{(N-1)} > \cdots > x_{(1)}$. Under some additional restrictions on the behavior of underlying distribution function, $\hat{\xi}_{k,N}$ is asymptotically Gaussian with mean ξ and variance $(\xi^2 k)^{-1}$. The $(1-x)$% confidence intervals can be computed as Equation 2.4.

$$\hat{\xi} \pm \lambda_{x/2}\frac{\hat{\xi}}{\sqrt{k}} \qquad (2.4)$$

where $\lambda_{x/2}$ is the $\left(1 - \frac{x}{2}\right)$ standard Normal quantile. Each different threshold value might lead to a different Hill estimator (Alfarano & Lux, 2010). The value of Hill estimator is then plotted in the figure, called Hill plot.

Based on simulation by Alfarano and Lux (2010), the appropriate tail fraction for the "best" estimator for the 'true' parameter α is not immediately obvious. The possible approach is searching for a region in the Hill plot where the estimated values are approximately constant, called eyeball method. This approach relied heavily on subjective graphical data analysis.

2.1.2 Mean Residual Life Plot

Mean Residual Life Plot (MRLP) is one of visual procedures to determine threshold in extreme value analysis. MRLP is subjective and sometimes difficult to interpret (Thompson, Cai, Reeve, & Stander, 2009). Mean excess function is a tool to help determining the threshold choice of u. The mean excess function of a random variable X with threshold u and endpoint x_F is defined as

$$e(u) = E(X - u|X > u), 0 \leq u < x_F \qquad (2.5)$$

The value $e(u)$ is the mean excess over the threshold value u. An appropriate value of the high threshold can be found by plotting the empirical mean excess function (Embrechts, Klüppelberg, & Mikosch, 1997). A mean excess plot, or can also be called MRLP in reliability cases, consist of the graph $\{(X_{k,n}, e_n(X_{k,n})) : k = 1, \dots, n\}$. For X following GPD with parameters $\xi < 1$ and β, for $u < x_F$, the mean excess function is defined as

$$e(u) = E(X - u|X > u) = \frac{\beta + \xi u}{1 - \xi}, \beta + u\xi > 0 \qquad (2.6)$$

From the function above, it can be inferred that the mean excess function of a GPD is linear. The empirical mean excess function of a given sample $X_1, \dots, X_n$ is defined by

$$e_n(u) = \frac{1}{N_u} \sum_{i \in \Delta_n(u)} (X_i - u), u > 0 \qquad (2.7)$$

where $N_u = \text{card}\{i : i = 1, \dots, n, X_i > u\} = \text{card}\Delta_n(u)$. This suggests a graphical approach for choosing u, choose u > 0 such that $e_n(u)$ is approximately linear for $x \geq u$. However, this is difficult because the term approximately is subjective to the observer.

2.2 Principal Component Analysis

Principal Component Analysis (PCA) is a method with the purpose of explaining variance-covariance structure of a set of variables through a few linear combinations of these variables. The linear combinations is called as principal components (PC). The objectives of PCA is for data reduction and interpretation (Johnson & Wichern, 2007). The PCA is the most popular and one of the oldest multivariate statistical technique and able to incorporate a large number of another multivariate methods, such as canonical analysis and linear disciminant analysis (Abdi & Williams, 2010) and also important in other statistical methods, such as linear regression (Joliffe & Cadima, 2016).

Suppose there are p random variable, denoted by $X_1, X_2, \dots, X_p$. Random vector $\boldsymbol{X}' = [X_1 \quad X_2 \quad \cdots \quad X_p]$ has variance and covariance matrix with eigen value $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Each eigenvalue has eigenvector denoted by $\boldsymbol{a}_i, i = 1, \dots, p$. Let $Y_i = \boldsymbol{a}_i'\boldsymbol{X} = a_1X_1 + a_2X_2 + \cdots + a_pX_p$. Principal component is linear combinations which maximizes $Var(a_i'X)$, subject to $a_i'a_i = 1$ and $Cov(a_i'X, a_k'X)$, k≠i.

Other than using the original data, principal components can also be computed using the standardized variables. The matrix notation for standardized variables is written below.

$$\boldsymbol{Z} = \left(\boldsymbol{V}^{1/2}\right)^{-1}(\boldsymbol{X} - \boldsymbol{\mu}) \tag{2.8}$$

where

$$\mathbf{V}^{1/2} = \begin{bmatrix} \sqrt{\sigma_{11}} & 0 & \cdots & 0 \\ 0 & \sqrt{\sigma_{22}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\sigma_{pp}} \end{bmatrix}$$

$\mathrm{E}(\mathbf{Z}) = \mathbf{0}$

$\mathrm{Cov}(\mathbf{Z}) = \left(\mathbf{V}^{1/2}\right)^{-1}\boldsymbol{\Sigma}\left(\mathbf{V}^{1/2}\right)^{-1} = \boldsymbol{\rho}$

The i-th principal components of $\mathbf{Z}$ with $\mathrm{Cov}(\mathbf{Z}) = \boldsymbol{\rho}$ is given by

$$Y_i = e_i'\left(\mathbf{V}^{1/2}\right)^{-1}(\mathbf{X} - \boldsymbol{\mu}), \qquad i = 1,2, \dots p.$$

with $\sum_{i=1}^{p} \text{Var}\,(Y_i) = \sum_{i=1}^{p} \text{Var}\,(Z_i) = p$ and $\rho_{Y_i,Z_k} = e_{ik}\sqrt{\lambda_i}$. The pair $(\lambda_1, e_1), (\lambda_2, e_2), \ldots, (\lambda_p, e_p)$ are eigenvalue and eigenvector of $\boldsymbol{\rho}$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Determining number of selected PC is essential for PCA. Jolliffe (2002) explained about several methods for deciding the number of principal components to be used. Let m be the number of chosen PCs. The first rule for finding appropriate m is cumulative percentage of total variation. The formula for this rule is written in Equation 2.9.

$$\text{Percentage of variance up to PC-m} = \frac{\sum_{i=1}^{m} \lambda_i}{p} \qquad (2.9)$$

This is the most obvious criteria for choosing the optimal m. The number of PCs, m, is the smallest value of m exceeding the chosen percentage. The total percentage usually can be set at 70%, 80% or 90%, depended on the desired value of contribution by PCs.

Another popular rule for determining number of PCs is using the scree graph or scree plot. Scree graph is a figure plotting eigenvalue or variance against component. It is even more subjective than the first rule, because it require visual observation to determine the optimal number of m. The optimal number, m, is selected by deciding where is the steep point of the graph. Figure 2.1 shows the example of scree graph (Jolliffe, 2002, fig 6.1).

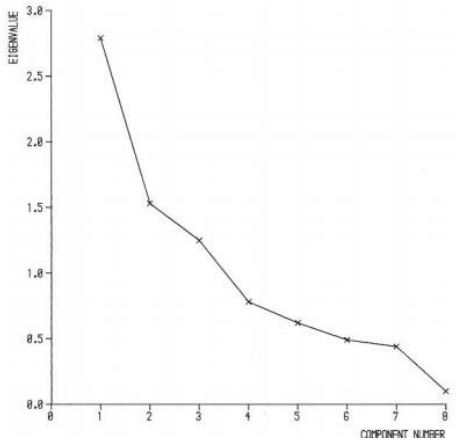


Figure 2.1 Scree graph

The number of selected PCs is not fixed, as different rule will produce different result of m. However, selecting optimal number of PCs should also consider the objectives of PCA, because it will set different requirements for how many PCs are needed. Simple rules of selection, as mentioned above, usually work well in application.

2.3 *K*-means Clustering

K-means is a clustering technique belong to nonhierarchical clustering methods. There are two common starting points for nonhierarchical clustering methods, the first is initial partition of items divided into groups or initial set of seed points to be core of the clusters (Johnson & Wichern, 2007). General procedure for *K*-means method is described below.

1. Partitioning the items into K initial clusters. Specify K initial centroids for each cluster.
2. Assigning an item to the cluster whose centroid is the nearest. Suppose there are p variable. For every single observation, let $\mathbf{c}_{\mathrm{i}} = [c_{i1} \quad c_{i2} \quad \cdots \quad c_{ip}]$ as the centroid for each variable in cluster i, and $\mathbf{x} = [x_1 \quad x_2 \quad \cdots \quad x_p]$ as the observation value for each variable, $i = 1, \ldots, K$. The distance is calculated using Euclidean distance:

$$d(c_i, x) = \sqrt{(c_{i1} - x_1)^2 + (c_{i2} - x_2)^2 + \cdots + \left(c_{ip} - x_p\right)^2} \quad (2.10)$$

3. Repeating step 2 until there are no more changes in cluster member.

   The result of final clusters is dependent to the value of initial centroid.

2.4 Gaussian Mixture Model

Mixture model is a method that can be used in problem where the population of sampling unit consists of a number of subpopulations within each of which a relatively simple model applies (Gelman, et al., 2013).

$$f(y|\boldsymbol{\theta}, \boldsymbol{\pi}) = \sum_{j=1}^{K} \pi_j p_j(y|\boldsymbol{\theta}_j) \quad (2.11)$$

where

$f(y|\boldsymbol{\theta}, \boldsymbol{\pi})$ = density function of mixture distribution

$p_j(y|\boldsymbol{\theta}_j)$ = density function j of *K* component of mixture distribution

$\boldsymbol{\theta}_j$ = vector parameter of each component of mixture distribution $(\theta_1, \theta_2, \dots, \theta_K), j = 1,2, \dots, K$

$\boldsymbol{\pi}$ = vector parameter of proportion $(\pi_1, \pi_2, \dots, \pi_K)$

π_j = proportion parameter of mixture distribution component, $\sum_{j=1}^{K} \pi_j = 1$ and $0 \leq \pi_j \leq 1, j = 1,2, \dots, K$

K = number of distribution in the mixture distribution

Based on Equation 2.15, $p_j(y|\boldsymbol{\theta}_j)$ depends on the distribution used for the model. In case of Gaussian Mixture Model (GMM), $p_j(y|\boldsymbol{\theta}_j)$ follows normal distribution.

There are several approaches to estimate the parameters of mixture model, namely Expectation and Maximization (EM), Neural Network, Maximum Likelihood and Bayesian. The approach used in this study is EM algorithm. There are two steps in EM algorithm, which done repeatedly for cluster forming, namely Expectation (E-step) and Maximization (M-step).

- E-step will generate expectation of parameter of the data based on data distribution.
- M-step will calculate parameter estimation using expected value of the previous E-step. The formulation for M-step will be obtained through Maximum Likelihood Estimation (MLE).

Those two steps will be repeated continuously until converge or reach certain tolerance value.

Main problems in GMM is determining probability of a single observation of x_i belong to certain group. GMM belong to soft clustering, using probability to assign observation to certain cluster. However, standard GMM has been proven to have disappointing result when the size of dataset is too small compared to the number of parameter. A method called High Dimensional Data Clustering (HDDC) has been developed by Bouveyron (2007) to address this issue. This method uses reparameterization to limits the number of parameters to estimate while proposing a flexible modeling of the data (Bouveyron, Girard, & Schmid, High-Dimensional Data Clustering, 2007a).

In the classical GMM model, the class are assumed to follow normal distribution $N_p(\mu_k, \Sigma_k)$. Let Q_k be the orthogonal matrix with the eigenvectors of Σ_k as columns and Δ_k be the diagonal matrix which contains the eigenvalues of Σ_k such that $\Delta_k = Q_k^t \Sigma_k Q_k$. The matrix Δ_k is covariance matrix of the *k*-th class in its eigenspace. In the sequence of HDDC, the model will be denoted as $[a_{kj} b_k Q_k d_k]$, with each parameter denotes:

$a_{k1}, a_{k2}, \dots, a_{kd_k}$ = model of variance of the actual data of *k*-th class

b_k = model of variance of noise. $a_{kj} > b_k, j = 1, \dots, d_k$

d_k = intrinsic dimension of latent subspace of the *k*-th group which spanned by the d_k first column vectors of Q_k. d_k is equal to $(p-1)$ for all $k = 1, \dots, K$

There are several types of model as well, the details of full available model can be found in works by Bouveyron (2007). The advantages of this model over the classical GMM is in the number of parameter to be estimated. The model in HDDC had much less parameters than the classical GMM. Table 2.1 gave information about the number of parameters used in classical GMM and model $[a_{kj} b_k Q_k d_k]$ in HDDC.

Table 2.1. Number of parameters used in classical GMM and HDDC

Clustering method	Number of parameters	Asymptotic order
Full GMM (Classical GMM)	$\rho + Kp(p+1)/2$	Kpd
Model $[a_{kj} b_k Q_k d_k]$	$\rho + \bar{\tau} + 2K + D$	$Kp^2/2$

Information of number of parameters in Table 2.1 contain details of notation where:

K = number components in the mixture model

p = number of variables

ρ = number of parameters for estimating means and proportions, $\rho = Kp + K - 1$

$\bar{\tau}$ = number of parameters required for estimating orientation matrices Q_k $\bar{\tau} = \sum_{k=1}^{K} d_k[p - (d_k + 1)/2]$

D = sum of intrinsic dimensions, $D = \sum_{k=1}^{K} d_k$

The process of maximizing the likelihood in HDDC is also using EM algorithm. The EM algorithm will stops when the difference between estimated values of the likelihood at two consecutive iterations is smaller than a certain threshold (Bergé, Bouveyron, & Girard, 2012).

2.5 Clusters Evaluation

Evaluating the result of clustering is an important aspect of the clustering methods. This also include the problem of determining the best number of clusters for the data. There are several methods available for evaluating cluster results. The detail for each method will be described below.

1. Silhouette

   Silhouette value is based on comparison of cluster tightness and separation, showing which objects fit well within the cluster and which objects are somewhere between the clusters. The average Silhouette width contain information about evaluation of clustering validity. The value can be used to help determine an appropriate number of clusters (Rousseeuw, 1987). The Silhouette width $s(i)$ for each $i \in I$ is defined as written in Equation 2.12.

   $$s(i) = \frac{b(i)-a(i)}{max(a(i),b(i))} \quad (2.12)$$

   where

   $a(i)$ = the average distance between i and other entities of the cluster to which i belongs

   b(i) = minimum of the average distances between i and all the entities in each other cluster.

   The range of Silhouette width value lies between -1 and 1. If the value is closer to 1, it means the observations are well clustered (Kodinariya & Makwana, 2013).

2. Pseudo-F statistic

   Pseudo-F statistic, also known as Calinski-Harabasz index is an informal indicator for suggesting best number of clusters using variance ratio criterion (Calinski & Harabasz, 1974). It is one of clustering evaluation method

calculated using sum squares of within and between clusters. The formula is given below.

$$C = \frac{\frac{SSbetween}{K-1}}{\frac{SSwithin}{N-K}} \quad (2.13)$$

with K as the number of clusters and N the number of observations (Desgraupes, 2013). Normally, the best number of clusters are chosen based on the highest value of *C*. Relatively large values indicate a good split of data (Soldek, Saeed, & Pejas, 2012).

3. Bayesian Information Criterion (BIC)

   Bayesian Information Criterion (BIC) is a criterion for model selection. BIC is closely related to Akaike Information Criterion (AIC) and the difference between them are BIC is giving bigger penalty term for the number of parameter in the model compared to AIC (Schwarz, 1978). The preferred model is the model with lowest BIC value. BIC for GMM is calculated using formula below.

$$BIC = 2\,ln\,L_{max} - 2\,ln(N)\left(K\frac{1}{2}(p+1)(p+2) - 1\right) \quad (2.14)$$

2.6 Bootstrap

Bootstrap is a data-based simulation method to draw a conclusion based on the data (Efron & Tibshirani, 1993). Bootstrap is a repeated sampling procedure from a set of the data. The advantage of using nonparametrical approach is no assumption of data distribution need to be fulfilled. Each bootstrap sampling would result in a different value. Theoretically, because the sample of bootstrap is taken repeatedly, the asymptotic distribution would follow the actual distribution of the data.

Take $\hat{F}$ as the original empirical distribution of a set of data, x, then each point of the data has the same probability of being taken as sample. A sample of bootstrap is defined as n random sample taken from $\hat{F}$. The sample of $x^* = (x_1^*, x_2^*, \dots, x_n^*)$. The asterisk mark indicates that those are not actual data, but the resampling data from x. Measurement taken from bootstrap can be shown in Equation 2.15.

$$\hat{\theta}^* = s(x^*) \quad (2.15)$$

where $s(x^*)$ are the result of applying function $s(.)$ to x^*. Function $s(.)$ can be replaced with any function needed for the analysis, such as mean, median or other measurement.

2.7 Precipitation

Precipitation is amount of liquid water depth of the water substance that has fallen at given point over a specified period of time, usually expressed in millimeters or inches (American Meteorological Society, 2018). Precipitation is usually measured using rain gauge in meteorological or observational station.

2.8 Radar image

Radar image describe potential intensity of rainfall detected by weather radar. The precipitation intensity is measured based on the amount of radar energy reflected by droplets in the clouds, described by reflectivity product in dBZ. (BMKG, 2018). The range of dBZ scale is 5-75, denoted by color gradation of sky blue to light purple. The range of precipitation intensity can be described using Table 2.2.

Table 2.2 Rainfall intensity based on dBZ score

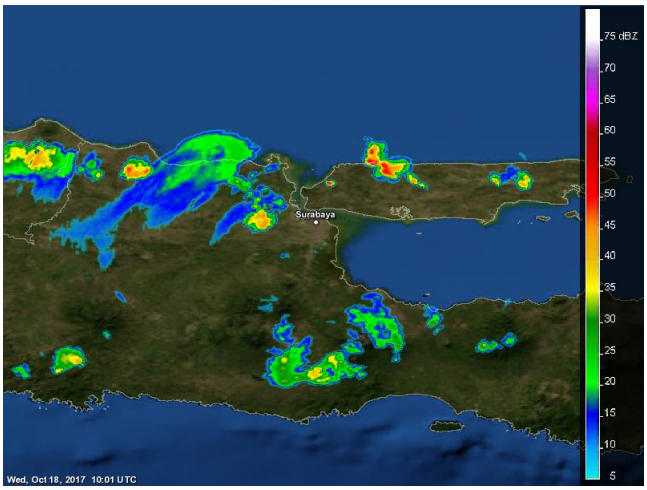
Precipitation Intensity	dBZ score	mm/hr
Light rain	30 to 38	1 to 5
Medium rain	38 to 48	5 to 10
Heavy rain	48 to 58	10 to 20
Very heavy rain	>58	>20

(this page is intentionally left blank)

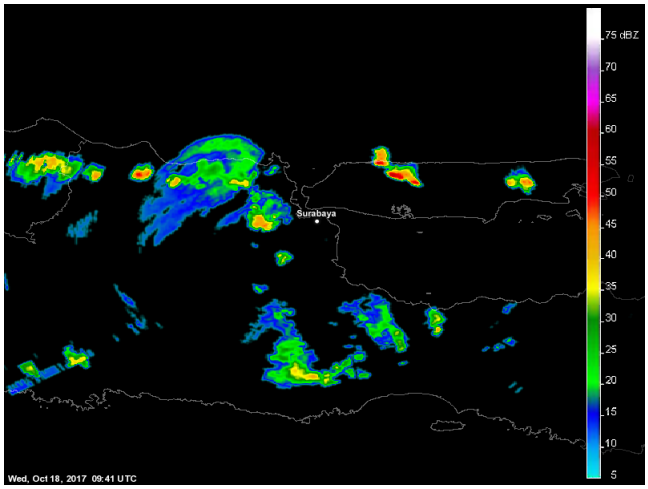
CHAPTER 3
RESEARCH METHODOLOGY

3.1 Data Source

There are two sets of data used in this research. The first is radar image precipitation over East Java, taken from BMKG site (http://radar.bmkg.go.id/bmkg2/imageQC/). The image is usually updated every 10 minutes and stays in the web for a day before replaced by new image. There are two types of image i.e, the normal colored image and the image with black background. Illustration of radar images are given in Figure 3.1.



(a)



(b)

Figure 3.1. (a) Radar image with colored background (b) Radar image with black background

Figure 3.1 shows the radar image in East Java on October 18, 2017 at 10.01 a.m UTC time. The image used in this study is image with black background, as it is easier to process without the problem of background noise. In the black image, the color of radar reflectivity can be clearly seen.

The second dataset is precipitation data recorded from rain gauge installed at the Institut Teknologi Sepuluh Nopember (ITS) Surabaya. The data can be accessed online at http://www.pskbpi.its.ac.id/its-weather-station/. There are several variables available, such as temperature, dew point, humidity, wind direction, wind speed, wind gust, pressure and accumulation of precipitation. The data is updated every 5 minutes. The precipitation data shown in the website is the accumulation of precipitation which will be reset by midnight. To obtain the precipitation for each 5 minutes period, the difference between each period is calculated.

3.2 Research Variable

The variable used in this research is described as follows.

1. Precipitation data from ITS, obtained from ITS Weather Station. The station is located in S 7°16‘48”, E 112°47’41, 28m above the sea level. The data is recorded per five minutes in WIB time (UTC+7), starting from 18-10-2017 to 31-03-2018.
2. Radar images of East Java, obtained from BMKG site. The image is updated every ten minutes in UTC time, starting from 18-10-2017 to 31-03-2018.

3.3 Step of Analysis

The steps of analysis in this research are described below.

1. Describing the precipitation data.
2. Aggregating the precipitation data for every 10 minutes to match the frequency of radar images.
3. Determining threshold for extreme precipitation by using MRLP and Hill Plot.
4. Selecting aggregated data that exceeds the threshold.
5. Using the date and time from the selected data to choose corresponding images at the time.

6. Preprocessing step for radar images.
   (i). Cutting the area of images. The original image is 600×800 pixels. For this analysis, the chosen area is smaller, size of 150×150 pixels with Surabaya as the center.
   (ii). Separating the components of the image. There are three component of R (Red), G (Green) and B (Blue) for each image. The red component is chosen because it signify the region of heavy precipitation.
   (iii). The process is done repeatedly on all selected images.
7. Obtaining the value of R component from the images, stored in matrix **X**. The size of **X** is 150×150. There are 161 matrices **X** available, correspond to each selected image. The structure of matrix **X** is shown on Table 3.1.

Table 3.1 Structure of matrix **X**

Pixels	**1**	**2**	**...**	**150**
1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,150}$
2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,150}$
...	...	...	...	...
150	$x_{150,1}$	$x_{150,2}$	...	$x_{150,150}$

8. Convert matrix **X** to vector. Each vector is then put in matrix of R component data called **Y**. The size of matrix **Y** is 161×22,500. The structure of **Y** is shown in Table 3.2.

Table 3.2 Structure of matrix **Y**

Observation	$\mathbf{Y_1}$	$\mathbf{Y_2}$	**...**	$\mathbf{Y_{22500}}$
zoomBlack_20180110_1341.png	$y_{1,1}$	$y_{1,2}$	...	$y_{1,22500}$
zoomBlack_20180111_1021.png	$y_{2,1}$	$y_{2,2}$	...	$y_{2,22500}$
...	...	...	...	...
zoomBlack_20180108_1401.png	$y_{160,1}$	$y_{160,2}$	...	$y_{160,22500}$
zoomBlack_20180108_1551.png	$y_{161,1}$	$y_{161,2}$	...	$y_{161,22500}$

9. Running HDDC on the dataset, with K starting from 2 to 10. The best number of cluster is selected using BIC value, and the average image for each cluster is displayed.
10. Running PCA and *K*-means on the dataset.

(i). Applying PCA to the dataset. Choosing number of components used for the next analysis.

(ii). Running *K*-means clustering analysis using selected PCs for *K* starting from 2 to 10. The result of each *K* is stored.

(iii). Calculating Pseudo-F and Silhouette of *K*-means result.

(iv). Comparing the value obtained from (iii) for evaluating *K*-means result.

11. Determining the number of *K* for the next step using Pseudo-F and Silhouette value from step 10(iii) and BIC from step 9.

12. Running modified HDDC with $K = 2$ and $K = 3$ on the dataset.

(i). Selecting 70% of the total observations for each of bootstrap replications.

(ii). Running modified HDDC on selected data. Calculating average value of each cluster and sorting the cluster number by smallest to largest.

(iii). Storing the result on matrix **B**. The dimension of B is 161×1000. The result is stored based on the replication sequence, with the remaining unselected images of 49 labelled as NA.

(iv). Repeating steps (i)-(iii) for 1000 replication.

(v). Determining final cluster of each observation using majority vote.

13. Comparing the result of modified HDDC using 2 and 3 clusters.

CHAPTER 4
RESULTS AND DISCUSSION

4.1 Characteristics of Precipitation in ITS Surabaya

Precipitation data in ITS is obtained from observational rain gauge installed in ITS in which the record is updated every five minutes. To match the frequency of radar images, precipitation data is aggregated to ten minutes period. Because the image is time-stamped in Coordinated Universal Time (UTC) and the precipitation data is recorded in Indonesia Western Standard Time (WIB, UTC+7), a time adjustment is made for the precipitation data. The time series plot of the precipitation in the period of 10 minutes is depicted in Figure 4.1.

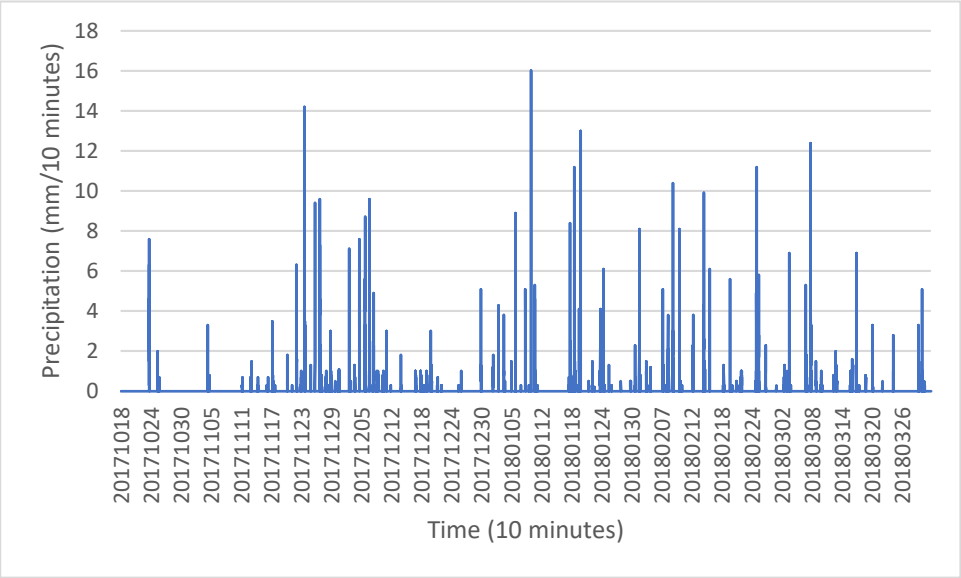
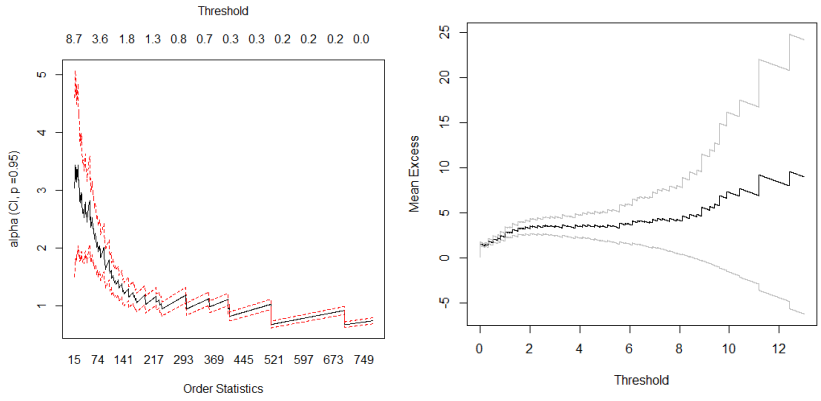


Figure 4.1 Time series plot of aggregated precipitation in ITS

Aggregated precipitation in Figure 4.1 shows that precipitation in Surabaya is relatively low in October and increased in November until mid-February. Precipitation in late-February and March is gradually lower, which is consistent with the ending of wet season in Indonesia.

After being aggregated for 10 minutes to match the time span of radar images, the next step is finding threshold for extreme precipitation. BMKG (2018) described that precipitation is categorized as heavy rainfall if the intensity is greater than 10 mm/hr. Since the dataset is updated every 10 minutes, the average of rain

intensity for each 10 minutes is around 1.67 mm. To check the proper threshold for the analysis, Mean Residual Life Plot (MRLP) and Hill Plot for the aggregated data is shown in Figure 4.2.



(a) (b)

Figure 4.2 (a) Hill Plot (b) Mean Residual Life Plot (MRLP)

According to the Hill plot of aggregated data, shown in Figure 4.2 (a), the threshold for heavy precipitation is between 1.3 mm and 1.8 mm. Meanwhile, from MRLP in Figure 4.2 (b), the point of where the plot look linear cannot be seen clearly, but from the plot it suggests that the threshold would be somewhere between 1 and 2 mm. Determining threshold value by using plots is relatively subjective, as it needed judgement and common sense (Embrechts, Klüppelberg, & Mikosch, 1997). Another value to consider for threshold selection is percentile value. For non-zero data (time where precipitation happened), the upper percentile are given in Table 4.1.

Table 4.1 Percentile of precipitation data

Percentile				
75%	80%	85%	90%	95%
1.125	1.5	2	3	4.3

Based on the previous value of average per ten minutes and value suggested by MRLP and Hill plot, the value of 80% percentile is inside the threshold range. Therefore, the value precipitation of 1.5 mm per ten minutes is set as threshold of extreme precipitation in ITS, Surabaya. However, when this value is calculated to an hour period, 1.5×6 is equal to 9 mm/hr. This value is below the range of heavy

precipitation in the definition of BMKG, and actually is in the upper bound of medium rain.

The next step is to check whether threshold of 1.5 mm per 10 minutes is a good fit for the precipitation data in ITS. After fitting the data into Generalized Pareto Distribution, the shape parameter of the precipitation data is -0.114 and the scale parameter is 3.471. Figure 4.3 shows plot of excess distribution of GPD of the precipitation data compared to its empirical value.

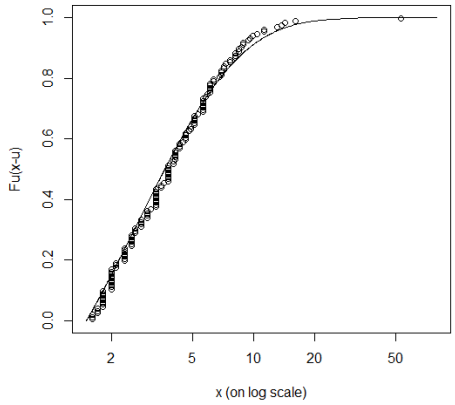


Figure 4.3 Plot of excess distribution of GPD of the precipitation data to its empirical value

Based on Figure 4.3, it can be seen that GPD function fits well to the empirical value, therefore the choice of threshold 1.5 mm/10 minutes is suitable for determining heavy precipitation data in ITS. There are 193 observations of precipitation selected by using this threshold. The date and time for selected data is then noted for selecting radar images of heavy precipitation in Surabaya.

4.2 Preprocessing of Radar Images

Radar images contain different kind of data structure compared to the usual dataset. Image is defined as instant illusion of a picture on flat surface. Picture are set of pixels, the smallest element of picture, arranged in rectangular array to form a complete image. Each pixel is represented in RGB (abbreviation for Red, Green, Blue), indicating how much of each red, green and blue included in the color (Graf, 1999).

The size of radar images in East Java is 600×800 pixels, covering entire East Java area. The center of the image is the weather radar located in BMKG Juanda.

The radius of radar coverage is 240 km (BMKG, 2018). However, because the analysis is focused on Surabaya and the image selection process is done using precipitation data in ITS Surabaya, using the whole image is not reasonable. Precipitation in Banyuwangi, for example, is less likely to be affected by precipitation in Surabaya. Therefore, the image is cut into smaller area, focusing in Surabaya. The illustration of radar image of November 22, 2017 at 9.32 UTC can be seen on Figure 4.4.

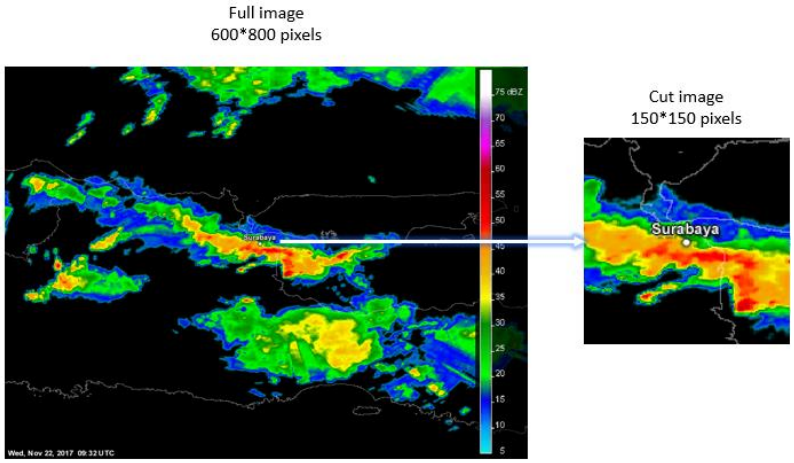


Figure 4.4 Process of selecting Surabaya area

The image is then separated into three component in RGB color model. The three components are Red (R), Green (G) and Blue (B). Each component has values of 0-255. The value of RGB is often denoted in (R,G,B) format. For example, the RGB value of white is (255,255,255) and black is (0,0,0). The RGB value of red is (255,0,0), green is (0,255,0) and blue is (0,0,255). With each image having three sets of data, the number of data for analysis are multiplied by three. Table 4.2 gives information about the legend in radar images and its color in R, G and B representation.

Table 4.2 Properties of legend in radar image

Image Properties	
dBZ	5 10 15 20 25 30 35 40 45 50 55 60 65
Legend	
R component	
G component	
B component	

Based on the description of dBZ scale in BMKG website, the color gradation in the image contains information about intensity of rainfall. Heavy precipitation is denoted by dBZ scale of 48 and above, and it is red in color. Out of the three components, G and B components mainly consist of zero values for the mentioned dBZ range, so the two components were not closely related to heavy precipitation event. On the other hand, R components in the dBZ range of 48 and above contain non-black color. Figure 4.5 shows comparison of images in RGB, R, G and B component.

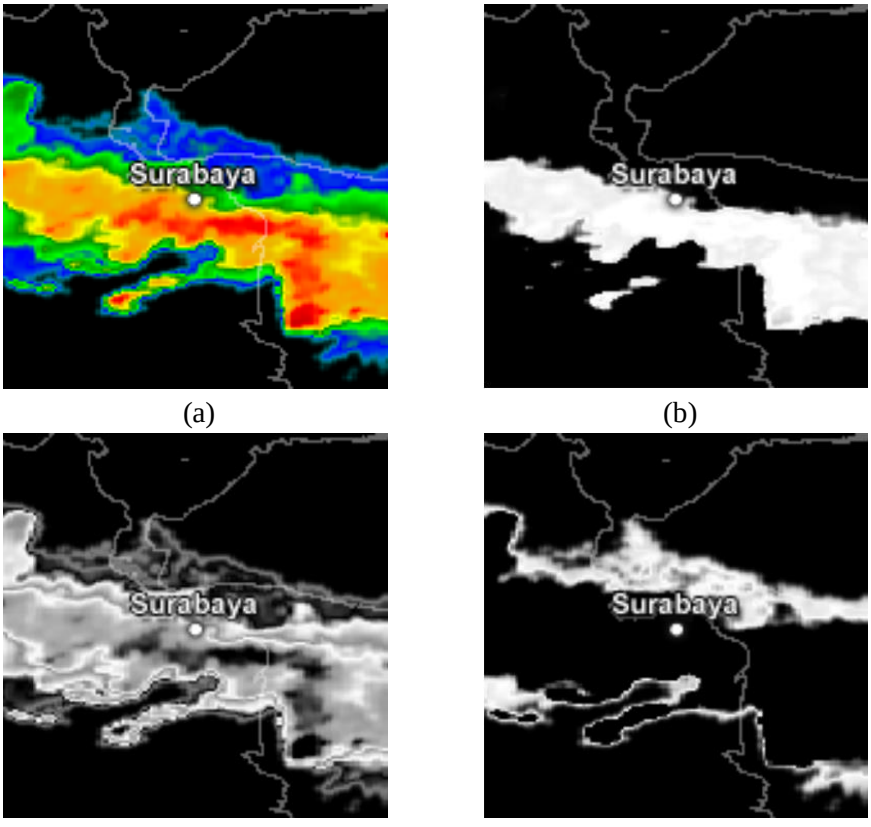


Figure 4.5 (a) RGB image (b) R component (c) G component (d) B component

Figure 4.5 (a) shows the cut image in Surabaya. From the picture, it can be seen that R area are spread across the maps from west to east. This red and orange area is where heavy precipitation happened. From the three images in (b), (c) and (d), the central area is represented better using R, in accordance with the result of identification in Table 4.2. Therefore, to simplify the process of analysis, only data from R component of the image will be used for further analysis.

The data from R component is in rectangular array sized 150×150, listed by its names containing information about the date and time for each images. The selected date and images from Section 4.1 is then used for selecting radar images. From 193 selected dates, there are only 161 images available. The 161 images are used as final data for clustering to check the shape of precipitation system. Figure 4.6 shows several chosen images in R component by matching the selected extreme dates to the radar images.

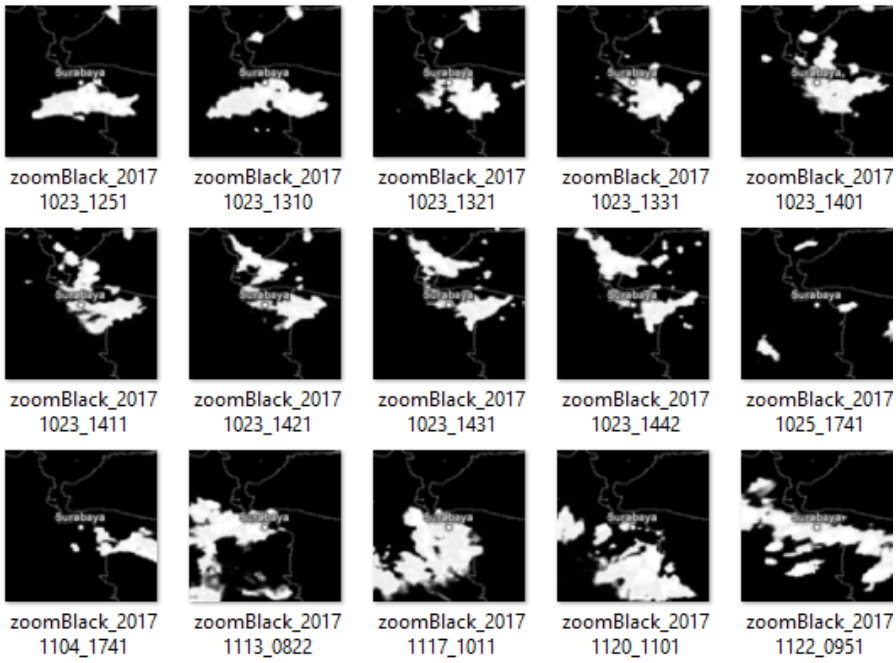


Figure 4.6 Several chosen images from the threshold

The complete selected images can be found in Enclosure 9. The focus in this study is to find general shape of precipitation system of heavy precipitation event in East Surabaya.

4.3 Gaussian Mixture Model for Heavy Precipitation Radar Images in Surabaya

Image data, as a set of pixels, is usually classified as high dimensional data. In the case of heavy precipitation radar images data, there are 161 observations with 22,500 features. Bouveyron (2007b) found that the classical GMM show a disappointing behavior when the size of the dataset is too small compared to the numbers of parameter to be estimated. Therefore, he proposed a specialized

Gaussian Mixture Model called High Dimensional Data Clustering (HDDC). The advantage of HDDC over the classical GMM is in parameter estimation. By using the subspace clustering in GMM, HDDC allowed for estimating less number of parameters. The comparison between number of parameters in the HDDC model and Classical GMM are shown in Table 4.3.

Table 4.3 Comparison between number of parameters between HDDC and Classical GMM

Number of clusters (*K*)	HDDC (model $[a_{kj}b_kQ_kd_k]$)	Classical GMM
2	157,501	506,317,501
3	202,504	759,476,252
4	427,484	1,012,635,003
5	607,440	1,265,793,754
6	674,964	1,518,952,505
7	1,124,813	1,772,111,256
8	1,327,303	2,025,270,007
9	1,372,356	2,278,428,758
10	1,484,809	2,531,587,509

The numbers for HDDC model shown in Table 4.3 were calculated using formula shown in Table 2.1, with the details of d_k for each cluster shown in Enclosure 10. From the numbers in Table 4.3, it can be seen that the number of parameters that need to be estimated in HDDC are much less than parameters for Classical GMM. The result of HDDC for *K* starting from 2 to 10 can be compared using the value of BIC generated by each model. Details on the BIC value of the HDDC is shown in Table 4.4.

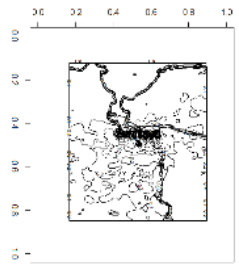
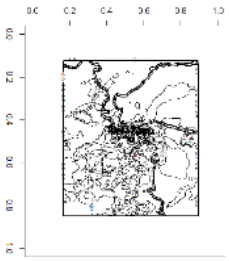
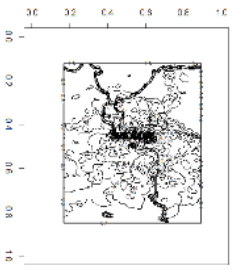
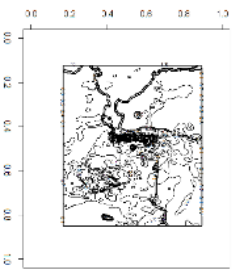
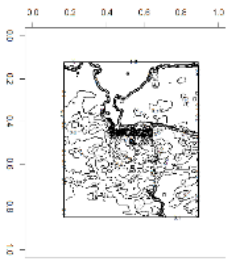
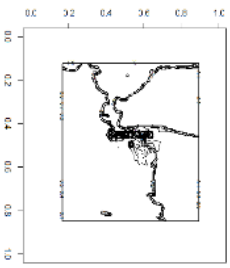
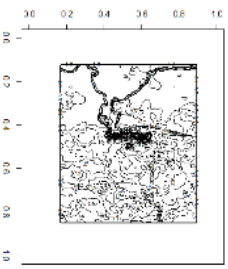
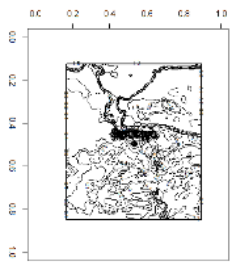
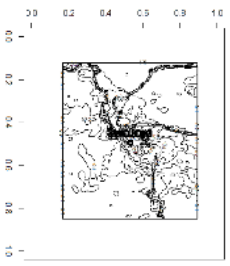
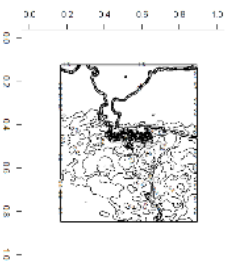
Table 4.4 Result of HDDC in radar image data

Number of clusters	BIC
2	39,636,666
3	39,466,383
4	39,932,456
5	40,156,657
6	36,619,357
7	27,589,760
8	27,960586
9	27,398,505
10	22,296,092

Based on the lowest BIC value shown in Table 4.4, the result shows 10 clusters as the optimum clusters. The images were grouped into their respective

cluster, and the average image for each cluster were displayed. The average images are displayed in contour plot in order to get clearer information about pattern shown in the image. The result for HDDC with 10 clusters are shown in Table 4.5.

Table 4.5 Contour plot of cluster member for 10 clusters in HDDC

Cluster 1	Cluster 2	Cluster 3	Cluster 4
			
Cluster 5	**Cluster 6**	**Cluster 7**	**Cluster 8**
			
	Cluster 9	**Cluster 10**	
			

Based on the result of clustering stored in Table 4.5, it can be seen that several clusters shows similar pattern. For example, pattern in Cluster 1 was similar to Cluster 5, and pattern in Cluster 2 was similar to Cluster 4. The only cluster with clear different pattern is Cluster 7, with small circle on the eastern side of Surabaya. Because of these similar patterns, it could be concluded that 10 clusters are not suitable for the radar image data in Surabaya. It is suggested that the number of optimal clusters should be smaller than 10, to avoid the grouping of similar pattern.

There is also second problem of HDDC, with the method showing inconsistency of cluster member. When the process is repeated, an image could be clustered into different cluster.

4.4 PCA and *K*-means clustering for Heavy Precipitation Radar Images in Surabaya

To evaluate the result of HDDC, another clustering method will be applied to the data. *K*-means is one of the most popular clustering method. However, *K*-means is distance-based clustering method, therefore it often does not work well for high dimensional data. This problem is caused by the curse of dimensionality, mainly for *K*-means with squared Euclidean distance. Aside than using alternative distance function, another way to solve the problem of high dimensionality for *K*-means is by employing dimension reduction. One of the way to do this is by using Principal Component Analysis (PCA) (Wu, 2012). The complete result of PCA is written in Enclosure 11. Selecting the number of components is an essential step for PCA. The usual method for selecting the number of components is by using the help of scree graph. Figure 4.7 shows screeplot for the first 50 PCs.

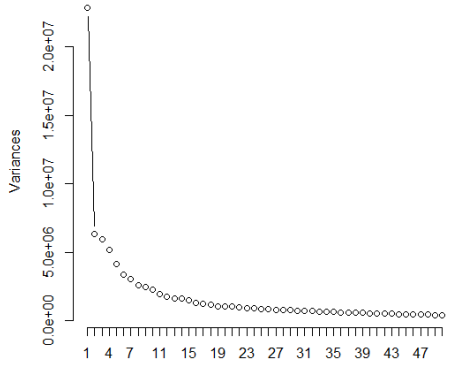


Figure 4.7 Screeplot for first 50 PCs

The first PC explain 20.92% variance of the data, which is the largest proportion out of all the 161 PCs. However, using only 20.92% is too small to represent all the data, therefore additional component are used. By using 41 components, 80,15% of variance in the selected radar images are explained. The new variable obtained by using the chosen 41 PCs were used for *K*-means clustering. The result of PCA and *K*-means for radar image data is shown in Table 4.6.

Table 4.6 Result of PCA and *K*-means for radar image data

Number of clusters	Silhouette	Pseudo-F
2	0.33	37.74
3	0.25	26.56
4	0.23	22.71
5	0.23	19.43
6	0.21	17.74
7	0.22	16.97
8	0.23	16.47
9	0.18	14.93
10	0.19	14.52

Table 4.6 shows that average Silhouette value for 2 clusters is the largest, with the trend of decreasing with the increase of cluster number. Similar conclusion can be drawn for the value of Pseudo-F. However, the Silhouette values for *K* starting from 3 to 10 are similar and the Pseudo-F values for *K* starting from 5 to 10 are also in the similar range, so taking consideration of GMM result and in order to compare the results generated by both methods, 10 clusters will be used. The contour plots for average images for each cluster are shown in Table 4.7.

Table 4.7 Contour plot of cluster member for 10 clusters in PCA and *K*-means

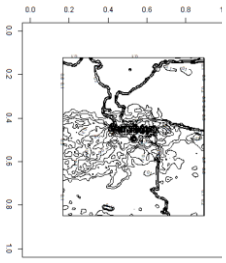
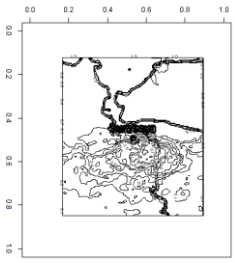
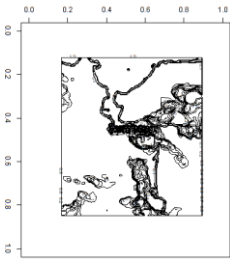
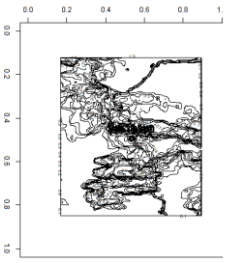
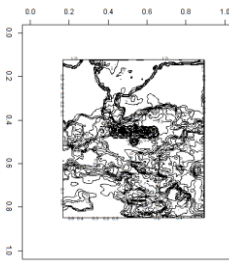
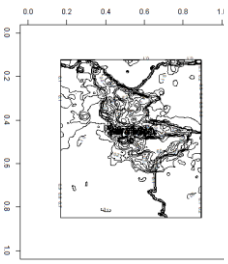
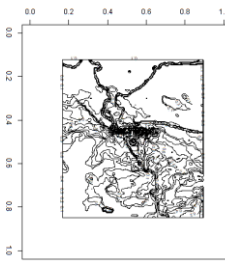
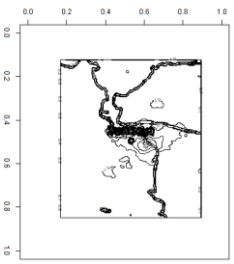
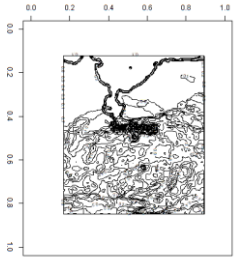
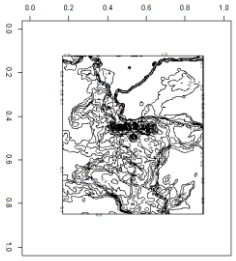
Cluster 1	Cluster 2	Cluster 3	Cluster 4
			
Cluster 5	**Cluster 6**	**Cluster 7**	**Cluster 8**
			

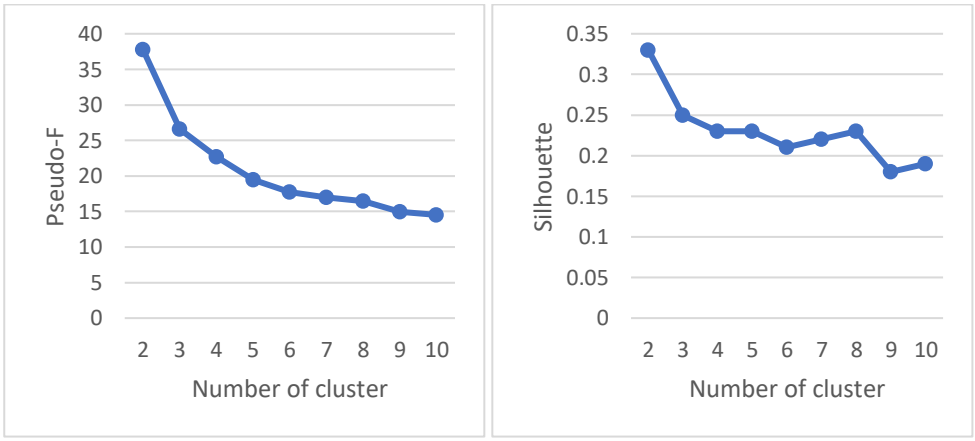
Table 4.7 Contour plot of cluster member for 10 clusters in PCA and *K*-means (cont'd)

Cluster 9	Cluster 10
	

Based on figures shown in Table 4.7, there are also clusters with similar pattern. For example, cluster 1 and cluster 2. There is also similarity between cluster 6, cluster 8 and cluster 10. The result of *K*-means are showing the same conclusion with HDDC. *K*-means also suggested that the number of optimal clusters should be smaller than 10. The second problem faced by *K*-means clustering is also in terms of inconsistency of cluster member.

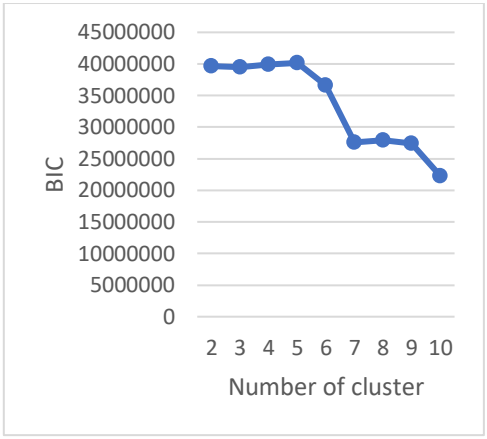
4.5 Modified High Dimensional Data Clustering

The previous description about the result of HDDC and *K*-means suggest that 10 clusters are noticeably too big for clustering the radar images. In order to determine best number of cluster, the result of HDDC and *K*-means for 2 to 10 clusters will be reviewed. Figure 4.8 shows the plot of Pseudo-F and Silhouette of *K*-means and BIC for HDDC for *K* starting from 2 to 10.



(a) (b)

Figure 4.8 Evaluation criteria for clustering result (a) Pseudo-F (b) Silhouette (c) BIC



(c)

Figure 4.8 Evaluation criteria for clustering result (a) Pseudo-F (b) Silhouette (c) BIC (cont'd)

In Figure 4.8 (a), the highest value of Pseudo-F was obtained when $K = 2$ clusters. When $K = 3$ clusters, the value of Pseudo-F dropped. This was the largest drop of Pseudo-F value, suggesting that 2 clusters might be suitable for radar images. Looking at Figure 4.8 (b), the largest Silhouette value also happened when $K = 2$. Because 2 clusters were more reasonable and supported by the value of Pseudo-F and Silhouette, it is decided that the method will be applied using $K = 2$.

The other problem of HDDC and *K*-means is inconsistency of cluster member. With the observation classified into different cluster whenever the process is replicated, it is difficult to get a conclusion of which cluster exactly is the pattern belong. To fix this problem, a new method is proposed, i.e. by applying ensemble concept to HDDC. The HDDC is chosen because *K*-means require PCA for reducing the dimension of the data. PCA as a form of feature transformation is indeed useful for identifying important features but in the case of high dimensional data, it does not help since the relative distance and the effect of irrelevant dimension were still there (Parsons, Haque, & Liu, 2004).

In the modified HDDC, bootstrap resampling process were applied to the data to select sample from the observations. Because the total observations are 161 images, the sample for each replication is decided to be less than the total observations. Each replication would select 70% number of sample from total data, about 112 images, to build an HDDC model with $K = 2$. The illustration for the resampling process is shown in Figure 4.9.

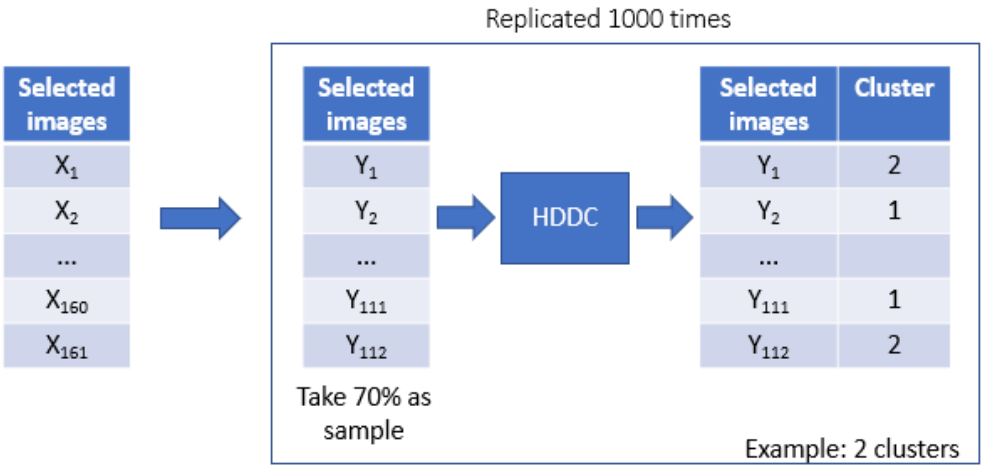


Figure 4.9 Illustration for resampling process of HDDC

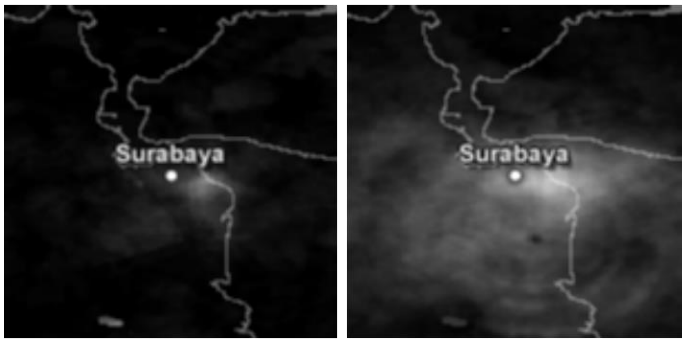
Each replication resulted in cluster member for each observation. After replicating the process for 1,000 times, there would be hundreds of cluster member for each observation, and the majority vote were used to decide which cluster does the observation belong to. In the case of $K = 2$, the illustration of this process is shown in Figure 4.10.

Selected images	Rep 1	Rep 2	...	Rep 999	Rep 1000	Final cluster
X_1	2	NA	...	2	2	**2**
X_2	NA	1	...	2	1	**1**
...	...	...	...	...	...	**...**
X_{50}	2	NA	...	1	NA	**2**
X_{51}	2	1	...	NA	2	**2**
...	...	...	...	...	...	**...**
X_{100}	1	1	...	2	1	**1**
X_{101}	1	1	...	NA	2	**2**
X_{102}	1	1	...	2	2	**2**
...	...	...	...	...	...	**...**
X_{160}	1	2	...	1	NA	**2**
X_{161}	2	2	...	NA	2	**2**

Final cluster chosen by majority vote

Figure 4.10 Illustration for selecting final cluster of modified HDDC

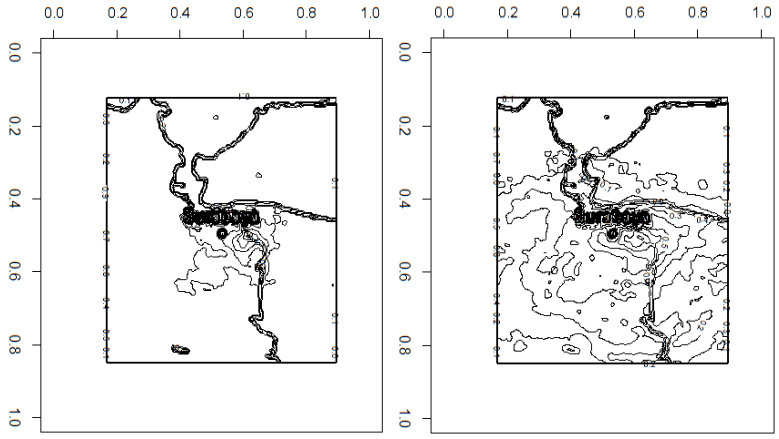
From the total of 161 images, there were 70 images belong to cluster 1 and 91 images belong to cluster 2. Figure 4.11 shows the average image of radar images for each cluster.



(a) (b)

Figure 4.11 Average image of cluster member of (a) Cluster 1 and (b) Cluster 2

The image in Figure 4.11 (a) shows small white area near the eastern part of Surabaya, and Figure 4.11 (b) shows white area in wider and bigger area compared to Cluster 1. However, the difference between clusters in terms of their shapes cannot be seen clearly using the average image. Therefore, contour plot will be employed to help identify the shape. The contour plot of average images of both clusters are shown in Figure 4.12.



(a) (b)

Figure 4.12 Contour plot of average image of (a) Cluster 1 and (b) Cluster 2

Based on contour plot in Figure 4.12, the pattern of the shape is clearer than the average image in Figure 4.11. Figure 4.12 (a) suggested that the shape in Cluster 1 is smaller circle, with the center part of the circle is in eastern part of Surabaya. This is expected because the heavy precipitation is selected using the precipitation data of ITS, which is located in East Surabaya. Meanwhile, Figure 4.12 (b) shows bigger circle and more smaller circle inside, spreading across Surabaya area.The figure suggests that the pattern in this cluster mostly are small circle near center

part of Surabaya, several big circles can be seen in the plot. The range of circle is wider compared to the contour image in cluster 1, with the outermost part of the circle is spread across the area of Surabaya, from the western part to the eastern part of Surabaya.

The next point in the analysis is duration of heavy precipitation event. Several images are a sequence of events happening consecutively, therefore the duration of heavy precipitation event can be calculated by looking at the time of heavy precipitation event started until it ended. The comparison of precipitation duration between clusters are shown in Figure 4.10.

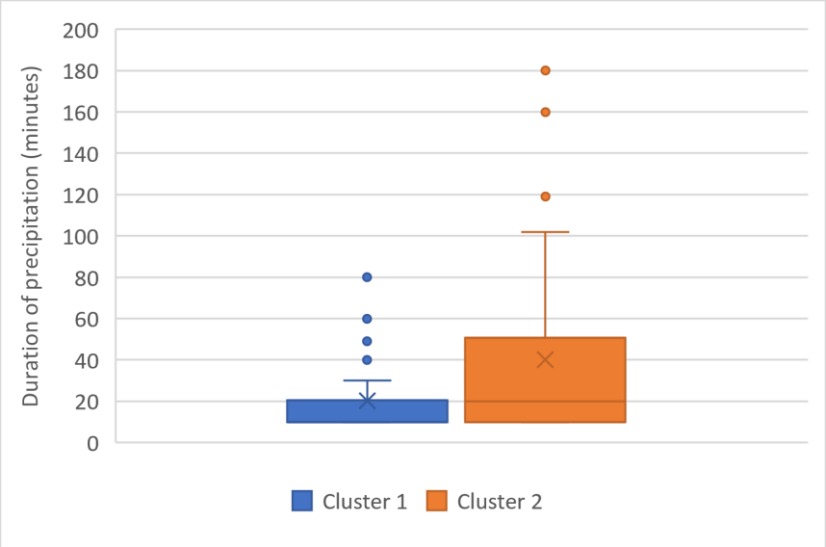


Figure 4.13 Comparison of precipitation duration between clusters in $K = 2$

Based on Figure 4.13, it can be seen that the average duration of precipitation between clusters are slightly different, though it is not statistically significant. The average of precipitation duration is 20.49 minutes for Cluster 1 and 40.8 minutes for Cluster 2. In other words, if the shape of cloud in radar images belong to cluster 1, then there will be heavy precipitation happening for an average of 20.49 minutes. Meanwhile, if the shape belong to Cluster 2, the heavy precipitation will last longer, with the average of 40.8 minutes.

Based on the shape of precipitation system captured in the average image of each cluster member, Cluster 1 is a cluster for radar images on the occasion of heavy precipitation with small circle shape and short period of heavy rain. Meanwhile for

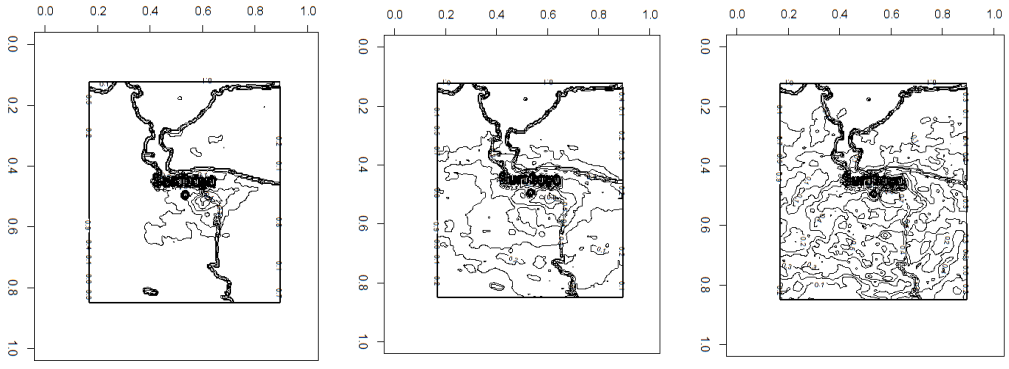
Cluster 2, the area of the circle is bigger, almost covering entire area of Surabaya and the period of rain is longer.

To further confirm whether $K = 2$ is suitable for clustering the pattern in heavy precipitation radar images in Surabaya, the HDDC is then applied again to the data for $K = 3$. By using three clusters, there were 63 images belong to cluster 1, 76 images belong to cluster 2, and 22 images belong to cluster 3. Figure 4.14 shows the average image of radar images for each cluster in $K = 3$.

(a) (b) (c)

Figure 4.14 Average image of cluster member of (a) Cluster 1 (b) Cluster 2, and (c) Cluster 3

The image in Figure 4.14 (a) is similar to the image in Figure 4.11 (a) and the image in Figure 4.14 (b) is similar to the image in Figure 4.11 (b). The white area in the image for Cluster 3 in Figure 4.14 (c) is bigger than Cluster 2, showing larger area of heavy precipitation. The contour plot of average image of the three clusters are shown in Figure 4.15.



(a) (b) (c)

Figure 4.15 Contour plot of average image of (a) Cluster 1, (b) Cluster 2 and (c) Cluster 3

Figure 4.15 (a) and (b) suggested that the pattern is similar to the contour plot shown in Figure 4.12. The area with bigger circle and more complicated pattern

that belong to Cluster 3 is shown in Figure 4.15 (c). The pattern in Cluster 3 almost covered entire maps area. Figure 4.10 depicted the comparison of precipitation duration between clusters are shown in.

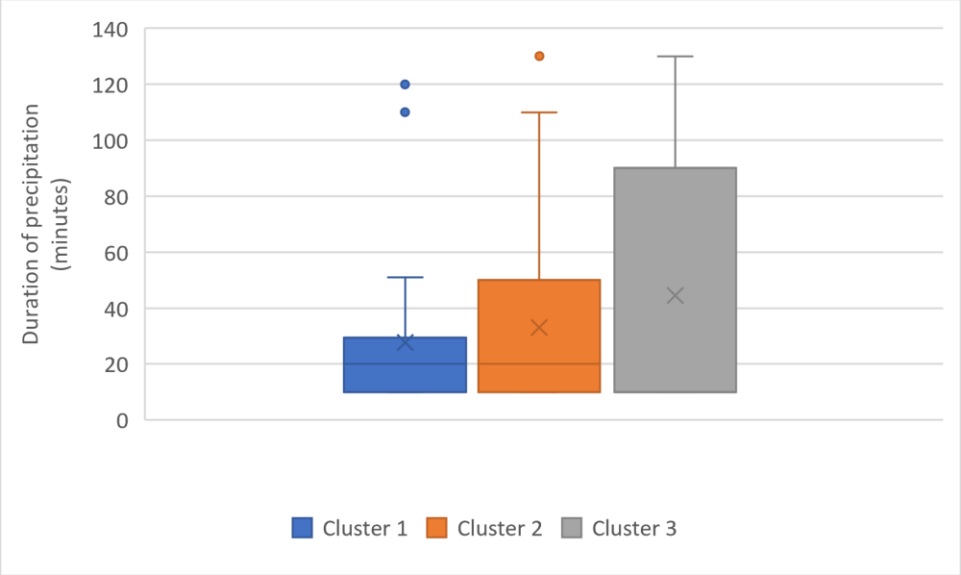


Figure 4.16 Comparison of precipitation duration between clusters in $K = 3$

Based on Figure 4.16, the average duration of precipitation between the three clusters are also slightly different but not statistically significant. The average of precipitation duration is 27.61 minutes for Cluster 1, 33.03 minutes for Cluster 2, and 44.56 minutes for Cluster 3. However, the range of precipitation duration in Cluster 2 and Cluster 3 did not differ much. The number of radar images in Cluster 3 is also the smallest between clusters. With the similar pattern and similar range of precipitation duration between Cluster 2 and Cluster 3, it can be concluded that the addition of another cluster did not give significant change on the clustering result of radar images. Hence, the suitable number of cluster for the heavy precipitation radar images are 2 clusters.

(this page is intentionally left blank)

CHAPTER 5
CONCLUSION AND SUGGESTION

5.1 Conclusion

A threshold of 1.5 mm per 10 minutes was determined for heavy precipitation in ITS. This threshold was then used as criteria for selecting radar images for the clustering process to identify the shape of precipitation system in Surabaya, East Java. The result of both HDDC and *K*-means came with 10 clusters which was noticeably too big for radar images data, as there were several clusters having the same pattern of precipitation system. The second problem with those two method was inconsistent cluster member when the analysis is replicated. To solve this problem, ensemble concept was applied to HDDC. By using 2 clusters, this method provided consistent cluster member. In addition, there were remarkable different characteristics found in each cluster. The first cluster was represented by small-shaped precipitation system in the center of Surabaya with shorter duration of heavy precipitation. The second cluster had bigger circle-shaped precipitation system, almost covering the entire area of Surabaya and had longer duration of heavy precipitation.

5.2 Suggestion

The problem faced in this research is the lack of dates and time of heavy precipitation event. The current data used in this research is precipitation data in ITS, so the precipitation recorded in this dataset is only rainfall happened in ITS area. Surabaya is a big city consisted of several districts, and the change of weather in some districts can be really different in similar time. The addition of precipitation data in other area of Surabaya will help on increasing the number of selected samples and will help to capture more shapes of precipitation system in Surabaya.

(this page is intentionally left blank)

REFERENCES

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 433-459. doi: 10.1002/wics.101

AghaKouchak, A., Nasrollahi, N., Li, J., Imam, B., & Sorooshian, S. (2010). Geometrical Characterization of Precipitation Patterns. Journal of Hydrometeorology, 12, 274-285.

Aldrian, E., & Susanto, R. D. (2003). Identification of three dominant rainfall regions within Indonesia and their relationship to sea surface temperature. Int J. Climatol, 23, 1435-1452.

Alfarano, S., & Lux, T. (2010). Extreme Value Theory as a Theoretical Background for Power Law Behavior. Kiel Working Paper 1648.

American Meteorological Society. (2018, 05 23). Precipitation. Retrieved from Glossary of Meteorology: http://glossary.ametsoc.org/wiki/Precipitation

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., & Gottardo, R. (2010). Combining Mixture Components for Clustering. Journal of Computational and Graphical Statistics, 19(2), 332-353.

Beirlant, J., Goegebeur, Y., Teugels, J., Waal, D. D., & Ferro, C. (2014). Statistics of Extremes. Chichester, West Sussex, England: John Wiley & Sons.

Bergé, L., Bouveyron, C., & Girard, S. (2012). HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High Dimensional Data. Journal of Statistical Software, 46(6). doi:10.18637/jss.v046.i06

BMKG. (2017, 03 14-16). Review of current observations in the MC Region: Current status of BMKG observing systems and data products. Accessed 06 04, 2018, from https://cdn.bmkg.go.id/Web/Day1_11_DonaldiPermana.pdf

BMKG. (2018). Citra Radar | BMKG. Taken from https://www.bmkg.go.id/cuaca/citra-radar.bmkg

Bommier, E. (2014). Peaks-Over-Threshold Modelling of Environmental Data. Taken from https://uu.diva-portal.org/smash/get/diva2:760802/FULLTEXT01.pdf

Bouveyron, C., Girard, S., & Schmid, C. (2007a). High-Dimensional Data Clustering. Computational Statistics & Data Analysis, 52(1), 502-519.

Bouveyron, C., Girard, S., & Schmid, C. (2007b). High-Dimensional Discriminant Analysis. Communications in Statistics: Theory and Methods, 36(14), 2607-2623.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1-27.

Coles, S. (2001). An Introduction to Statistical Modeling of Extreme Values. London: Springer.

Cooley, D. (2009). Extreme value analysis and the study of climate change. D. Climatic Change(97), 77-83.

Cooley, D. S. (2005). Statistical Analysis of Extremes Motivated by Weather and Climate Studies: Applied and Theoretical Advances. PhD theses, University of Colorado.

D'Arrigo, R., & Wilson, R. (2008). El Nino and Indian Ocean influences on Indonesian drought: implications for forecasting rainfall and crop productivity. International Journal of Climatology, 28, 611-616. doi:10.1002/joc.1654

Davison, A. C., & Smith, R. L. (1990). Models for Exceedances over High Thresholds. Journal of the Royal Statistical Society. Series B (Methodological), 52(3), 393-442. Retrieved from www.jstor.org/stable/2345667

Desgraupes, B. (2013). Clustering Indices. University of Paris Ouest - Lab Modal'X, 1-34.

Drees, H., Haan, L. D., & Resnick, S. (2000). How to make a hill plot. The Annals of Statistics, 28(1), 254-274.

Efron, B., & Tibshirani, R. (1993). An Introduction to the Bootstrap. London: Chapman & Hall.

Embrechts, P., Klüppelberg, C., & Mikosch, T. (1997). Modelling Extremal Events for Insurance and Finance (1 ed.). Verlag Berlin Heidelberg: Springer.

Fisher, R. A., & Tippett, L. H. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. Mathematical Proceedings of

the Cambridge Philosophical Society, 24, pp. 180-290. doi:10.1017/S0305004100015681

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). Bayesian Data Analysis Third Edition. Boca Raton: Chapman & Hall/CRC Press.

Ghosh, S., & Resnick, S. (2010). A discussion on mean excess plots. Stochastic Process and their Applications(120), 1492-1517.

Gilli, M., & Kellezi, E. (2006). An Application of Extreme Value Theory for Measuring Financial Risk. Computational Economics, 27(1), 1-23.

Goldstein, J., Mirza, M., Etkin, D., & Milton, J. (2003). Hydrologic assessment: Application of extreme value theory for climate extremes scenarios construction. 14th symposium on global change and climate variations, American meteorological society 83rd annual meeting.

Graf, R. F. (1999). Modern Dictionary of Electronics (7th Edition ed.). USA: Newnes.

Harrison, D. L., Driscoll, S. J., & Kitchen, M. (2000). Improving precipitation estimates from weather radar usign quality control and correction techniques. Meteorol. Appl., 6, 135-144.

Hendon, H. H. (2003). Indonesian Rainfall Variability: Impacts of ENSO and Local Air-Sea Interaction. Journal of Climate, 16, 1775-1790.

Hidayat, R., & Kizu, S. (2010). Influence of the Madden-Julian Oscillation on Indonesian rainfall variability in austral summer. Int. J. Climatol., 30, 1816-1825. doi:10.1002/joc.2005

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. Annals of Statistics 3, 5, 1163-74.

Ilhamsyah, Y. (2013). The utilization of weather and climate information to support marine and fisheries activities. DEPIK, 2(3), 200-209.

Imaduddina, A. H., & W, W. H. (2014). Sea Level Rise Flood Zones: Mitigating Floods in Surabaya Coastal Area. Procedia - Social and Behavioral Sciences, 135, 123-129.

Islam, M. N., Hayashi, T., Terao, T., Uyeda, H., & Kikuchi, K. (2005). Characteristics of Precipitation Systems Analyzed from Radar Data over Bangladesh. Journal of Natural Disaster Science, 27(1), 17-23.

Johnson, R. A., & Wichern, D. W. (2007). Applied Multivariate Statistical Analysis Sixth Edition. Upper Saddle River, NJ: Pearson Prentice Hall.

Joliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical Transactions of Royal Society A, 374. doi:10.1098/rsta.2015.0202

Jolliffe, I. T. (2002). Principal Component Analysis (2nd ed.). Verlag New York: Springer. doi:10.1007/b98835

Kalti, K., & Mahjoub, M. (2014). Image Segmentation by Gaussian Mixture Models and Modified FCM Algorithm. The International Arab Journal of Information Technology, 11(1), 11-18.

Kodinariya, T. M., & Makwana, D. P. (2013). Review on determining number of cluster in K-means Clustering. International Journal of Advance Research in Computer Science and Management Studies, 1(6).

KOMPAS. (2017a). Banjir Terjang Sejumlah Kawasan di Surabaya. Accessed 06 04, 2018, from https://regional.kompas.com/read/2017/02/17/20453061/banjir.terjang.sejumlah.kawasan.di.surabaya

KOMPAS. (2017b). Banjir Hingga 50 cm di Surabaya, Risma Sebut Banyak Sampah di Lokasi. Accessed 06 04, 2018, from https://regional.kompas.com/read/2017/11/25/18413441/banjir-hingga-50-cm-di-surabaya-risma-sebut-banyak-sampah-di-lokasi

Langousis, A., Mamalakis, A., Puliga, M., & Deidda, R. (2016). Threshold detection for the generalized Pareto distribution: Review of representative methods and application to the NOAA NCDC daily rainfall database. Water Resources Research, 52(4), 2659-2681. doi:10.1002/2015WR018502

Leadbetter, M. R. (1991). On a basis for 'Peaks over Threshold' modeling. Statistics & Probability Letters, 12, 357-362.

Lee, S., & Kim, J. H. (2018). Exponentiated generated Pareto distribution: Properties and applications towards extreme value theory. Communications in Statistics-Theory and Methods. doi:10.1080/03610926.2018.1441418

Ling, H., & Zhu, K. (2017). Predicting Precipitation Events Using Gaussian Mixture Model. Journal of Data Analysis and Information Processing, 5, 131-139.

Marimotou, V., Raggad, B., & Trabelsi, A. (2006). Extreme Value Theory and Value at Risk: Application to Oil Market. GREQAM Working Paper No. 2006-38.

Montfort, M. A., & Witter, J. V. (1986). The Generalized Pareto distribution applied to rainfall depths. Hydrological Sciences Journal, 31(2), 151-162. doi:10.1080/02626668609491037

Ostojic, D. R., Bose, R. K., Krambeck, H., Lim, J., & Zhang, Y. (2013). Energizing Green Cities in Southeast Asia. Washington DC: The World Bank.

Parsons, L., Haque, E., & Liu, H. (2004). Subspace clustering for high dimensional data: a review. ACM SIGKDD Explorations Newsletter - Special issue on learning from imbalanced datasets, 6(1), 90-105. doi:10.1145/1007730.1007731

Paski, J. A. (2017). Impact of remote sensing data assimilation (radar and satellite) on numerical weather prediction for rainfall estimation. Jurnal Penginderaan Jauh dan Pengolahan Data Citra Digital, 14(2), 79-88.

Pfaff, B., & McNeil, A. (2018). evir: Extreme Values in R. Retrieved from https://CRAN.R-project.org/package=evir

Rahayu, A. (2013). Identification of Climate Change with Generalized Extreme Value (GEV) Distribution Approach. Journal of Physics: Conference Series(423).

REPUBLIKA. (2016). Flood hits several areas in Surabaya. Accessed 06 04, 2018, from http://en.republika.co.id/berita/en/national-politics/16/05/31/o80vi4317-flood-hits-several-areas-in-surabaya

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

SAS Institute Inc. (2018). SAS/STAT(R) 9.22 User's Guide. Retrieved 06 04, 2018, from https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_cluster_sect021.htm

Schwarz, G. (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461-464.

Soldek, J., Saeed, K., & Pejas, J. (2012). Advanced Computer Systems: Eighth International Conference, ACS' 2001 Mielno, Poland October 17–19, 2001 Proceedings. Mielno: Springer Science & Business Media.

Supari, Tangang, F., Juneng, L., & Aldrian, E. (2016). Observed changes in extreme temperature and precipitation over Indonesia. Int J Climatol, 37, 1979-1997. doi:10.1002/joc.4829

TEMPO. (2016, 4 15). Diguyur Hujan 5 Jam, Surabaya Banjir. Accessed 7 11, 2018, from https://nasional.tempo.co/read/762868/diguyur-hujan-5-jam-surabaya-banjir

TEMPO. (2016). Surabaya Inundated with Flood Water. Accessed 06 04, 2018, from https://en.tempo.co/read/news/2016/05/31/307775400/Surabaya-Inundated-with-Flood-Water

Thompson, P., Cai, Y., Reeve, D., & Stander, J. (2009). Automated threshold selection methods for extreme wave analysis. Coastal Engineering, 56, 1012-1021.

Wang, P., Smeaton, A., Songyang, L., O'Connor, E., Ling, Y., & O'Connor, N. (2009). Short-term rainfall nowcasting: Using rainfall radar imaging. Eurographics Ireland 2009.

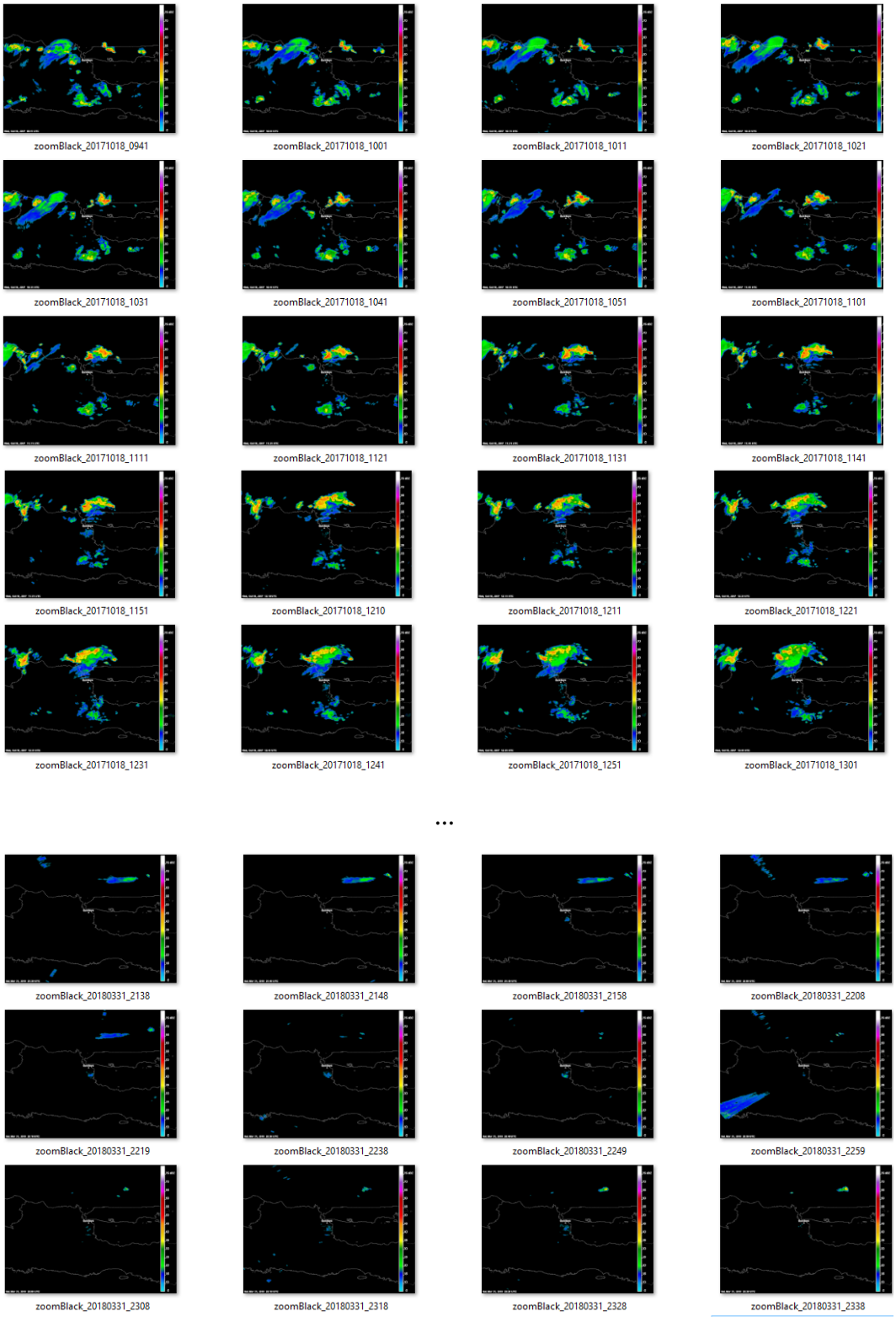
Wu, J. (2012). Advances in K-means Clustering. Berlin Heidelberg: Springer Theses. doi:10.1007/978-3-642-29807-3

ENCLOSURE

Enclosure 1. Precipitation data in ITS

No	Date	UTC	Precip Accumulation
1	2017-10-18	0:03	0
2	2017-10-18	0:09	0
3	2017-10-18	0:14	0
4	2017-10-18	0:19	0
5	2017-10-18	0:24	0
6	2017-10-18	0:30	0
7	2017-10-18	0:35	0
8	2017-10-18	0:40	0
9	2017-10-18	0:46	0
10	2017-10-18	0:51	0
11	2017-10-18	0:56	0
12	2017-10-18	1:02	0
13	2017-10-18	1:07	0
14	2017-10-18	1:12	0
15	2017-10-18	1:18	0
16	2017-10-18	1:23	0
⋮	⋮	⋮	⋮
41415	2018-03-31	22:18	0
41416	2018-03-31	22:24	0
41417	2018-03-31	22:29	0
41418	2018-03-31	22:34	0
41419	2018-03-31	22:40	0
41420	2018-03-31	22:45	0
41421	2018-03-31	22:50	0
41422	2018-03-31	22:55	0
41423	2018-03-31	23:01	0
41424	2018-03-31	23:06	0
41425	2018-03-31	23:11	0
41426	2018-03-31	23:17	0
41427	2018-03-31	23:22	0
41428	2018-03-31	23:27	0
41429	2018-03-31	23:33	0
41430	2018-03-31	23:38	0
41431	2018-03-31	23:43	0
41432	2018-03-31	23:49	0
41433	2018-03-31	23:55	0

Enclosure 2. Example of radar image



Enclosure 3. Syntax of R for preprocessing precipitation data and radar image

```
keputih=read.csv('d:/Radar/precipITS-UTC.csv',header=T)
#Data of precipitation in keputih
keputih[,2]=substr(gsub(":","",keputih[,2]),1,3)
keputih[,1]=gsub("-","",keputih[,1])

#Radar image
library(png)
library(grid)
library(gridExtra)
library(imager)
library(magick)

setwd('d:/Radar/colorcut')
library(data.table)
file1=list.files(pattern='.png',full.names=FALSE)
files=file1
files=gsub("zoomBlack_","",files)
files=gsub(".png","",files)
files=substr(files,1,12)

precip=aggregate(keputih[,4],list(keputih[,1],keputih[,2]),'sum')
time=paste(precip[,1],'_',substr(precip[,2],1,3))
time=gsub(" ","",time)
data=cbind(precip,time)

#data which has precip>=1.5
id=which(precip[,3]>=1.5)
n=length(id)
hprecip=matrix(nrow=n,ncol=3)
for (i in 1:n){
  num=which(files==time[id[i]])
  num=ifelse(length(num)>1,num[1],ifelse(length(num)<1,NA,num))
  hprecip[i,1]=num
  hprecip[i,2]=time[id[i]]
  hprecip[i,3]=precip[id[i],3]
}
hprecip=transform(hprecip)
colnames(hprecip)=c('num','name','prcp')
hprecip.fix=na.omit(hprecip)

#selecting image
setwd('d:/Radar/colorcut')
img=NULL
file.precip=file1[as.numeric(levels(hprecip[,1]))]
for (i in 1:length(file.precip)){
 img[[i]]=readPNG(file.precip[i])
}

#matrix data for R
dataclust=matrix(nrow=length(file.precip),ncol=150*150)
for (i in 1:length(file.precip)){
 dataclust[i,]=as.vector(img[[i]][,,1]*255)
}
rownames(dataclust)=file.precip

library(evir)
cek.gpd=read.csv('d:/precip161.csv',header=FALSE)
output=gpd(cek.gpd[,2],1.5)
plot.gpd(output)
param=output$par.ests
```

Enclosure 4. Syntax of R for modified HDDC

```
N=length(dataclust[,1])
n=floor(0.7*N)
no=c(1:N)
nama=rownames(dataclust)
res=matrix(ncol=100,nrow=N)
for (i in 1:100){
 sampel=sample(no,n)
 clust=hddc(dataclust[sampel,],K=2,init='mini-em')
 res.temp=clust$class
 nama.sampel=rownames(dataclust[sampel,])
 nama.id=match(nama,nama.sampel)
 for (j in 1:N){
  res[j,i]=res.temp[nama.id[j]]
 }
}
write.csv(res,'d:/result/2/r8.csv')
res=read.csv('d:/result/3/gmm3.csv',header=TRUE)

#sorting the cluster
res1=res
for (j in 1:100){
 m=max(res[,j],na.rm=TRUE)
 avg=matrix(nrow=m)
 for (i in 1:m){
  cl=which(res[,j]==i)
  avg[i]=mean(as.vector(dataclust[cl,]))
 }
 temp=sort(avg)
 for (k in 1:m){
  id.old=which(avg==temp[k])
  id.new=which(res[,j]==id.old)
  res1[id.new,j]=k
 }
}
write.csv(res1,'d:/result/3/gmm3-res1.csv')

#function for computing SS
ss <- function(x) sum(scale(x, scale = FALSE)^2)
ssclust=function(x,kelas){
totss=ss(x)
centers=matrix(ncol=length(x[1,]),nrow=max(kelas))
for (i in 1:max(kelas)){
 kelas.id=which(kelas==i)
 centers[i,]=colMeans(x[kelas.id,])
}
betweenss=ss(centers[kelas,])
output=list(totss,betweenss)
names(output)=c('totss','betweenss')
return(output)
}

getmode <- function(v) {
   uniqv <- unique(na.omit(v))
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
final.cl=NULL
for (i in 1:N){
 final.cl[i]=getmode(t(res1[i,]))
}
write.csv(final.cl,'d:/result/3/gmm3-finalcl.csv')
```

Enclosure 5. Syntax of R for PCA and *K*-means

```
#PCA
pc=prcomp(dataclust)
summary(pc)
screeplot(pc,npcs=50,type='lines')
#pilih hingga PC 41, 80% variance
dataclust.pc=pc$x[,1:41]

library(cluster)
cl.kmeans.pc=NULL
tabel.pc=matrix(ncol=4,nrow=10)
for (i in 2:10){
 cl.kmeans.pc[[i]]=kmeans(dataclust.pc,i,iter.max=100)
 tabel.pc[i,1]=cl.kmeans.pc[[i]]$tot.withinss
 tabel.pc[i,2]=cl.kmeans.pc[[i]]$betweenss
 temp=silhouette(cl.kmeans.pc[[i]]$cluster,dist(dataclust.pc))
 tabel.pc[i,3]=mean(temp[,3])
 tabel.pc[i,4]=(tabel.pc[i,2]/(i-1))/(tabel.pc[i,1]/(n-i))
}
colnames(tabel.pc)=c('withinss','betweenss','silhouette','pseudoF')
```

Enclosure 6. Syntax of R for processing clustering result

```
img1=NULL
for (i in 1:length(file.precip)){
setwd('d:/Radar/colorcut')
 dev.off()
 img1[[i]]=readPNG(nama[i])
 grid.raster(img1[[i]][,,1])
 setwd('d:/result/3/finalcl')
 name=paste(final.cl[i],'-',file.precip[i],sep='')
 dev.print(png,filename=name,width=150,height=150)
}

library(magick)
img2=NULL
setwd('d:/result/2/finalcl')
daftar=list.files(pattern='.png',full.names=FALSE)
no.clust=substr(daftar,1,1)
i=1
img2=NULL
setwd('d:/result/2/finalcl')
no=which(no.clust==i)
img2=image_read(daftar[no])
img.avg=image_average(img2)
plot(as.raster(img.avg))
setwd('d:/result/2')
name=paste('cluster',i,'.png',sep='')
dev.copy(png,name)
dev.off()

#contour map
a=readPNG('d:/result/2/cluster1.png')
contour(a[,,1])
name=paste('contour2.png',sep='')
dev.copy(png,name)
dev.off()
```

Enclosure 7. Selected dates above the threshold

No	Date	Time	Precipitation	No	Date	Time	Precipitation
1	20171023	12:50	7.6	42	20171127	06:00	9.6
2	20171023	13:00	1.5	43	20171127	06:10	4.1
3	20171023	13:10	3.3	44	20171127	08:00	2.3
4	20171023	13:20	4.6	45	20171129	10:00	3
5	20171023	13:30	1.7	46	20171129	10:10	2.1
6	20171023	14:00	1.8	47	20171203	09:20	1.5
7	20171023	14:10	6.3	48	20171203	09:30	7.1
8	20171023	14:20	2	49	20171205	09:10	7.6
9	20171023	14:30	4.6	50	20171205	09:20	3.3
10	20171023	14:40	2.5	51	20171205	09:30	5.6
11	20171025	17:40	2	52	20171205	09:40	5.6
12	20171104	17:40	3.3	53	20171206	12:30	3.8
13	20171113	08:20	1.5	54	20171206	12:40	5.6
14	20171117	10:10	3.5	55	20171206	12:50	4
15	20171120	11:00	1.8	56	20171206	13:00	8.7
16	20171122	09:50	3.1	57	20171206	13:20	3.3
17	20171122	10:10	6.3	58	20171207	09:10	9.6
18	20171122	10:20	1.8	59	20171207	09:20	2.5
19	20171122	10:30	1.5	60	20171207	09:40	6.4
20	20171122	10:40	1.5	61	20171207	10:00	5.6
21	20171122	10:50	1.6	62	20171207	10:10	3
22	20171124	04:30	4.3	63	20171207	10:20	7.9
23	20171124	04:40	6.1	64	20171208	05:30	2.5
24	20171124	04:50	14.2	65	20171208	08:20	1.6
25	20171124	05:00	5.1	66	20171208	08:30	1.7
26	20171124	05:10	13.7	67	20171208	08:50	4.9
27	20171124	05:20	2.6	68	20171208	09:10	2.3
28	20171124	05:30	3.3	69	20171211	10:40	3
29	20171124	05:40	2.3	70	20171214	10:00	1.8
30	20171124	06:00	2.3	71	20171214	10:10	1.8
31	20171124	06:10	3.5	72	20171220	08:20	2.6
32	20171124	06:20	1.8	73	20171220	08:30	3
33	20171124	06:30	3.3	74	20171230	04:50	1.5
34	20171124	06:50	2	75	20171230	05:00	5.1
35	20171124	07:20	2.1	76	20180101	14:10	1.8
36	20171126	05:50	1.5	77	20180102	16:50	4.3
37	20171126	06:00	9.4	78	20180102	17:00	3.8
38	20171127	05:10	7.3	79	20180103	20:30	2
39	20171127	05:30	4.6	80	20180103	20:50	3.8
40	20171127	05:40	3.8	81	20180105	12:40	1.5
41	20171127	05:50	8.9	82	20180106	07:30	8.4

Enclosure 7. Selected dates above the threshold (cont'd)

No	Date	Time	Precipitation	No	Date	Time	Precipitation
83	20180106	07:40	8.9	124	20180209	08:20	1.7
84	20180106	07:50	8.1	125	20180209	08:50	10.4
85	20180106	08:00	4.1	126	20180209	09:00	1.8
86	20180106	08:10	6.8	127	20180209	09:10	3
87	20180108	14:00	5.1	128	20180210	13:40	6.1
88	20180108	15:50	4.8	129	20180210	14:00	3.8
89	20180110	13:40	16	130	20180210	14:10	8.1
90	20180111	10:20	5.3	131	20180210	14:20	5.6
91	20180111	10:30	1.8	132	20180213	06:20	2.3
92	20180111	10:40	1.5	133	20180213	07:20	3.8
93	20180111	11:50	4.6	134	20180213	07:30	1.5
94	20180118	13:30	2	135	20180215	06:30	4.3
95	20180118	13:40	8.4	136	20180215	09:30	1.5
96	20180118	13:50	6.9	137	20180215	09:40	9.9
97	20180118	15:10	2.5	138	20180215	09:50	7.1
98	20180119	08:20	5.9	139	20180215	10:00	2.5
99	20180119	08:30	11.2	140	20180215	10:10	2.1
100	20180119	08:40	6.1	141	20180215	10:20	2.5
101	20180119	08:50	2.8	142	20180215	10:30	2.6
102	20180119	09:00	1.5	143	20180215	10:40	1.5
103	20180119	11:10	1.5	144	20180215	11:30	1.5
104	20180120	08:00	3.8	145	20180216	14:50	6.1
105	20180120	08:10	4.1	146	20180220	13:10	5.6
106	20180120	08:20	2.5	147	20180220	13:20	2.5
107	20180120	15:30	5.3	148	20180220	13:30	2.3
108	20180120	15:40	13	149	20180225	17:20	2
109	20180122	22:30	1.5	150	20180225	17:30	4.9
110	20180122	22:40	1.5	151	20180225	17:40	4.5
111	20180124	12:40	2.8	152	20180225	18:10	11.2
112	20180124	13:20	4.1	153	20180226	04:00	2.8
113	20180125	05:50	6.1	154	20180226	04:10	1.7
114	20180131	09:10	2.3	155	20180226	04:20	2.6
115	20180201	03:50	2	156	20180226	04:30	5.8
116	20180201	04:00	5.6	157	20180226	04:40	4.1
117	20180201	04:20	8.1	158	20180227	13:00	2.3
118	20180201	04:30	3.6	159	20180304	04:10	6.9
119	20180202	12:30	1.5	160	20180304	04:20	4
120	20180207	10:00	5.1	161	20180307	10:00	5.3
121	20180208	09:00	3.8	162	20180307	10:10	2.3
122	20180208	09:30	2.1	163	20180307	10:20	1.8
123	20180208	09:40	1.5	164	20180307	12:20	1.8

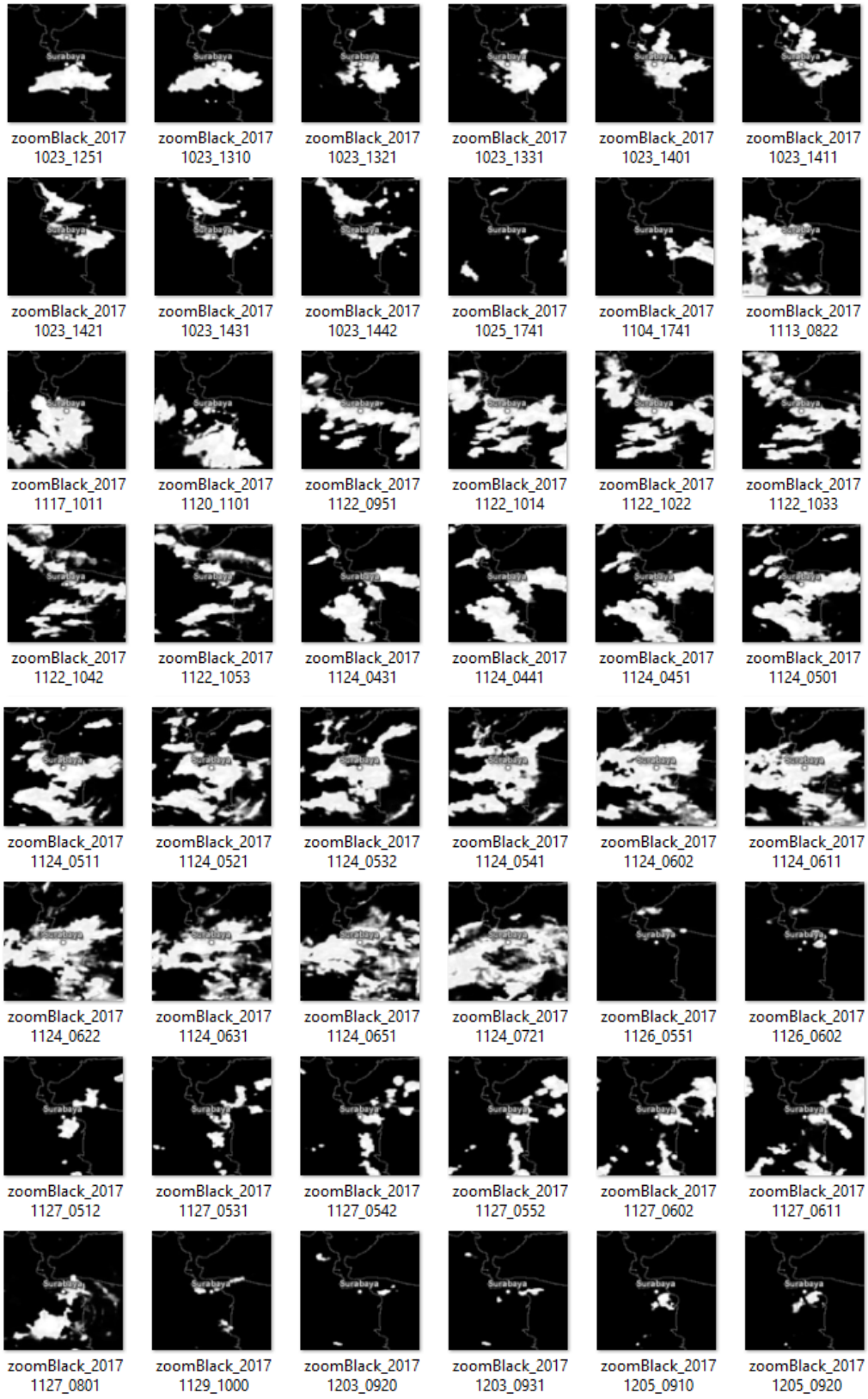
Enclosure 7. Selected dates above the threshold (cont'd)

No	Date	Time	Precipitation	No	Date	Time	Precipitation
165	20180308	10:10	2.8	180	20180324	15:00	2.8
166	20180308	10:20	3.8	181	20180329	12:10	2
167	20180308	10:30	6.6	182	20180329	12:20	3.3
168	20180308	10:40	12.4	183	20180329	13:50	1.5
169	20180308	10:50	8.9	184	20180330	04:40	5.1
170	20180308	11:00	4.3	185	20180330	04:50	2
171	20180308	11:20	3.3	186	20180401	11:30	2
172	20180308	11:30	1.8	187	20180401	13:00	1.6
173	20180309	11:40	1.5	188	20180401	13:10	2.7
174	20180313	08:40	2	189	20180402	06:20	5.6
175	20180316	13:20	1.6	190	20180402	06:30	4.8
176	20180317	08:30	3.3	191	20180402	06:40	2
177	20180317	08:40	6.9	192	20180402	09:10	9.2
178	20180320	13:10	3.3	193	20180402	09:20	5.8
179	20180324	14:50	1.5				

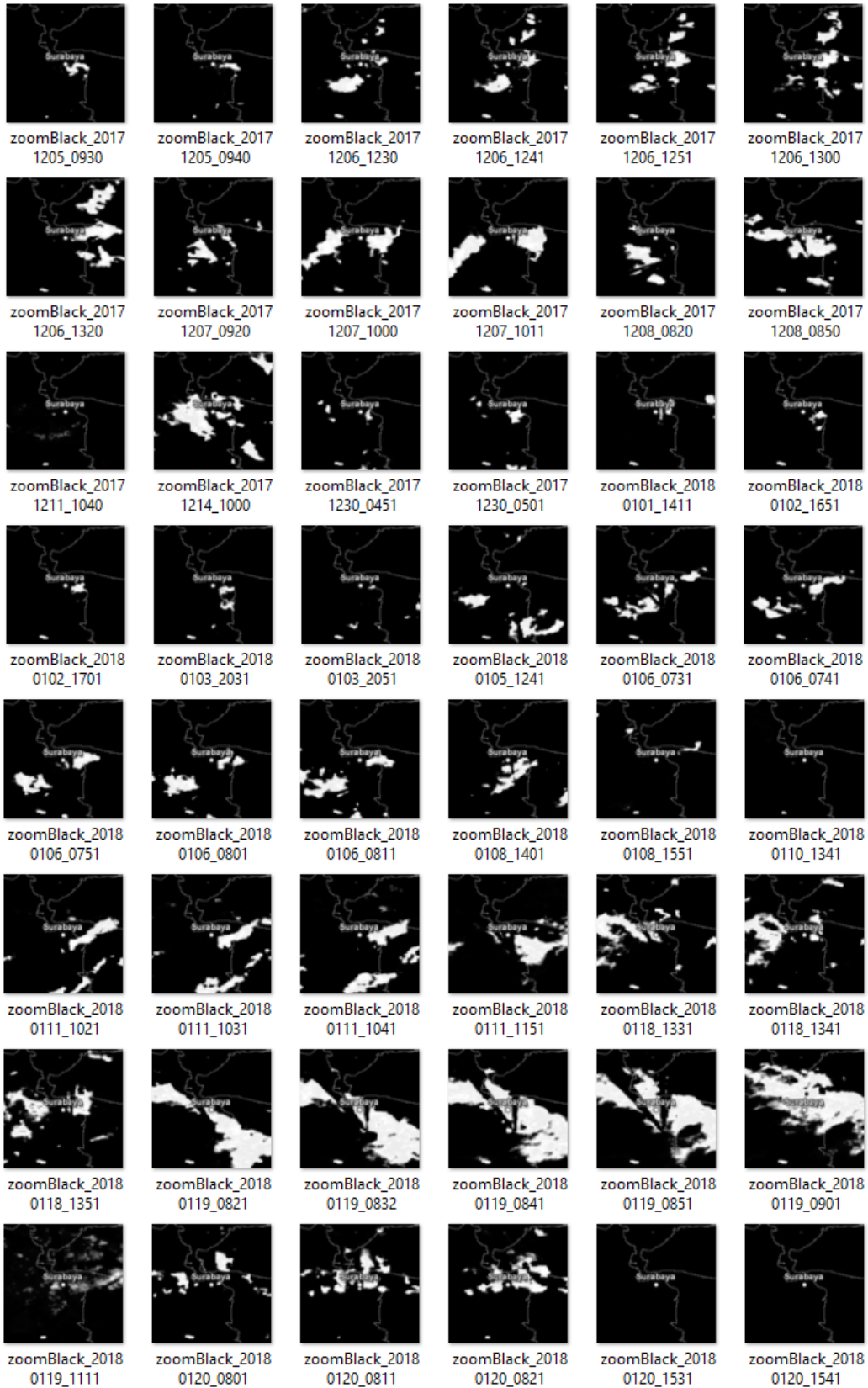
Enclosure 8. Output of gpd function in R (from 'evir' package)

```
$n
[1] 161

$data
  [1] 16.0  5.3 13.0  4.1  6.1  5.1  3.8  9.9  5.6  2.8  1.7  2.6  5.8  6.9  4.0
 [16]  2.0  2.8  2.0  3.3  5.1  2.0  3.5  9.4  3.0  7.1  7.6  3.3  5.6  5.6  5.6
 [31]  4.0  8.7  2.5  3.0  1.6  3.0  7.6  1.8  6.3  2.0  4.6  2.5  5.1  1.8  4.3
 [46]  3.8  2.0  3.8  8.9  8.1  4.1  2.0  5.1  4.8  5.3  1.8  4.6  2.0  8.4  6.9
 [61]  5.9 11.2  6.1  2.8  3.8  4.1  2.5  2.8  2.3  2.0  8.1  3.8  2.1 10.4  1.8
 [76]  3.0  6.1  3.8  8.1  2.3  4.3  7.1  6.1  2.5  2.3  2.0  4.9  4.5 11.2  4.1
 [91]  2.3  3.3  3.3  3.3  3.5  1.8  6.3  1.8  1.6  4.3  6.1 14.2  5.1 13.7  2.3
[106]  2.3  3.3  7.3  4.6  3.8  8.9  3.8  3.3  5.6  4.9  3.3  4.6  1.7  8.4  6.8
[121]  5.6  3.6  1.7  5.6  2.5  2.1  2.5  2.6  1.6  6.9  3.1  2.6  3.3  1.8  2.0
[136]  2.1  9.6  4.1  2.3  1.8

$threshold
[1] 1.5

$p.less.thresh
[1] 0.1304348

$n.exceed
[1] 140

$method
[1] "ml"

$par.ests
        xi       beta
-0.1145074  3.4711514

$par.ses
        xi       beta
0.07947918 0.40131937

$varcov
            [,1]        [,2]
[1,]  0.00631694 -0.02444071
[2,] -0.02444071  0.16105723

$information
[1] "observed"

$converged
[1] 0

$nllh.final
[1] 298.186

attr(,"class")
[1] "gpd"
```

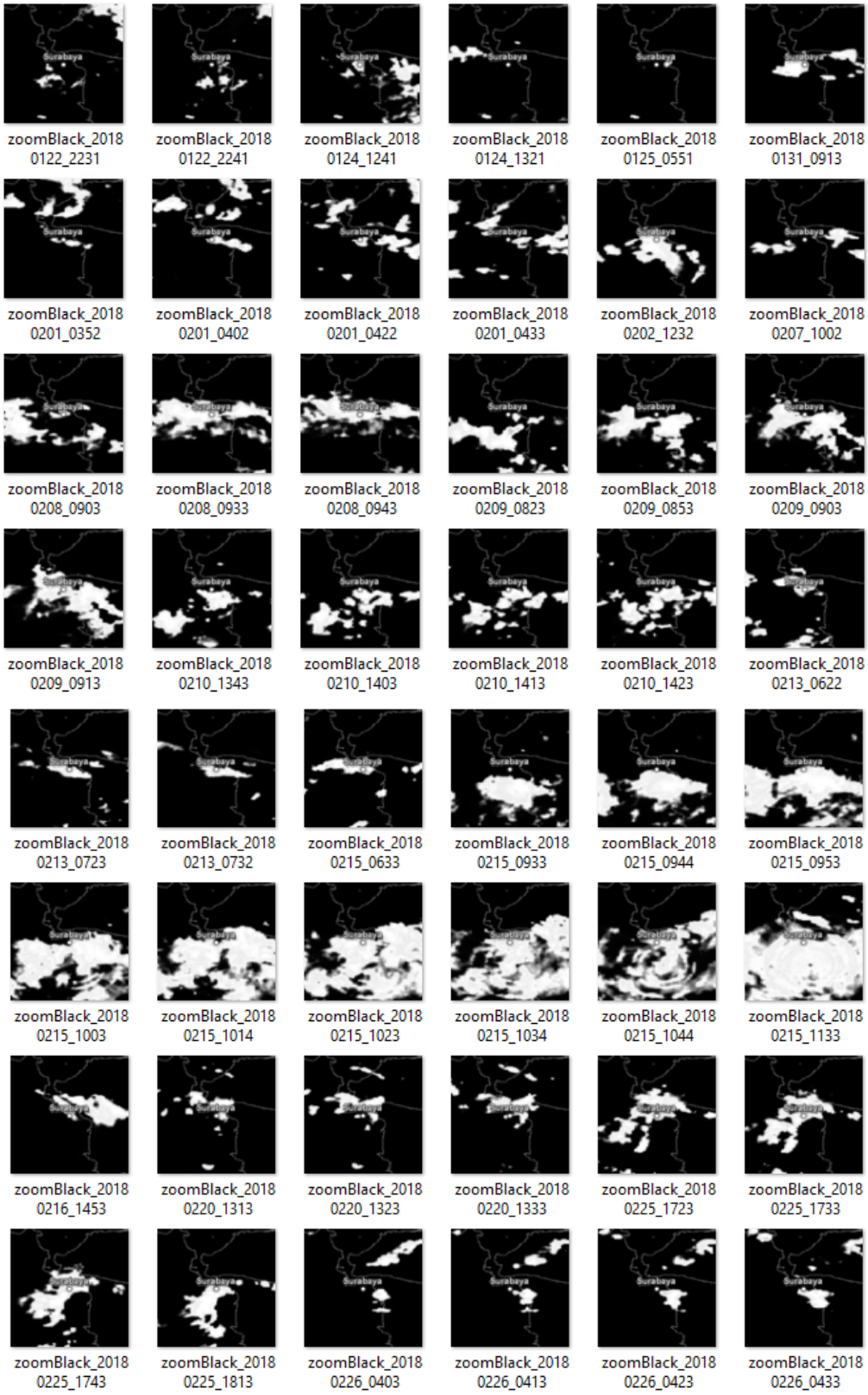
Enclosure 9. Selected images for cluster analysis



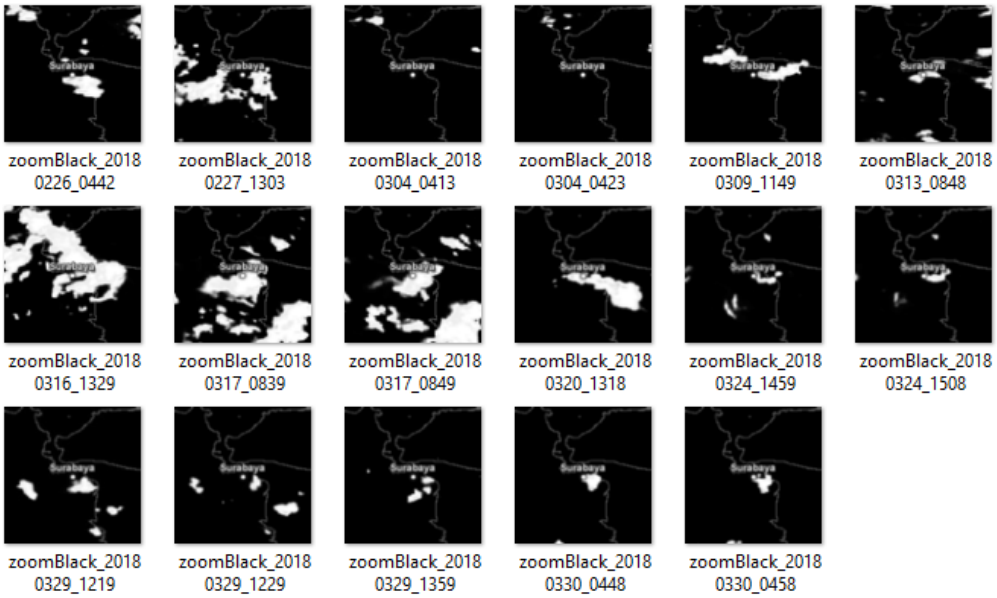
Enclosure 9. Selected images for cluster analysis (cont'd)



Enclosure 9. Selected images for cluster analysis (cont'd)



Enclosure 9. Selected images for cluster analysis (cont'd)



Enclosure 10. Result of HDDC for $K = 2$ to 10

```
2 Cluster
HIGH DIMENSIONAL DATA CLUSTERING
MODEL: AKJBKQKDK
  Posterior probabilities of groups
       1     2
   0.453 0.547
      Intrinsic dimensions of the classes:
       1 2
  dim: 2 3

Class       a1       a2      a3
    1 13434150  6231530       .
    2 23799683 10395782 9794068
       1    2
Bk: 1443 4385
BIC:  -39636666

3 Cluster
HIGH DIMENSIONAL DATA CLUSTERING
MODEL: AKJBKQKDK
  Posterior probabilities of groups
       1     2     3
   0.199 0.404 0.398
      Intrinsic dimensions of the classes:
       1 2 3
  dim: 1 3 3

Class       a1       a2      a3
    1 22609884        .       .
    2 21613930 13689825 9874091
    3 19223280  8414224 7061814
      1    2    3
Bk: 528 4428 2466
BIC:  -39466383

4 Cluster
HIGH DIMENSIONAL DATA CLUSTERING
MODEL: AKJBKQKDK
  Posterior probabilities of groups
        1     2     3     4
   0.0683 0.273 0.335 0.323
      Intrinsic dimensions of the classes:
       1 2 3 4
  dim: 1 4 4 6

Class       a1       a2       a3      a4      a5      a6
    1 84370750        .        .       .       .       .
    2 23440149 11558599  7854835 6744402       .       .
    3 14305450  9640678  5580017 5260105       .       .
    4 16856223 12387891 10322929 7202967 6681015 5901534
      1   2    3    4
Bk: 3033 793 2463 3064
BIC:  -39932456

5 Cluster
HIGH DIMENSIONAL DATA CLUSTERING
MODEL: AKJBKQKDK
  Posterior probabilities of groups
       1     2     3    4      5
   0.292 0.205 0.143 0.28 0.0807
      Intrinsic dimensions of the classes:
       1 2 3 4  5
  dim: 3 2 2 3 12

Class       a1       a2       a3       a4      a5      a6      a7      a8
    1 30462564 13082696 11176745        .       .       .       .       .
    2 17413446  6197114        .        .       .       .       .       .
    3 38775435 15658996        .        .       .       .       .       .
    4 20578707  9088262  7280975        .       .       .       .       .
```

Enclosure 10. Result of HDDC for $K = 2$ to 10 (cont'd)

```
  5 25373148 24266273 13315993 11456497 9687602 6614858 5064277 4781268

Class      a9     a10     a11     a12
    1       .       .       .       .
    2       .       .       .       .
    3       .       .       .       .
    4       .       .       .       .
    5 3924876 3201302 2611896 2463049
       1   2    3    4     5
Bk: 3948 521 2466 2362 1e-08
BIC:  -34350589
Information: b < 10e-6

6 Cluster
HIGH DIMENSIONAL DATA CLUSTERING
MODEL: AKJBKQKDK
  Posterior probabilities of groups
      1     2     3     4      5      6
   0.13 0.416 0.267 0.106 0.0559 0.0248
      Intrinsic dimensions of the classes:
       1 2 3 4 5 6
  dim: 1 6 3 4 8 2

Class       a1       a2       a3       a4      a5      a6      a7      a8
    1 50750690        .        .        .       .       .       .       .
    2  8302861  5866061  5042080  4015506 3121975 2701138       .       .
    3 22839573 12415134 10301784        .       .       .       .       .
    4 26093769 24725207 12922298 10411906       .       .       .       .
    5 36345689 19737285 12845920 11279218 9975880 8028252 7690498 5110341
    6 49913106 42972163        .        .       .       .       .       .
       1    2    3    4     5    6
Bk: 4352 1065 3418 1987 1e-08 1480
BIC:  -36619357
Information: b < 10e-6

7 Cluster
HIGH DIMENSIONAL DATA CLUSTERING
MODEL: AKJBKQKDK
  Posterior probabilities of groups
       1      2     3     4      5    6     7
   0.112 0.0435 0.242 0.199 0.0745 0.13 0.199
      Intrinsic dimensions of the classes:
        1 2 3 4  5 6 7
  dim: 17 3 3 4 11 2 3

Class       a1       a2       a3       a4       a5       a6      a7      a8
    1 16809441 12559051 11680382  8260305  7812441  7209705 5551318 5187663
    2 62938643 30668812 15794837        .        .        .       .       .
    3 17925954 11773871  8774191        .        .        .       .       .
    4 21609450 14980101 12817927  8275121        .        .       .       .
    5 36651954 24018474 22732333 16125981 12100587 11168832 9682366 7789437
    6 39412201 18048630        .        .        .        .       .       .
    7 13448000  5960719  4009901        .        .        .       .       .

Class      a9     a10     a11     a12     a13     a14     a15     a16     a17
    1 4218283 4060392 3540960 3413633 3038300 2661918 2087905 1794286 1766535
    2       .       .       .       .       .       .       .       .       .
    3       .       .       .       .       .       .       .       .       .
    4       .       .       .       .       .       .       .       .       .
    5 6697302 6111558 5692719       .       .       .       .       .       .
    6       .       .       .       .       .       .       .       .       .
    7       .       .       .       .       .       .       .       .       .
       1   2    3    4     5    6   7
Bk: 1e-08 766 2000 3015 1e-08 3116 388
BIC:  -27589760
Information: b < 10e-6
```

Enclosure 10. Result of HDDC for $K = 2$ to 10 (cont'd)

```
8 Cluster
HIGH DIMENSIONAL DATA CLUSTERING
MODEL: AKJBKQKDK
  Posterior probabilities of groups
       1      2      3      4      5      6      7      8
   0.224 0.0373 0.224 0.0932 0.0932 0.211 0.0807 0.0373
      Intrinsic dimensions of the classes:
       1 2 3  4  5 6 7 8
  dim: 5 2 5 14 14 5 4 2

Class       a1       a2       a3       a4       a5      a6      a7      a8
    1 15686151 11772520 10979803  6456708  5299372       .       .       .
    2 60299723 25870419        .        .        .       .       .       .
    3  9722796  5714025  4664890  3806705  2589490       .       .       .
    4 30155196 16967517 14912304 13267192 11126319 9737353 8982053 8549735
    5 29568068 20948013 14307363 12811022 11110191 9068779 7520557 5628422
    6 24158347 13940083  8905481  8454097  6548175       .       .       .
    7 33157854 21260016 14230881 11474569        .       .       .       .
    8 80814360 22290591        .        .        .       .       .       .

Class      a9     a10     a11     a12     a13     a14
    1       .       .       .       .       .       .
    2       .       .       .       .       .       .
    3       .       .       .       .       .       .
    4 7420185 6323352 5478425 4873873 3923585 3243008
    5 5164454 4454518 4197295 3506745 3305305 2791641
    6       .       .       .       .       .       .
    7       .       .       .       .       .       .
    8       .       .       .       .       .       .
       1    2   3     4     5    6    7   8
Bk: 2150 1151 401 1e-08 1e-08 2092 1743 596
BIC:  -27960586
Information: b < 10e-6

9 Cluster
HIGH DIMENSIONAL DATA CLUSTERING
MODEL: AKJBKQKDK
  Posterior probabilities of groups
        1     2     3     4     5      6     7      8      9
   0.0621 0.087 0.193 0.149 0.149 0.0807 0.155 0.0683 0.0559
      Intrinsic dimensions of the classes:
       1 2 3 4 5  6 7 8 9
  dim: 9 1 4 6 5 12 4 3 8

Class       a1       a2       a3       a4       a5       a6      a7      a8
    1 20163560 16576022 14181839 11288327  9438233  8053944 7118962 6183126
    2 63961007        .        .        .        .        .       .       .
    3 14360058  7292060  5327030  4659397        .        .       .       .
    4 18199343 14616402  8954940  6530158  5502916  4436560       .       .
    5 21841258 16063134 11184080  9816494  7654057        .       .       .
    6 24229371 18443839 12243058  9910733  7368539  6355837 5809734 5213549
    7 22058770 15405906 11123362  8635775        .        .       .       .
    8 38419394 19020257 15258158        .        .        .       .       .
    9 36989665 20439508 18914324 16065675 11975970 10376616 9540917 8544020

Class      a9     a10     a11     a12
    1 4486179       .       .       .
    2       .       .       .       .
    3       .       .       .       .
    4       .       .       .       .
    5       .       .       .       .
    6 4346815 3471025 2790030 1444817
    7       .       .       .       .
    8       .       .       .       .
    9       .       .       .       .
       1    2   3    4    5     6    7    8     9
Bk: 1e-08 4028 355 1078 2273 1e-08 2365 1290 1e-08
BIC:  -27398505
Information: b < 10e-6
```

Enclosure 10. Result of HDDC for $K = 2$ to 10 (cont'd)

```
10 Cluster
HIGH DIMENSIONAL DATA CLUSTERING
MODEL: AKJBKQKDK
  Posterior probabilities of groups
        1      2     3      4      5     6     7      8     9    10
   0.0994 0.0559 0.087 0.0745 0.0683 0.199 0.087 0.0435 0.124 0.161
      Intrinsic dimensions of the classes:
       1 2  3 4 5 6  7 8 9 10
  dim: 1 8 13 2 6 2 13 6 2  3

Class       a1       a2       a3       a4       a5       a6      a7      a8
   1  40869266        .        .        .        .        .       .       .
   2  14910322 12504079  9773614  7698456  5921972  4505648 4253933 3567383
   3  27967427 20747315 15374835 11604701  9834446  9238314 7190629 6759226
   4  43659226 16910200        .        .        .        .       .       .
   5  34270706 24351845 20395215 15009952 11880413 10287985       .       .
   6   9289980  4667513        .        .        .        .       .       .
   7  23321935 18742729 12936128  9440552  8514701  7382178 6042093 5656567
   8  35799583 34437851 25495006 18272976 16048319 13590211       .       .
   9  33376742 14925246        .        .        .        .       .       .
   10 19600942 10053818  9805429        .        .        .       .       .

Class      a9     a10     a11     a12     a13
   1        .       .       .       .       .
   2        .       .       .       .       .
   3  6164259 5952322 4985708 3876302 3339843
   4        .       .       .       .       .
   5        .       .       .       .       .
   6        .       .       .       .       .
   7  5162021 3934019 3778506 3185260 2784079
   8        .       .       .       .       .
   9        .       .       .       .       .
   10       .       .       .       .       .
       1     2     3    4   5   6     7     8    9   10
Bk: 4360 1e-08 1e-08 1579 942 418 1e-08 1e-08 3161 2006
BIC:  -22296092
Information: b < 10e-6
```

Enclosure 11. Result of PCA analysis in R

```
Importance of components:
                             PC1       PC2       PC3       PC4       PC5
Standard deviation     4781.3952 2.511e+03 2.445e+03 2.279e+03 2.036e+03
Proportion of Variance    0.2092 5.771e-02 5.468e-02 4.751e-02 3.791e-02
Cumulative Proportion     0.2092 2.669e-01 3.216e-01 3.691e-01 4.070e-01
                              PC6       PC7       PC8       PC9      PC10
Standard deviation     1.836e+03 1.748e+03 1.609e+03 1.571e+03 1.507e+03
Proportion of Variance 3.084e-02 2.794e-02 2.369e-02 2.258e-02 2.077e-02
Cumulative Proportion  4.378e-01 4.658e-01 4.895e-01 5.120e-01 5.328e-01
                             PC11      PC12      PC13      PC14      PC15
Standard deviation     1398.8390 1.332e+03 1276.2975 1.262e+03 1.230e+03
Proportion of Variance    0.0179 1.623e-02    0.0149 1.458e-02 1.385e-02
Cumulative Proportion     0.5507 5.669e-01    0.5818 5.964e-01 6.103e-01
                             PC16      PC17      PC18      PC19      PC20
Standard deviation     1.132e+03 1.102e+03 1.072e+03 1029.6839 1.016e+03
Proportion of Variance 1.172e-02 1.112e-02 1.051e-02    0.0097 9.450e-03
Cumulative Proportion  6.220e-01 6.331e-01 6.436e-01    0.6533 6.628e-01
                             PC21      PC22      PC23      PC24      PC25
Standard deviation     1.002e+03 979.20228 948.59215 943.31120 923.4875
Proportion of Variance 9.190e-03   0.00877   0.00823   0.00814   0.0078
Cumulative Proportion  6.720e-01   0.68073   0.68896   0.69710   0.7049
                             PC26      PC27      PC28      PC29      PC30
Standard deviation     901.20721 892.06351 879.96002 875.53115 856.35847
Proportion of Variance   0.00743   0.00728   0.00708   0.00701   0.00671
Cumulative Proportion    0.71234   0.71962   0.72670   0.73372   0.74043
                             PC31      PC32      PC33      PC34      PC35
Standard deviation     838.92241 832.29866 816.14414 806.38642 784.94301
Proportion of Variance   0.00644   0.00634   0.00609   0.00595   0.00564
Cumulative Proportion    0.74687   0.75321   0.75930   0.76525   0.77089
                             PC36      PC37      PC38      PC39      PC40
Standard deviation     775.93119 759.54691 751.50161 743.61360 730.60356
Proportion of Variance   0.00551   0.00528   0.00517   0.00506   0.00488
Cumulative Proportion    0.77640   0.78168   0.78684   0.79190   0.79679
                             PC41      PC42      PC43      PC44      PC45
Standard deviation     715.68903 713.76373 704.33743 691.32892 684.73574
Proportion of Variance   0.00469   0.00466   0.00454   0.00437   0.00429
Cumulative Proportion    0.80147   0.80613   0.81067   0.81505   0.81934
                             PC46      PC47      PC48      PC49      PC50
Standard deviation     673.85239 664.51147 652.02394 648.74430 640.88782
Proportion of Variance   0.00415   0.00404   0.00389   0.00385   0.00376
Cumulative Proportion    0.82349   0.82753   0.83142   0.83527   0.83903
                             PC51      PC52      PC53      PC54      PC55
Standard deviation     633.57988 629.58827 622.41089 615.92971 608.95971
Proportion of Variance   0.00367   0.00363   0.00354   0.00347   0.00339
Cumulative Proportion    0.84270   0.84633   0.84988   0.85335   0.85674
                             PC56      PC57      PC58      PC59      PC60
Standard deviation     606.48874 599.09740 592.24642 586.06861 579.22140
Proportion of Variance   0.00337   0.00328   0.00321   0.00314   0.00307
Cumulative Proportion    0.86011   0.86339   0.86660   0.86974   0.87281
                             PC61      PC62      PC63      PC64      PC65
Standard deviation     570.95682 570.37051 561.14527 554.75708 550.18244
Proportion of Variance   0.00298   0.00298   0.00288   0.00282   0.00277
Cumulative Proportion    0.87579   0.87877   0.88165   0.88447   0.88724
                             PC66      PC67      PC68      PC69      PC70
Standard deviation     547.70924 539.53777 535.26703 534.72827 529.56877
Proportion of Variance   0.00274   0.00266   0.00262   0.00262   0.00257
Cumulative Proportion    0.88998   0.89265   0.89527   0.89788   0.90045
```

Enclosure 11. Result of PCA analysis in R (cont'd)

```
                           PC71      PC72      PC73      PC74      PC75
Standard deviation     516.47732 513.86873 508.11815 506.78368 504.37252
Proportion of Variance   0.00244   0.00242   0.00236   0.00235   0.00233
Cumulative Proportion    0.90289   0.90531   0.90767   0.91002   0.91235
                           PC76      PC77      PC78      PC79      PC80
Standard deviation     499.39413 494.03129 491.09791 488.12404 484.03791
Proportion of Variance   0.00228   0.00223   0.00221   0.00218   0.00214
Cumulative Proportion    0.91463   0.91686   0.91907   0.92125   0.92339
                           PC81      PC82      PC83      PC84      PC85
Standard deviation     473.93877 470.19661 466.52393 464.94782 460.76382
Proportion of Variance   0.00206   0.00202   0.00199   0.00198   0.00194
Cumulative Proportion    0.92545   0.92747   0.92946   0.93144   0.93338
                           PC86      PC87      PC88      PC89      PC90
Standard deviation     451.41851 450.03758 447.65132 445.27667 440.50776
Proportion of Variance   0.00186   0.00185   0.00183   0.00181   0.00178
Cumulative Proportion    0.93525   0.93710   0.93893   0.94075   0.94252
                           PC91      PC92      PC93      PC94      PC95
Standard deviation     437.35704 431.77461 430.4734 426.23640 420.39118
Proportion of Variance   0.00175   0.00171   0.0017   0.00166   0.00162
Cumulative Proportion    0.94427   0.94598   0.9477   0.94934   0.95095
                           PC96      PC97      PC98      PC99     PC100
Standard deviation     418.6572 410.95903 410.46789 402.45315 397.44239
Proportion of Variance   0.0016   0.00155   0.00154   0.00148   0.00145
Cumulative Proportion    0.9526   0.95410   0.95564   0.95713   0.95857
                          PC101     PC102     PC103     PC104     PC105
Standard deviation     394.71594 391.91962 389.99446 385.41945 380.32126
Proportion of Variance   0.00143   0.00141   0.00139   0.00136   0.00132
Cumulative Proportion    0.96000   0.96140   0.96279   0.96415   0.96548
                          PC106     PC107     PC108     PC109     PC110
Standard deviation     377.2296 373.99233 370.45200 365.36008 359.23297
Proportion of Variance   0.0013   0.00128   0.00126   0.00122   0.00118
Cumulative Proportion    0.9668   0.96806   0.96931   0.97054   0.97172
                          PC111     PC112     PC113    PC114     PC115
Standard deviation     355.01133 352.50170 348.38936 346.1138 339.08563
Proportion of Variance   0.00115   0.00114   0.00111   0.0011   0.00105
Cumulative Proportion    0.97287   0.97401   0.97512   0.9762   0.97727
                          PC116     PC117     PC118     PC119     PC120
Standard deviation     337.39665 331.89730 329.19305 327.03359 320.81330
Proportion of Variance   0.00104   0.00101   0.00099   0.00098   0.00094
Cumulative Proportion    0.97831   0.97931   0.98031   0.98128   0.98223
                          PC121     PC122     PC123     PC124     PC125
Standard deviation     317.41606 314.62494 306.55071 302.54390 299.39469
Proportion of Variance   0.00092   0.00091   0.00086   0.00084   0.00082
Cumulative Proportion    0.98315   0.98405   0.98491   0.98575   0.98657
                          PC126     PC127     PC128     PC129     PC130
Standard deviation     296.2917 294.47230 285.55440 283.04105 273.54878
Proportion of Variance   0.0008   0.00079   0.00075   0.00073   0.00068
Cumulative Proportion    0.9874   0.98817   0.98891   0.98965   0.99033
                          PC131     PC132     PC133     PC134     PC135
Standard deviation     269.46041 263.12607 260.19898 246.07109 238.67062
Proportion of Variance   0.00066   0.00063   0.00062   0.00055   0.00052
Cumulative Proportion    0.99100   0.99163   0.99225   0.99280   0.99332
                          PC136     PC137     PC138     PC139     PC140
Standard deviation     237.11214 230.68151 225.24613 217.90825 215.85690
Proportion of Variance   0.00051   0.00049   0.00046   0.00043   0.00043
Cumulative Proportion    0.99384   0.99433   0.99479   0.99522   0.99565
                          PC141     PC142     PC143     PC144     PC145
Standard deviation     212.36885 209.3089 206.42377 193.69258 189.68894
Proportion of Variance   0.00041   0.0004   0.00039   0.00034   0.00033
Cumulative Proportion    0.99606   0.9965   0.99685   0.99720   0.99753
```

Enclosure 11. Result of PCA analysis in R (cont'd)

```
                          PC146     PC147     PC148     PC149     PC150
Standard deviation     182.2520 179.05878 173.92591 169.34523 163.99489
Proportion of Variance   0.0003   0.00029   0.00028   0.00026   0.00025
Cumulative Proportion    0.9978   0.99812   0.99840   0.99866   0.99891
                          PC151     PC152     PC153     PC154     PC155
Standard deviation     155.63490 146.4362 135.85436 129.75332 112.40443
Proportion of Variance   0.00022   0.0002   0.00017   0.00015   0.00012
Cumulative Proportion    0.99913   0.9993   0.99950   0.99965   0.99977
                          PC156    PC157    PC158    PC159  PC160     PC161
Standard deviation     101.9180 84.25538 78.44677 44.37400 0.2908 1.081e-11
Proportion of Variance   0.0001  0.00006  0.00006  0.00002 0.0000 0.000e+00
Cumulative Proportion    0.9999  0.99993  0.99998  1.00000 1.0000 1.000e+00
```

(this page is intentionally left blank)

BIOGRAPHY

Kiki Ferawati, born in Gresik, June 21st 1993. Kiki completed her bachelor degree at 2015 in ITS. In 2016, she enrolled in master program of Statistics in ITS. In her third semester, she participated as an exchange student in Department of Earth and Environmental Science in Kumamoto University for 6 months, starting from October 2017 to March 2018 to study more about meteorological aspects in order to get a better understanding of her thesis project. Kiki has been focused and involved on climate-related project since her bachelor degree. Kiki is also interested in computational statistics, statistical programming, and field related to data mining.

kiki.ferawati@gmail.com