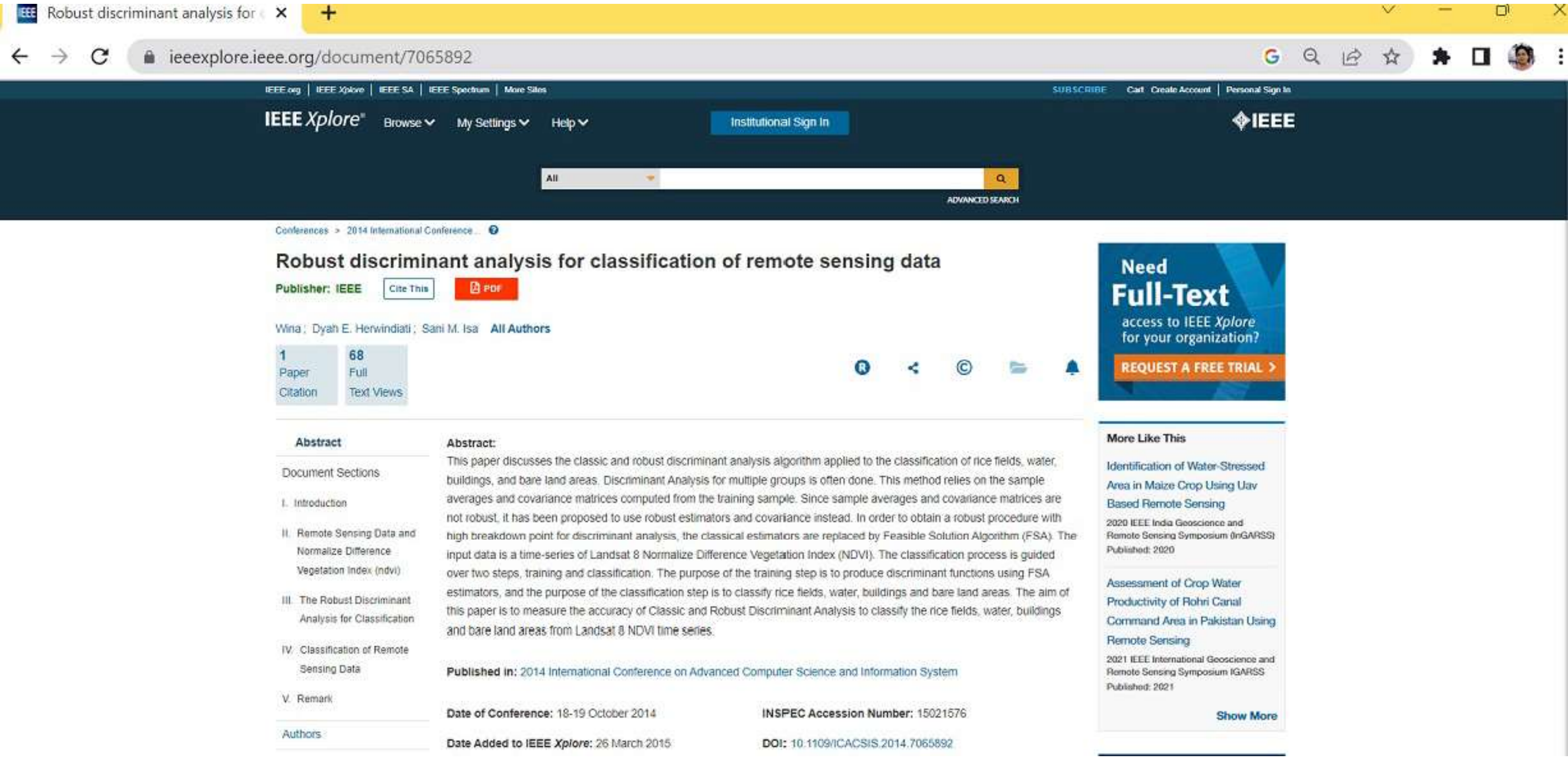
Robust discriminant analysis for
ieeexplore.ieee.org/document/7065892
IEEE.org | IEEE Xplore | IEEE SA | IEEE Spectrum | More Sites
SUBSCRIBE
Cart Create Account | Personal Sign In
IEEE Xplore®
Browse
My Settings
Help
Institutional Sign In
IEEE
All
ADVANCED SEARCH
Conferences > 2014 International Conference...

Robust discriminant analysis for classification of remote sensing data

Publisher: IEEE

Wina ; Dyah E. Herwindiati ; Sani M. Isa All Authors

1 Paper Citation

68 Full Text Views

Abstract
Document Sections
I. Introduction
II. Remote Sensing Data and Normalize Difference Vegetation Index (ndvi)
III. The Robust Discriminant Analysis for Classification
IV. Classification of Remote Sensing Data
V. Remark
Authors

Abstract:
This paper discusses the classic and robust discriminant analysis algorithm applied to the classification of rice fields, water, buildings, and bare land areas. Discriminant Analysis for multiple groups is often done. This method relies on the sample averages and covariance matrices computed from the training sample. Since sample averages and covariance matrices are not robust, it has been proposed to use robust estimators and covariance instead. In order to obtain a robust procedure with high breakdown point for discriminant analysis, the classical estimators are replaced by Feasible Solution Algorithm (FSA). The input data is a time-series of Landsat 8 Normalize Difference Vegetation Index (NDVI). The classification process is guided over two steps, training and classification. The purpose of the training step is to produce discriminant functions using FSA estimators, and the purpose of the classification step is to classify rice fields, water, buildings and bare land areas. The aim of this paper is to measure the accuracy of Classic and Robust Discriminant Analysis to classify the rice fields, water, buildings and bare land areas from Landsat 8 NDVI time series.

Published in: 2014 International Conference on Advanced Computer Science and Information System

Date of Conference: 18-19 October 2014

INSPEC Accession Number: 15021576

Date Added to IEEE *Xplore*: 26 March 2015

DOI: 10.1109/ICACSIS.2014.7065892

More Like This

Identification of Water-Stressed Area in Maize Crop Using Uav Based Remote Sensing
2020 IEEE India Geoscience and Remote Sensing Symposium (InGARSS)
Published: 2020

Assessment of Crop Water Productivity of Rohri Canal Command Area in Pakistan Using Remote Sensing
2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS
Published: 2021

Robust Discriminant Analysis for Classification of Remote Sensing Data

Wina, Dyah E. Herwindiati, and Sani M. Isa
Faculty of Computer Science, Tarumanagara University
Email: yap_cho_huan@yahoo.co.id, herwindiati@untar.ac.id, sani.fti.untar@gmail.com

Abstract--**This paper discusses the classic and robust discriminant analysis algorithm applied to the classification of rice fields, water, buildings, and bare land areas. Discriminant Analysis for multiple groups is often done. This method relies on the sample averages and covariance matrices computed from the training sample. Since sample averages and covariance matrices are not robust, it has been proposed to use robust estimators and covariance instead. In order to obtain a robust procedure with high breakdown point for discriminant analysis, the classical estimators are replaced by Feasible Solution Algorithm (FSA). The input data is a time-series of Landsat 8 Normalize Difference Vegetation Index (NDVI). The classification process is guided over two steps, training and classification. The purpose of the training step is to produce discriminant functions using FSA estimators, and the purpose of the classification step is to classify rice fields, water, buildings and bare land areas. The aim of this paper is to measure the accuracy of Classic and Robust Discriminant Analysis to classify the rice fields, water, buildings and bare land areas from Landsat 8 NDVI time series.**

***Index Terms*--discriminant analysis, feasible solution algorithm (FSA), high breakdown point, normalize difference vegetation index (NDVI), outlier, robust discriminant analysis.**

I. Introduction

Discriminant analysis was introduced by Fisher (1938) as a statistical method for separating two groups of populations. Rao (1948)extended this multivariate technique to multiple populations [1]. The main purpose of discriminant analysis is to predict group membership from a set of predictors using discriminant functions. But new observations might be classified incorrectly if the data which being analyzed consist of outliers.

Outlier is a case with such an extreme value on one variable (aunivariate outlier) or such astrange of combination of scores on two or more variable that it distorts statistics [2].In order to make discriminant analysis work optimally within the classification though in the condition of data which contains of outlier, robust estimator is needed. The great majority of robust estimator is concentrated on replacing mean vectors and covariance matrices in discriminant analysis by robust counterparts.

Many robust estimators of multivariate and scatter have been proposed in the literature, including Croux and Dehon with S-estimators [3], Basak with M-estimators [4], Chork and Rousseeuw with MVE estimators [5],Herwindiati, Djauhari and Mashuri with MVV estimators [6], Hubert and Debruynewith MCDestimators [7] and Adrian with FSA and FMCD estimators [15]. They proposed the robust measure based on the same criteria, that is, to detect the outlier in multivariate data sets.

In order to obtain a robust procedure with high breakdown point for discriminant analysis, the classical estimators are replaced by Feasible Solution Algorithm (FSA). The FSA was introduced by Hawkins [8] to obtain the MCD estimators for a given data set. It provides a high breakdown multivariate estimator for use. The algorithm is probabilistic; it involves taking random starting ‘trial solutions’ and refining each to a local optimum satisfying the necessary condition for the MCD optimum.

This paper discusses robust estimators for discriminant analysis classification, in this case for the classification of rice fields, water, buildings and bare land areas using remote sensing. The classification process is performed in two steps, training step and classification step. The training step is used to produce discriminant functions using FSA estimators, and the classification step is used to classify rice fields, water, buildings and bare land areas.

A time-series of Landsat 8 NDVI data was prepared. Landsat 8 is the latest NASA satellite that began in 1972, providing global, synoptic, and repetitive coverage of the Earth’s land surfaces, continues at a scale where natural and human-induced changes can be detected, differentiated, characterized, and monitored over time [9]. The Landsat Data Continuity Mission (named Landsat 8 after on-orbit initialization and verification) launched from Vandenberg Water Force Base in California on February 11, 2013, atop an Atlas V rocket. As with previous partnerships, this collaboration between the U.S. Geological Survey (USGS) and National Aeronautics and Space Administration (NASA) continues the mission to acquire high-quality data that meet both USGS and NASA scientific and operationalrequirements for observing land use and land cover change [10].

The NDVI layers are composited from data acquired over a 10-day period, where the 1st, the 11th, and the 21st of each month define the time limit for each compositing period [11]. We select a reference NDVI cycle that represents the growth pattern of rice fields. The selection of a reference NDVI cycle requires ground truthing but may also be based on information derived from satellite images as will be explained below. Reference selection is done by longitude and latitude adjustment with Google earth.

The aim of this paper is to measure the accuracy of Classic and Robust Discriminant Analysis to classify the rice fields, water, buildings and bare land areas from Landsat 8 NDVI time series. The algorithm of the Classic Discriminant Analysis and the FSA estimators for Robust Discriminant Analysis are shown in section III-A and III-B. The results of the classification appear at the end of the paper.

II. Remote Sensing Data and Normalize Difference Vegetation Index (NDVI)

Remote sensing is the science (and to some extent, art) of acquiring information about the Earth's surface without actually being in contact with it. This is done by sensing and recording reflected or emitted energy and processing, analyzing, and applying that information [12].

The Normalized Difference Vegetation Index (NDVI) is a numerical indicator that uses the visible and near-infrared bands of the electromagnetic spectrum, and is adopted to analyze remote sensing measurements and assess whether the target being observed contains live green vegetation or not. NDVI is often directly related to ground parameters such as percent of ground cover, photosynthetic activity of the plant, surface water, leaf area index and the amount of biomass. Generally, healthy vegetation will absorb most of the visible light that falls on it, and reflects a large portion of the near-infrared light. Unhealthy or sparse vegetation reflects more visible light and less near-infrared light [13].

The study area is the zones in Karawang, West Java, Indonesia (107º02' to 107º40', 5º56' to 6º34', 1,737.30 km^2). The multispectral data of Karawang were obtained from the Landsat 8 NDVI time series. Ten multi-date Landsat 8 images listed in Table I are used for this study.

III. The Robust Discriminant Analysis for Classification

Karawang is the place of industrial activity (such as factories). However, as it continues to evolve, it has seen a heavy influx of residential development and a surge of people. One of the state-owned strategic industries also have a facility in the industrial area, i.e. Money Printing Public Company of the Republic of Indonesiawhich prints paper money, coins, and valuable documents such as passports, excise stamps, stamp duty and so forth. In agriculture, Karawang is known as the rice granary.

This paper investigates the classification of rice fields, water, buildings and bare land areas in Karawang. In order to make discriminant analysis work optimally within the classification though in the condition of data which contains of outlier, the robust estimators has been chosen. The classis discriminant analysis estimators will be replaced with robust estimators and called Robust Discriminant Analysis. The FSA has been chosen as the robust estimators to replace the classic estimators.

Table I
Ten multi-date Landsat 8 images

No	File Name	Date
1	LC81220642013141LGN01	21 May 2013
2	LC81220642013157LGN00	6 June 2013
3	LC81220642013173LGN00	22 June 2013
4	LC81220642013189LGN00	8 July 2013
5	LC81220642013205LGN00	24 July 2013
6	LC81220642013221LGN00	9 August 2013
7	LC81220642013237LGN00	25 August 2013
8	LC81220642013253LGN00	10 Sept 2013
9	LC81220642013269LGN00	26 Sept 2013
10	LC81220642013285LGN00	12 Oct 2013

All of the Landsat 8 satellite images were transformed into NDVI time-series following the equation below [13].

$$\mathrm{NDVI} = \frac{(\mathrm{NIR} - \mathrm{RED})}{(\mathrm{NIR} + \mathrm{RED})} \tag{1}$$

The Landsat 8 NDVI time series of 10-day period image is shown in Fig. 1.

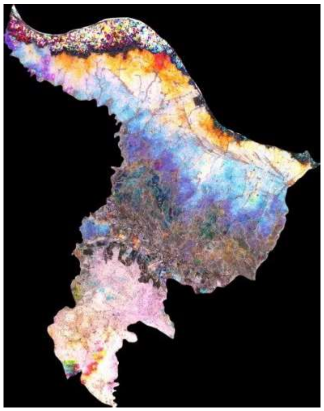

Fig. 1.The NDVI time series of Karawang, West Java.

A. The Classic Discriminant Analysis

The main purpose of a discriminant analysis is to predict group membership. The procedure begins with a set of observations where both group membership and the values of the interval variables are known. The end result of the procedure is a model that allows prediction of group membership. A second purpose of discriminant function analysis is an understanding of the data set, as a careful examination of the prediction model that results from the procedure can give insight into the relationship between group membership and the variables used to predict group membership.

The classification equation for the *j*th group is [2]

$$C_j = C_{j0} + C_{j1}X_1 + C_{j2}X_2 + \cdots + C_{jp}X_p \qquad (2)$$

Classification coefficients C_j are found from the means of the *p* predictors and the pooled within-group variance-covariance matrix *W*. the within-group covariance matrix is produced by dividing each element in the cross-product matrix S_{wg} by the within-group degrees of freedom *N*-*k*[2].

$$C_j = W^{-1}M_j \qquad (3)$$

The classification coefficients for group C_j found by multiplying the inverse of the within-group variance-covariance matrix W^{-1} by a column matrix of means for group *j* on the *p* variables ($M_j = X_{j1}, X_{j2}, \ldots X_{jp}$). The constant for group C_{jo} is found as follows [2]:

$$C_{j0} = \left(-\frac{1}{2}\right) C_j^T M_j \qquad (4)$$

The constant C_{j0} is formed by multiplying -1/2 times the transpose of the classification coefficients for group *j* times the column matrix of means for group *j*.

To assign cases into groups, a classification equation is developed for each group. Data for each case are inserted into each classification equation to develop a classification score for each group for the case. The case is assigned to the group for which it has the highest classification score.

B. The Feasible Solution Algorithm for Robust Discriminant Analysis Estimators

The Feasible Solution Algorithm (FSA) was introduced by Hawkins for the MCD estimator in multivariate data. The MCD requires a decision on *h*, the number of cases to trim. The true number of outliers is generally unknown, giving rise to two possible approaches. One is to substitute for h a 'worst-case' assumption - the value of *h* that provides the maximum breakdown point and so accommodates the maximum possible number of potential outliers. The other approach is to trim some smaller number of cases in the quite common anticipation that no more than a few cases might be outliers.

The FSA is constructed based on the swapping iteration. It investigates all possible pairwise swaps in which one of the retained cases is trimmed, being replaced by one of the trimmed cases. If the determinant of the covariance matrix of the resulting new subset is smaller than that of the old, then the old subset could not be the MCD solution and may be replaced by the new one. If there is more than one swap, that will lead to a reduction in the determinant and it will lead to the greatest reduction. If there is no possible pairwise swap that would lead to a reduction in the covariance determinant, then the current trial subset satisfies the necessary condition for the MCD optimum.

The Algorithm for the proposed method is as follows:

1. Add each X_i with a 1, defining $Z_i = (1{:}X_i)$
2. Consider *n* and *p* is the size and the dimension of the data. Compute how many data will be used to calculate the robust function [8] $h = \left[\frac{n+p+1}{2}\right]$.
3. Consider a trial partition of the cases into trimmed and retained cases. Let $J = \{i_1, i_2, i_3, \ldots, i_{n-h}\}$ be the set of indices of the currently retained cases and write the partitioned matrix. $Z_j = (Z_{i1}, Z_{i2}, Z_{i3}, \ldots, Z_{in-h})$.
4. Let $A = Z_J Z_J^T$. The FSA involves evaluating pairwiseexchanges between a retained case and a trimmed case. This can be facilitated using determinantalidentity [8][$(1 - u^T A^{-1}u)(1 - v^T A^{-1}v) + (u^T A^{-1}v)^2$].Here*u* and *v* represent the *Z* vector of the cases to be respectively trimmed and restored. The pairwiseexchange will lead to a reduction in the covariance determinant if the term in the brackets is less than 1, and the best swap to make is that for which the bracketed term is a minimum.
5. If the swap is to be made, it can be implemented using standard formulas [8]$A = A + vv^T - uu^T$for updating and downdating of the inverse.
6. Repeat process 3, 4 and 5 with random trial subsets. Following each to a feasible solution. Take the smallest determinant.
7. Compute the mean vector and covariance matrix of the subset that has the smallest determinant.
8. Repeat process 1-7 for each group (in this case: rice fields, water, building and bare land).
9. Consider *S* to be the covariance matrices of each group, then the pooled within-group variance-covariance matrix for discriminant analysis is as follows [14].

$$W = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_i - 1)s_i^2}{n_1 + n_2 + \cdots + n_i - i} \qquad (5)$$

The mean vectors, covariance matrices, and the within-group variance-covariance matrix will be used to form the discriminant functions for each group.

IV. Classification of Remote Sensing Data

This section discusses the classification of the Landsat 8 NDVI time series data of 10-period day. There are two steps involved to classify the new observations. The first step is training step. The goal of the training step is to generate discriminant functions using Classic Discriminant Analysis and Robust Discriminant Analysis for each group namely rice fields, water, buildings, and bare land that will be used in the classification step. In training step, because Robust Discriminant Analysis could generate a different function using the same data, we generate 15 different functions and 1 functions for Classic Discriminant Analysis because it will generate the same functions. We used 100 data for each group to generate each functions rice fields, water, buildings and bare land.

The second step is classification step. The classification step is performed by multiplying the raw score on each predictor X by its associated classification function coefficient C_j summing over all predictors and adding a constantC_{j0}. The case is assigned to the group for which it has the highest classification score. In this case, we used 100 data for each group.Fig. 2, 3, 4, and Table II show the Classification Result using Robust Discriminant Analysis and Fig. 5 and Table III show the Classification Result using Classic Discriminant Analysis. Table IIshows the average percentage of 15 robust experiments and Table IIIshows the percentage of classic experiment.

TABLE II
THE AVERAGE PERCENTAGE OF 15 ROBUST EXPERIMENTS

Category	Percentage
Rice Fields	74.86%
Water	65.13%
Buildings	88.73%
Bare Land	73.60%
Mean	**75.58%**

From the experiments we see that the robust discriminant analysis takes about 15 minutes to 20 minutes to generate 1 function and the classic discriminant analysis takes about 1 minute to generate 1 function because the classic discriminant does not have to do the interchanging step. The times that used to generate 1 function depends on how much data is used and how many outliers in the data. Data with many outliers will takes longer time to find the stop condition. The mean percentage of Robust Discriminant Analysis is 75.58% and for classic Discriminant Analysis is 72.50%. Based on the data used for classification, the misclassification is caused by the similarity value of rice fields and bare land and also the similarity value of water and buildings. Rice fields NDVI value range is between about 0.1 to 0.4 while bare land range is about 0.1 to 0.3. Buildings typically range from 0.1 to 0.2 and water value has approximately less than 0.2.

V. REMARK

The Robust Discriminant Analysis has a better result than Classic Discriminant Analysis where the percentage for Robust Discriminant Analysis is 75.58% and 72.50% for Classic Discriminant Analysis.Both methodsalmost have the same percentage can be caused by an error in the determination of the type of land use using eyes visualization.

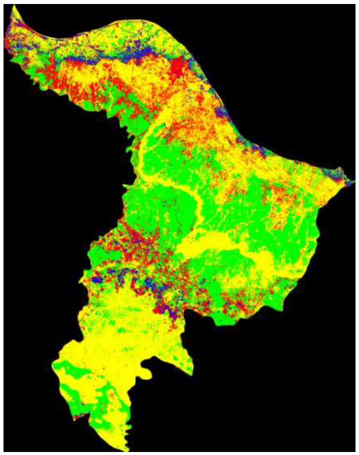

Fig. 2. Robust result number 3

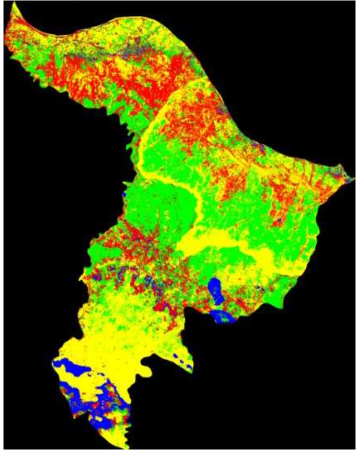

Fig. 3. Robust result number 2

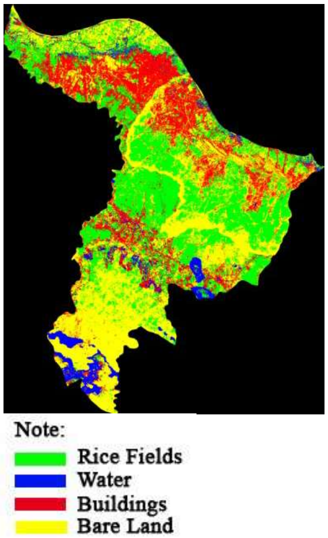

Fig. 4. Robust result number 3

TABLE III
THE PERCENTAGE OF CLASSIC EXPERIMENT

Category	Percentage
Rice Fields	75%
Water	61%
Buildings	91%
Bare Land	78%
Mean	**76.25%**

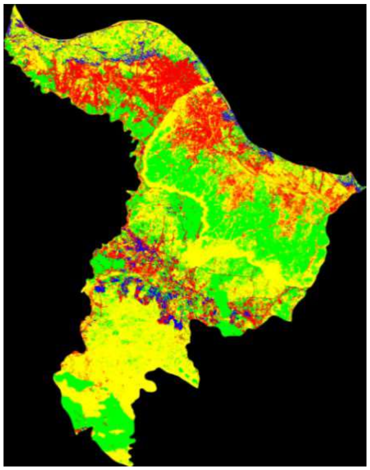

Fig. 5. Classic result

VI. ACKNOWLEDGMENT

We would like to thank the Faculty of Information Technology Tarumanagara University and the Agency for the Assessment and Application of Technology for supporting this research.

REFERENCES

[1] F.Filzmoser, K. Joossens, and C.Croux,"Multiple Group Linear Discriminant Analysis: Robustness and Error Rate", *Proceedings in Computational Statistics*, pp. 521-532, 2006.

[2] *Using Multivariate Statistics*, 6th ed., Pearson,B. T. Tabachnick, and L. S.Fidell , 2013.

[3] Croux, C. and Dehon, C. "Robust Liniear Discriminant Analysis using S-Estimators", *The Canadian Journal of Statistics,* Vol. 29, No. 7, 2001.

[4] A. College, and P.S. Altoona,"Robust M-Estimation in Discriminant Analysis", *The Indian Journal of Statistics*, Vol. 60, Series B, Pt. 2, pp. 246-268, 1998.

[5] C. Y. Chork, and P.J. Rousseeuw, "Integrating a High Breakdown Option into Discriminant Analysis in Exploration Geochemistry", *Journal of Geochemical Exploration*, 43, pp. 191-203, 1992.

[6] D. E.Herwindiati, M. A.Djauhari, and M.Mashuri, "Robust Multivariate Outlier Labelling", *Journal of Communication in Statistics Simulation and Computation*, Vol. 36 (6), pp. 1287-1294, 2007.

[7] M. Hubert, and M.Debruyne, "Minimum Covariance Determinant", *WIREs Computational Statistics,* Vol. 2, pp. 36-43, 2010.

[8] D. M. Hawkins, "The Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data",*Computational Statistics and Data Analysis*, 17, pp. 197-210, 1994.

[9] J. R. Irons, and J. L.Dwyer,"An Overview of the Landsat Data Continuity Mission", *Proceeding of SPIE*, Vol. 7695, 2010.

[10] U.S. Geological Survey:*Landsat 8*, September 2013, Available: http://pubs.usgs.gov/fs/2013/3060/pdf/fs2013-3060.pdf

[11] R.Geerken, B.Zaitchik, and J. P. Evans, "Classifying rangeland vegetation type and coverage from NDVI time series using Fourier Filtered Cycle Similarity", *International Journal of Remote Sensing*, Vol. 26, No. 24, pp. 5535-5554, 2005.

[12] Natural Resources Canada: *Fundamentals of Remote Sensing – Introduction*, 28 January 2014, Available: https://www.nrcan.gc.ca/earth-sciences/geomatics/satellite-imagery-air-photos/satellite-imagery-products/educational-resources/9363

[13] Food Security and Nutrition Analysis Unit: *Understanding The Normalize Difference Vegetation Index (NDVI)*, Available:http://www.fsnau.org/downloads/Understanding_the_Nomalized_Vegetation_Index_NDVI.pdf

[14] *PengantarStatistika*, 3rd ed., PT. GramediaPustakaUtama, Ronald E. Walpole, Jakarta, 1997.

[15] Adrian, "KlasifikasiPertumbuhanTanamanPadiMenggunakanModissSintetik", *JIKSI*, Vol. 1, No. 2, 2013.