Robust Statistics for Classification of Remote Sensing Data

Dyah E. Herwindiati, Maman A. Djauhari, Luan Jaupi

Robust Statistics for Classification of Remote Sensing Data

Dyah E. Herwindiati, *Tarumanagara University-Indonesia*, herwindiati@untar.ac.id
Maman A. Djauhari, *Universiti Teknologi Malaysia- Malaysia*, maman@utm.my
Luan Jaupi, *Conservatoire National des Arts et Metiers-France*, jaupi@cnam.fr

Abstract. The classification of remote sensing data from Landsat 7 satellite is considered, and an area under investigation is Jakarta Province. The supervised land classification is done with two processes: the training sites and classification process. A robust computationally efficient approach is applied for training site to deal with the large remote sensing data set of Jakarta. The objective of this paper is to introduce the depth function for robust estimation of a multivariate location parameter minimizing vector variance for classification of green space at Jakarta Province.

Keywords. Data depth, minimum vector variance, Mahalanobis distance, remote sensing, robust estimation

1 Introduction

Lillesand et.al [7] defined remote sensing as the science and art of obtaining information about an object through the analysis of data acquired by a device that is not in contact with the object under investigation. The instruments used for this special technology are known as remote sensors and include photographic cameras, mechanical scanners, and imaging radar systems. In many aspects, remote sensing can be thought of as a reading process, using various sensors. The analysis of the information remote sensing data is made though visual and digital image processing.

This paper discusses the robust classification of remote sensing data from Landsat 7 satellite. The supervised land classification is done with two steps; i.e. the training sites process and the classification process. The outcomes of training site are the spectral imaging references of green space area, which contains the water catchment area and vegetation area. The values are useful for classification.

Since 2003, the Scan Line Corrector (SLC) of Landsat 7 failed and the failure appears to be permanent. The non-functioning SLC causes large gaps at the edges of the image. Aiming to restore an image a gap filling procedure is necessary. However, the filling gap strategy creates

a new problem that is the anomalous image characters. The anomalous observations caused by instrument error need comprehensive approach to get the estimator of spectral reference.

The robust approach is applied to classify the pixels after the gap filling process. Hampel et.al [4] state that the major goal of robust statistics is to develop methods that are robust against the possibility that a proportion of outliers may occur anywhere in the data. This paper introduces the robust estimation of location and covariance matrix for the training site to deal with the large remote sensing data set of Jakarta.

There are several robust algorithms, for example Rousseeuw [10] introduced the criteria of minimizing the determinant of covariance abbreviated as MCD, Hawkins [3] offered an algorithm which is called the feasible solution algorithm (FSA) which ensured the optimal solution for MCD through a probabilistic approach. Next, Rousseeuw and van Driesen [11] introduced an algorithm which is called the fast MCD, noted FMCD.

The FMCD is the good robust procedure, as it has robust property of high breakdown point (BP), but Werner [13] showed that FMCD might not be efficient for for large data of high dimension. Regarding with this aspect, Herwindiati et.al [5] proposed a criterion for robust estimation of location and covariance matrix, minimum vector variance (MVV) in outlier labeling. This method is good for the application of very large and high dimension data set. It is also robust and has the same breakdown point as the FMCD. Furthermore, its computational complexity is far less than that FMCD procedure. This paper proposes the modified MVV algorithm to estimate the green space pixels in training step.

The sample images of green space are selected for the training step. The images are selected by a visual image and the Global Mapper. The objective of paper is to introduce the depth function for robust estimation of a multivariate location-scale parameter which is used for classification of remote sensing data. A new depth function which is equivalent to Mahalanobis depth can be used to replace the inversion process of covariance matrix, see Djauhari [1]. The method is applied to supervised land classification in Jakarta Province.

2 Remote Sensing Data and Preprocessing for Classification

Data remote sensing is used in acquiring information about the Earth's surface without actually being in contact with the object . Remote sensors collect data by detecting the energy that is reflected from Earth. These sensors can be on satellites or aircrafts.

Landsat satellites have been providing multispectral images of the Earth continuously since the early 1970's. Landsat data have been utilized in a variety of government, public, private, and national security applications. The purpose of the Landsat program is to provide the world's scientists and application engineers with a continuing stream of remote sensing data for monitoring and managing the Earth's resources, NASA [8]. In the past, the way to investigate a given area was through direct observation and sampling on the ground. Observations were mainly limited to those features that could be seen, photographed, or measured in visible portions of the spectrum.It was difficult to obtain global interpretations or to document changing conditions over large areas.

The case of research is the Jakarta multispectral imaging from Landsat -7 satellite. Jakarta is the capital of Indonesia. Spread over an area of around 700 square kilometres, the population of Jakarta is around 9.5 million on 2010. Land use is changed without the good planning. The quality of the environment is becoming worse day by day.

The supervised classification is done for detection of Jakarta green space areas; i.e. the water catchment area and vegetation area. The tiff formatted imaging on the year 2002 and the year 2010 are used as input. Data is captured by sensor having 7 bands involving the visible spectral , near - IR, and mid IR. The spatial resolution of 6 bands (band 1 - 5, and band 7) are 30 square meters, the resolution of the sixth band is 60 square meters. The area under investigation is Jakarta Province covered by coordinate (5 19' 12" - 6 23' 54") South Latitude and (106 22' 42" - 106 58' 18") East Longitude.

On 31 May 2003 the Landsat 7 Enhanced Thematic Mapper (ETM) sensor had a failure of the Scan Line Corrector (SLC). The SLC is an electromechanical device for the forward motion by modifying the instrument's optical path. Since that time all Landsat ETM images have the wedge-shaped gaps. The impact of failure is an approximately 20% data loss. The gap filling is the preprocessing technique used for filling missing parts of remotely sensed imagery. We do the gap filling procedure with the multi source. Figure 1 reveals the multispectral with SLC of, the gap filling technique recovers the image as shown in Figure 2.

The problems of the gap filling are inconsistencies in lighting; season and cloud cover between images. Two images of the same area may differ significantly in these properties. The inconsistent digital number of pixel is considered as an anomolous observation. The robust method is applied to classify one or several the anomolous observations in the data set.

Figure 1. The Multispectral Jakarta 2010 with SLC of Landsat

Figure 2. The Multispectral Jakarta 2010 after Preprocessing of Gap Filling

3 The Methods of Classification

Robust Minimum Vector Variance

The minimum vector variance (MVV) is an efficient robust measure minimizing the square of parallelogram diagonal length. Herwindiati et.al [5] proposed minimum vector variance (MVV) to improve FMCD algorithm.

Consider random samples $\vec{X_1}$, $\vec{X_2}$, ..., $\vec{X_n}$ from p-variate distribution of location parameter μ and a positive definite covariance matrix Σ. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ be eigen values of covariance matrix S of size $(p \times p)$, the total variance (TV) is formulated as $Tr\,(S) = \lambda_1 + \lambda_2 + \cdots + \lambda_p$; the covariance determinant (CD) is $\mid S \mid = \lambda_1\, \lambda_2\, \cdots\, \lambda_p$. Herwindiati [5] stated that the vector variance (VV) is $Tr\left(S^2\right) = \lambda_1^2 + \lambda_2^2 + \ldots + \lambda_p^2$. The advantage of vector variance consist in the fact that it is able to measure multivariate dispersions even if the covariance matrix S is singular.

The robust minimum vector variance (MVV) estimators for location parameters and covariance matrix are defined as the pair $(T_{MVV},\ S_{MVV})$ minimizing $Tr\left(S^2_{MVV}\right)$ among all possible $h = \left[\frac{n+p+1}{2}\right]$ sets H, where $T_{MVV} = \frac{1}{h}\ \sum_{i \in H} \vec{X}_i$, $S_{MVV} = \frac{1}{h} \sum_{ieH} (\vec{X}_i - T_{MVV})(\vec{X}_i - T_{MVV})^t$ and $Tr\left(S^2_{MVV}\right) = s^2_{11} + s^2_{22} + \cdots + s^2_{pp} + 2\sum_{i=1}^{p} \sum_{j \neq i}^{p} s^2_{ij}$

Computations of MVV are efficient. The efficiency of MVV is of order $O\left(p^2\right)$ compare with MCD by using Cholesky decomposition which is of order $O\left(p^3\right)$.

The Depht Function

Djauhari and Umbara [2] proposed a new depth function which is equivalent to Mahalanobis depth for reducing the level complexity of FMCD and MVV.

Let $X_1, X_2, \cdots, X_n$ be a random sample from p-variate distribution where the second moment exists. The sample mean vector and sample covariance matrix are, respectively,

$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \vec{\bar{X}}\right)\left(X_i - \vec{\bar{X}}\right)^t$, $i = 1, 2, \cdots, n$
The sample version of Mahalanobis Depht of X_i is defines as

$$MD_i = \frac{1}{1 + \left(X_i - \vec{\bar{X}}\right)^t S^{-1}\left(X_i - \vec{\bar{X}}\right)} \tag{1}$$

(Liu, 1990 cited by Djauhari and Umbara [2])

If the T^2 Hotelling's Statistics or Mahalanobis distance d_i^2 is defined as
$T_i^2 = \left(\vec{X}_i - \vec{\bar{X}}\right)^t S^{-1}\left(\vec{X}_i - \vec{\bar{X}}\right)$, then

$$MD_i = \frac{1}{1 + T^2} \tag{2}$$

$$MD_i = \frac{1}{1 + d_i^2} \tag{3}$$

The part of denominator MD_i is Mahalanobis distance, we need the inversion of sample covariance matrix S. The computational complexity of the inversion is high. To cover the problem, Djauhari and Umbara intoduced a new depht function M_i which is less complicated than mahalanobis depth,

$$M_i = \begin{bmatrix} 1 & \left(X_i - \vec{\bar{X}}\right)^t \\ \left(X_i - \vec{\bar{X}}\right) & S \end{bmatrix}$$

M_i is a matrix of size $(p+1)\times(p+1)$ is associated with $X_1, X_2, \cdots, X_n$. By using the property of determinant of a partitioned matrix we have $T_i^2 = 1 - \frac{|M_i|}{|S|}$. If $|S|$ and $|M_i|$ are respectively determinant of S and M_i, then

$$MD_i = \frac{|S|}{2|S| - |M_i|} \tag{4}$$

From the equation (4), we see that $MD_i \leq MD_j$ if and only if $(2|s| - M_j) \leq (2|s| - M_i)$;
$|M_i|$ and MD_i are defined as the same multivariate ordering.
The good characteristic of $|M_i|$ is that its calculation does not need any matrix inversion.

4 The Algorithm of Training and Classification Step

The process of classification is done with two steps. The first step is the training site and the second one is the classification step. To conduct the training step, the Modified Robust Minimum Vector Variance (DMVV) is considered for calculating of green space spectral. The DMVV is robust estimation of a multivariate location parameter, which minimize a vector variance; and it uses the advance of a depth funtion to replace the inversion process of covariance matrix. The goal of training step is to predict the range of reference green space spectral area; i.e. water chatment area and vegetation area;
The algoritm is described as follows,

Algorithm 4.1.

Step 1 *Crop image of the vegetation area in size* $(a \times a)$ *pixel based on the RGB color space of multispectral visual and Normalized Difference Vegetation Index (NDVI) for a training data set.*

Step 2 *Assume that the data set of p-variate observations is* $\{\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n\}$

Step 3 *Let* $H_0 \subset \{1, 2, \cdots, n\}$ *with* $|H_O| = h$ *and* $h = \left[\frac{n+p+1}{2}\right]$

Step 4 *Compute the mean vector* $\vec{\bar{X}}_{H_O}$ *and covariance matrix* S_{H_O} *of* H_O

Step 5 *Compute* $M_i = \begin{vmatrix} 1 & \left(X_i - \vec{\bar{X}}_O\right)^t \\ \left(X_i - \vec{\bar{X}}_O\right) & S_O \end{vmatrix}$ *for* $i = 1, 2, \cdots, n$

Step 6 *Sort* M_i *in decreasing order,* $M_{\pi(1)} \geq M_{\pi(2)} \geq \cdots \geq M_{\pi(n)}$

Step 7 *Define* $H_W = \left\{\vec{X}_{\pi(1)}, \vec{X}_{\pi(2)}, \cdots, \vec{X}_{\pi(h)}\right\}$

Step 8 *Calculate the new mean vector and covariance matrix of* H_W*, they are noted as* $\vec{\bar{X}}_{H_W}$ *and* S_{H_W}

Step 9 *If* $Tr\left(S^2_{H_W}\right) = 0$ *the process is stopped. If* $Tr\left(S^2_{H_W}\right) \neq Tr\left(S^2_{H_O}\right)$ *repeat steps (2 – 8) the process is continued until the k-th iteration if* $Tr\left(S^2_k\right) - Tr\left(S^2_{k+1}\right) \leq \varepsilon$ *and* ε *is a small constant*

Step 10 *Let* $\vec{T}_{VV}$ *and* S_{VV} *be the location and covariance matrix given by that process. Robust squared MVV Mahalanobis distance for the data training set is defined as,*

$d^2_{VV}\left(\vec{X}_i\,,\,\vec{T}_{VV}\right) = \left(\vec{X}_i - \vec{T}_{VV}\right)^t S^{-1}_{VV}\left(\vec{X}_i - \vec{T}_{VV}\right)$*, for all i = 1, 2, . . . , n.*

Step 11 *Determine the range of each green space spectral area; i.e. water chatment area and vegetation area; called as* $c_1 \leq d_{VV} \leq c_2$ *from the robust distance as stated at point 10, where* c_1 *is the first quartile and* c_2 *is the third quartile.*

The illustration of cropped images in the Step 1 is shown in Figure 3.

Channel-1 Channel-3 Channel-5

Figure 3. The example of cropped vegetation images in Chanel 1,3 and 5

The range of vegetation area spectral is $4.570382 \leq d_{VV} \leq 12.839$. The Figure 4 and 5 illustrate the dispersion of d_{VV} and the dispersion of d_{VV} inside the interval $4.570382 \leq d_{VV} \leq 12.839$ respectivelly.

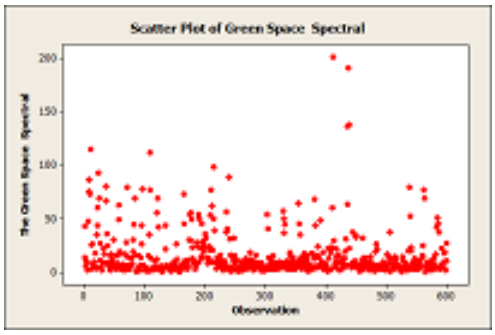

Figure 4. Scatter Plot of Green Space Spectral

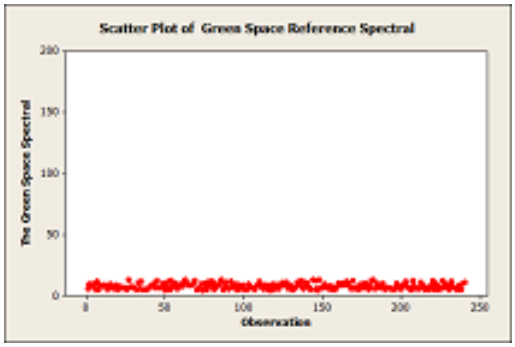

Figure 5. Scatter Plot of Green Space Reference Spectral

The classification step is done by using the reference spectral in the training step. Assume $Y_1, Y_2, \cdots, Y_M$ are the pixels of hole imaging Jakarta having p-variates. The distance $d^2_{Cls}\left(\vec{Y}_i, \vec{T}_{VV}\right)$ is done to classify each pixel in three classes; i.e. Water catchment area, vegetation area and impervious area, where $d^2_{Cls}\left(\vec{Y}_i, \vec{T}_{VV}\right) = \left(\vec{Y}_i - \vec{T}_{VV}\right)^t S^{-1}_{VV}\left(\vec{Y}_i - \vec{T}_{VV}\right)$, for $i = 1, 2, \cdots, M$. The impervious surface is defined as surface impenetrable by water including side walks, street, highway, parking lot and rooftops [4], observation $\vec{Y}_i$ is classified as the impervious area if $d^2_{Cls}\left(\vec{Y}_i, \vec{T}_{VV}\right)$ is not in the interval $c_1 \leq d_{VV} \leq c_2$.

Figure 6 and 7 are the classification of Jakarta on the years 2002 and 2010. The vegetation area is labeled with the green color, the water catchment area is colored in yellow, and the grey color is the impervious area.

On the year 2002, the percentage of Jakarta green space is around 10.2569% and the area are increased to 11.24568% on the year 2010. Table 1 gives the percentage of green space on the year 2002 and 2010.

The water cathment area on the year 2010 is significantly greater than on the year 2002. The biggest increased area is in Halim Perdana Kusuma. Figure 8 shows the changed used land, the blue color means the increasing water cathment area and the decreasing one is colored in red. Halim Perdana Kusuma is signed in white circle.

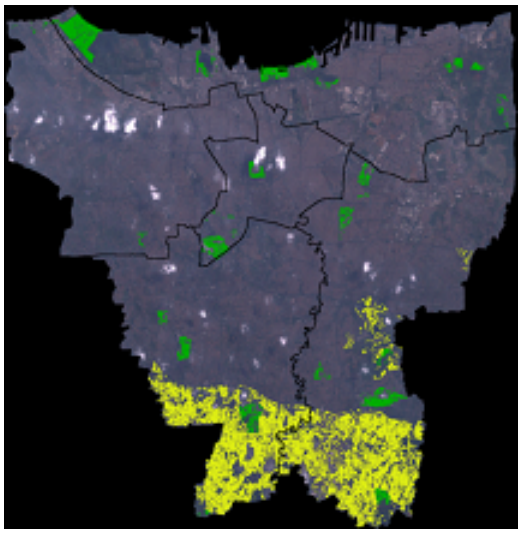

Figure 6. The Classification of Jakarta on 2002

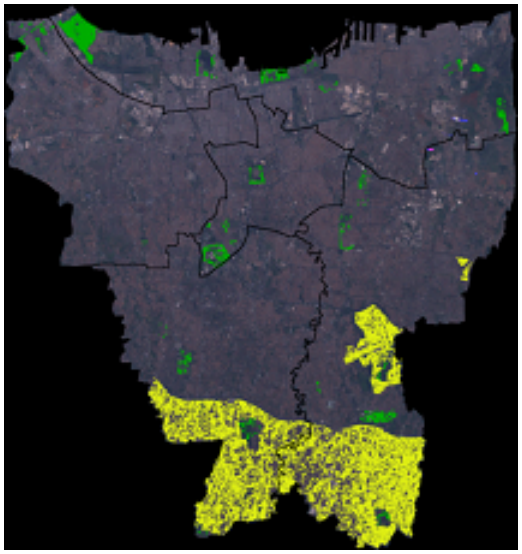

Figure 7. The Classification of Jakarta on 2010

Table 1. The Percentage of Green Space of Jakarta

The Year	The Green Space area		Total
	Water Cathment Area	Vegetation Area	
2002	8.16079%	2.09611%	10.2569%
2010	9.6937%	1.55198%	11.24568%

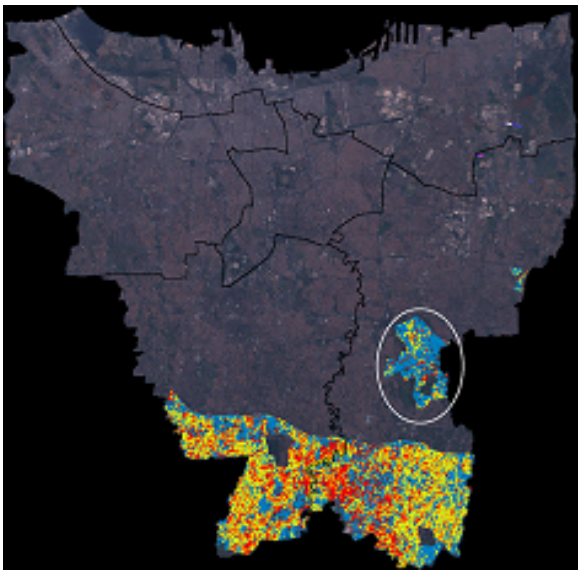

Figure 8. The Change Vegetation Area – Jakarta for year 2002-2010

5 The Evaluation of Results

The Visualization of Area

Figure 9 tells about the real visual Halim Perdana Kusuma after forestation and reforestation by Google Earth. The former Jakarta Governor working period on year 2000 – 2008 made the effort to repair and build Jakarta. The green land project budget was increased and the Governor also determined ***The Special Rules Capital Regional District Jakarta Province Number 8 of 2007*** for management and implementation of protected areas. We remembered that in May 1998 there was a riot in Jakarta, a massive destruction and burning in all areas of Jakarta.

Figure 9. Halim Perdana Kusuma – Jakarta on 2011

Simulation Experiment of DMVV in Training Process

The breakdown point is defined as the smallest amount of contamination that causes an estimator to brake down and take on infinite values. Consider random samples $\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n$ of random vectors $\vec{X}$ and a location estimator $T_n(\mathrm{X}) \in \Re^p$, suppose the m data for which the values are changed to be infinity and the location estimator $T_{n^*}(\mathrm{X}^*)$, so that
the maximum $bias\left(m, T, \vec{X}\right) = \sup_{\mathrm{X}^*} \| T_{n^*}(\mathrm{X}^*) - T_n(\mathrm{X})\|$. The breakdown point is defined as the smallest fraction of contamination that can cause the estimator T_{n^*} on values arbitrarily far from T_n, see Roussew and and Leroy [12].

We did the simulation of breakdown point by using normal data with shift at location parameter. The DMVV has a high break down point. For experiment, we generate $(n - n_{OUT})$ data points from $p-$dimensional random variable X following a distribution given as a mixture of normal distribution following of the form $(1-\alpha)\, N(O,\, I) + \alpha\, N(\delta e_1\,,\, \lambda I\,)$, where e_1 denotes the first unit vector, δ and λ are constants $(n = 100\,; p = 10)$.

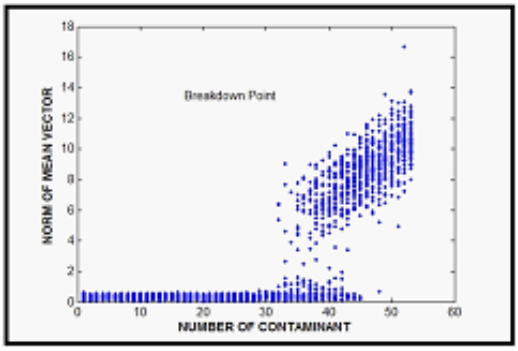

Figure 10. The Breakdown Point of DMVV

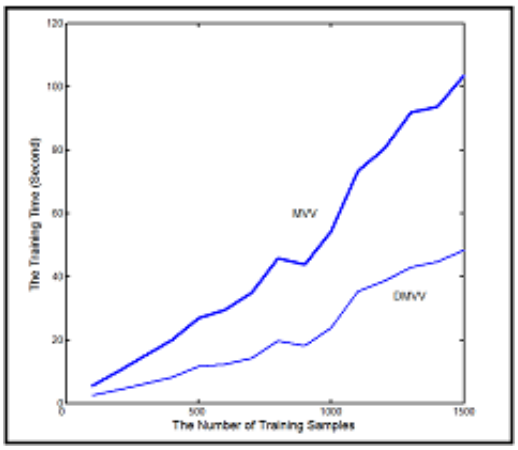

Figure 11. The Computation Time (Second) in Training Process

The DMVV is very efficient method for classification of a large remote sensing data. Figure 11 illustrates the comparison time (second) of MVV and DMVV to estimate the spectral reference of green space in training process. The DMVV has lower computation time (second) than MVV. The simulation is operated by MATLAB 8.00 and the spesification of processor is Intel(R)Core(TM)i7CPU-RAM 4.00 GB.

6 Additional Remark

An advantage of $|M_i|$ as a measure of the depth consists that does not need any matrix inversion in its computation. It only needs to compute the determinant of a symmetric matrix. The modified minimum vector variance with depth function (DMVV) is an efficient and effective robust approach for classification of remote sensing data. The empirical result of classification proved that the DMVV is able to reduce computational time and is a reliable robust method with high breakdown point.

Bibliography

[1] Djauhari, M.A (2008) *A Robust Estimation of Location and Scatter.* Malaysia Journal of Mathematical Sciences, 2(1),1-24.

[2] Djauhari M.A and Umbara R.F (2007) *A Redefinition of Mahalanobis Depth Function, Journal of Fundamental Sciences*, 3(1),150-157.

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

[4] Hampel, F.R., Ronchetti, E. M., Rousseuw, P.J. and Stahel, W.A. (1985) *Robust Statistics*, John Wiley, New York.

[5] Herwindiati, D.E., Djauhari, M.A. and Mashuri, M. (2007), *Robust Multivariate Outlier Labeling*, J. Communication in Statistics Simulation And Computation, 36, No 6, 1287-1294.

[6] Huber, P.J. (1980), *Robust Statistics*, Massachusetts, Wiley Series in Probability and Mathematical Statistics.

[7] Lillesand, T.M., Kieffer, R.W. and Chipman, J.W. (2007) *Remote Sensing and Image Interpretation*, Hoboken, NJ : John Wiley and Sons

[8] National Aeronautics and Space Administration *Landsat 7 Science Data Users Handbook* $http://landsathandbook.gsfc.nasa.gov/pdfs/Landsat7_Handbook.pdf$.

[9] Natural Resources Canada *Fundamental of Remote Sensing* $www.nrcan.gc.ca/sites/www.nrcan.gc.ca.../fundamentals_e.pdf$.

[10] Rousseeuw, P.J. (1985), *Multivariate Estimation with High Breakdown Point*, Paper appered in Grossman W., Pflug G., Vincze I. dan Wertz W., editors, Mathematical Statistics and Applications, B, 283-297. D. Reidel Publishing Company.

[11] Rousseeuw, P.J. and van Driessen, K. (1999), *A Fast Algorithm for The Minimum Covariance Determinant Estimator*, Technometrics, 41, 212-223.

[12] Rousseeuw, P.J. and Leroy A. M. (1987), *Robust Regression and Outlier Detection* , John Wiley, New York.

[13] Werner, M. (2003) *Identification of Multivariate Outliers in Large Data Sets*, PhD Thesis, University of Colorado at Denver.