Robust Reduction Dimension for Mapping of Rice Field

D. E. Herwindiati, L. Jaupi, and S. Mulyono

***Abstract*—Mapping of rice field is done with a conventional two step process: training process and classification.The results of mapping process are highly influenced by accuracy of spectral reference obtained in training process. Robust reduction dimension improvements are proposed for computing estimators. The first improvement consists in a modification of robust subset with preliminary data inspection. The inspection is useful for screening and removing the potential outliers. As a second improvement the replacement of process inversion of covariance matrix with a new depth function is proposed. The case study of research is rice fields located in Karawang, West Java. Data from MODIS (Moderate Resolution Imaging Spectroradiometer) satellite are used for rice field mapping.**

***Index Terms*— C-Step, depth function, minimum vector variance, principal component analysis, remote sensing, robustness.**

I. Introduction

Rice is the main staple consumption in Indonesia. It is a very important commodity for most Indonesian people. The growth of Indonesian population during the period from year 2000 to year 2010 is 1.49 percent. The census in 2010 stated that the population is approximately 237.56 million people. As the population growth rate continues to grow, the demand for rice will continue to increase every year. On the other hand the number of farmers during this period has decreased. Apriyana [11] stated Indonesian needs 13 millions hectare productive land on 2012 and it only has 8.1 millions hectare available.

Indonesia is one of the good rice producers in the world. The climate and geography of Indonesian are suitable with the rice plant, the problem of reducing rice production should not be found. Decreasing production capacity of rice have caused decreasing of capacity in food supply. Government policies to increase of rice production capacity have been done, such as: rehabilitation and extensification in irrigation infrastructure, expansion of new land for rice, and the acceleration of technology innovation, including revitalization of development research.

This paper discusses our research on rice mapping using data remote sensing. Lillesand et.al [18] 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. Data of research are supported by Indonesian Agency for the Assessment and Application of Technology (BPPT). The mapping process is done through two steps; the training process and the mapping or classification process. In the first process, data are collected by ground-based sensor and in the next process; i.e. mapping or classification process, data come from multispectral of MODIS sattelite (*Moderate Resolution Imaging Spectroradiometer*). The case study of this research is rice plantation field in Karawang, West Java, Indonesia.

Many data mining approaches have been used for classification processes. In this case we introduce the robust dimension reduction method for mapping of rice field. Classification is one technique of data mining used to predict group membership based on information on one or more characteristics of data. In this research, we use terminology ‘mapping’ for classification process

Several problems might appeare in the mapping process, such as unprecission of ground-based data collection, data of training process have the tendency to be collinear, and the inconsistency of the weather during the sattelite capturing of objects. The robust dimension reduction is believed to be the solution of the problems.

Principal component analysis (PCA) is the most commonly used for dimension reduction technique. The main idea PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set, Jollife[7].

Robust principal component (ROBPCA) is the famous robust dimension reduction method introduced by Huber et al [9]. ROBPCA is a method combining two

D. E. Herwindiati is with Tarumanagara University, Let.Jen.S.Parman 1, Jakarta 11440, Indonesia (corresponding author phone: +62215677464 fax: +622156941924, e-mail: herwindiati@untar.ac.id).

L. Jaupi is with Conservatoire National des Arts et Métiers, 292 rue Saint Martin, 75003 Paris, France (e-mail: jaupi@cnam.fr).

S. Mulyono is with The Indonesian Agency for the Assessment and Application of Technology (BPPT), M.H.Thamrin 8, Jakarta 10340, Indonesia (e-mail: sidik.mulyono@bppt.go.id).

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

advantages of both projection pursuit and robust method minimizing covariance determinant (MCD). ROBPCA has good properties but the computation is still elaborative for images classification. Herwindiati and Isa [4] introduced MVV robust PCA for reducing the time of computation.

MVV robust PCA is defined as robust method minimizing vector variance (VV) with dimension reduction. The vector variance (VV) is multivariate dispersion that is formulated as $Tr\left(\Sigma^2\right)$, geometrically VV is a square of the length of the diagonal of a parallelotop generated by all principal components of $\vec{X}$.

This paper deals with mapping rice field by using MODIS sattelite data with spatial resolution (500m x 500m). One pixel of MODIS can acquire (500m x 500m) of land covers, NASA [11]. It means that one pixel gives the great information of variety of land covers. Indeed for robust computation the convergence of estimator is not fast. The estimator is fluctuated till at last it meets its stability. An efficient and effective robust method must be used to encounter the problems. An improvement of MVV Robust PCA is introduced. The objective of this paper is to enhance MVV robust PCA for mapping rice field . The enhancements consist of the modification of robust subset selection and replacement of inversion process of covariance matrix with a new depth function proposed by Djauhari and Umbara [8].

II. Ground-Based Data and Outlier

Ground-based data that is used as input in our training process of mapping rice research are hyper spectral data of rice plants collected by using International Light (ILT900) spectrometer which has wavelength range between 250-900 nm and its sensor has 25 degree of field of view (FOV) angle.

To acquire the spectral of rice plant as an object, the sensor is located at least 2 meter height above the objects in order to be able to cover about 1m x 1m area. The spectral itself is represented by reflectance of light transmitted into the sensor, and its properties are then converted into digital number of 2048 channels. The obtaining data were logged directly into notebook computer which integrated with the equipment during observation. This manner is recurrently done 8 times during rice planting period of April – July 2012 in experiment farmland located in Subang district of West Java. Moreover, to simplify the data, all of the obtained data were adjusted and reformed into 4 channels of MODIS data by computational process.

Supervised ground-based data collection is done by user interaction. The potential anomolous observation is appeared in the training process. Beckman and Cook [17] divided outliers into two major categories (cited from Anscombe 1960). First, there might be errors in the data due to some errors; and second, outliers may be present from the inherent variability of the data. The robust method deals with a very real problem in statistical applications, the robust estimator provide a reliable classification when the data contain outliers.

There are some robust criteria proposed to get effective estimators. The most well known criterion is to minimize the volume of ellipsoid of a parallelotop. Among them, MVE (minimizing the volume of ellipsoid) and MCD (minimizing the covariance determinant) introduced by Rousseeuw [14] are the most popular. However, in recent years MCD receives much more attention than MVE dueto its performance in estimating the true location and scatter. Some improved versions of MCD algorithm are available, for example feasible solution algorithm in Hawkins [5] and Hawkins and Olive [6] Fast MCD (FMCD) algorithm in Rousseeuw and van Driessen [15], block adaptive computationally-efficient outlier nominators (BACON) in Billor et al. [13], improved FMCD algorithm in Hubert et al. [9]

The main objective of this paper is to introduce an effective and efficient method for mapping rice field using the supervised mapping field. The supervised mapping field is done with two steps: the training process process and the mapping or classificafion of land. The robust dimension reduction method is implemented to this research. The following discussion will explain the concept of the used method.

III. Principal Component Analysis

Feature selection or reduction in remote sensing has been used for different purposes. The most exploration is used for classification of multispectral or hyperspectral image. One of the most common forms of feature reduction is PCA. The main idea PCA is to reduce the dimensionality of data set consisting of a large number of inter related variable, while retaining as much as possible of variation in the data set, see Jollife[7]. S. Mulyono[1] used a genetic algorithm based new sequence principal component regression (GA-NSPCR) on how to effectively reduce the number of those bands with high accuracy for reliable rice yield prediction.

Computation of the principal components reduces to the solution of an eigenvalue-eigenvector problem for a positive-semidefinite symmetric matrix. Suppose that the random vector $\vec{X}$ of p components has the classical covariance matrix S which is a $p \times p$ symmetric and positive semi definite,

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{12} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Covariance matrix S has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and eigenvector U such that $U'SU = L$; where L is diagonal matrix.

Principal components are weighted linear combinations of $\vec{Y}$ whose variances are as large as possible. The first principal component is given by $\vec{Y}_1 = \vec{U}_1' X$ which has the largest proportion of total

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

variance. Technically, the pricipal components can be defined as a linear combination of optimally-weighted observed variable, they are orthogonal to and independent of other components.

The proportion of total variance of the $-k$ principal component is often explained by the ratio of the eigenvalues $\lambda_k = \sum_{i=i}^{k} \lambda_i$. The decision to choose the best low dimensional space can be determined by the largest proportion of total variance

The classical PCA, which is based on sample mean and sample covariance is very sensitive to outliers in the training data set. The k principal component consisting of the largest proportion of total variance S is often pushed toward the outliers, see Herwindiati and Isa [4]. The efficient and effective robust dimension reduction is described in the next section.

IV. The MVV Subset Modification Algorithm

Robust subset algorithms have been proposed in recent years for a large range of applications. In brief, the task of the robust subset algorithm can be interpreted as that of selection of subset from a data set that is then used to calculate the robust estimator. We know that Rousseeuw and van Driessen [15] approximated the FMCD estimator by searching among all subsets containing half of the data that is most tightly clustered together. This subset has minimum generalized variance or minimum covariance determinant. The FMCD algorithm is fast and high breakdown point robust procedure that is constructed based on the so-called concentration step (C-step).

Herwindiati et al. [3] proposed a criterion for robust estimation of location and covariance matrix minimizing vector variance (VV), the method is known as minimum vector variance (MVV). The subset of MVV is guided on C-step. The robust MVV is in progress to be implemented in the applications of problems in data mining; especially for cheap and fast computational time. For large data such as hyper spectral remote sensing data, the C-step is not efficient.

The subset modification is introduced for robust computation in training process. The preliminary of modification is conducted by Z-Score approach which is useful to screen and remove the potential outliers. The Z-Score is used for data screening before C-Step is started.

Iglewicz and Hoaglin [2] proposed the resistant Z-Score to remove the potential outliers. The resistant Z-Score is defined as

$$M_i = \frac{0.6745\left(x_i - \bar{x}\right)}{\text{MAD}} \qquad (1)$$

where : the estimator MAD (the median of the absolute deviation) $\text{MAD} = median_i \left\{ \left| x_i - \bar{x} \right| \right\}$ and the constant 0.675 is calculated from E(MAD) for large n. The observations are potential outliers if $M_i > D$ and D is constanta ($D = 3.5$) calculated from a simulation study.

The following experiments describe the performance of subset MVV and subset modification of MVV.

The numbers of 1000 data are generated from the multivariate normal mixture model,

$$(1-\varepsilon)\, N_3(\vec{\mu}_1, \mathrm{I}_3) \;+\; \varepsilon\, N_3(\vec{\mu}_2, \mathrm{I}_3)\,,$$

with $p = 3$, $\varepsilon = 0.05$, $\vec{\mu}_1 = \vec{0}$, $\vec{\mu}_2 = 8\,\vec{e}$, and $\vec{e}$ is a vector of dimension 3 and all of its components are having value 1. Figure 1 and Figure 2 show the MVV subset and-subset modification of MVV respectively. The ellipsoide of MVV subset modification is more concentrated than MVV Subset, consequently the minimum of vector varianve is faster to be convergence

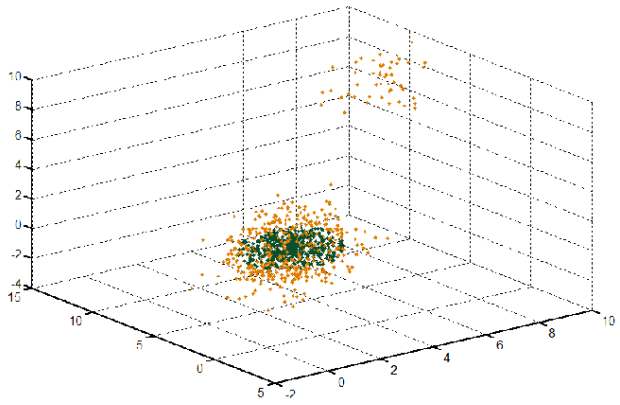

Figure 1. Subset of MVV

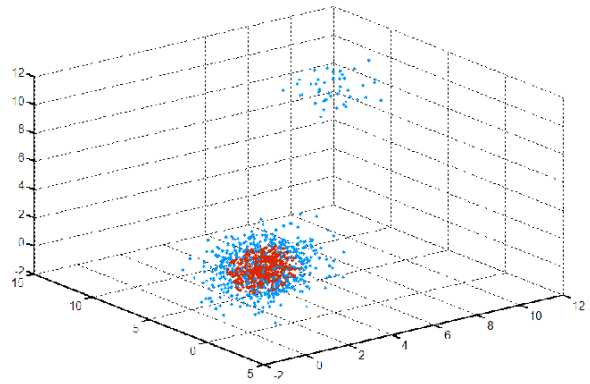

Figure2. Subset Modification of MVV

The next experiment is our experience when we to compute of robust estimator in the training process. The small data training (it has only 574 onservations) from 4 channels is treated. We use two subsets to compute the estimator. Here, we see the subsets selections of 'original' C-Step which are done with several replications and the minimum vector variance is oscillated to be convergence. Preliminary data inspection is used to reduce the high computational time. The result of experiment is shown in Figure 3.

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

As shown in this figure, the C-step is running on two approaches, and the modification subset algorithm is faster to the convergence since the modification subset of MVV only needs 9 replications of C-step, while MVV needs 15 replications of C-Step.

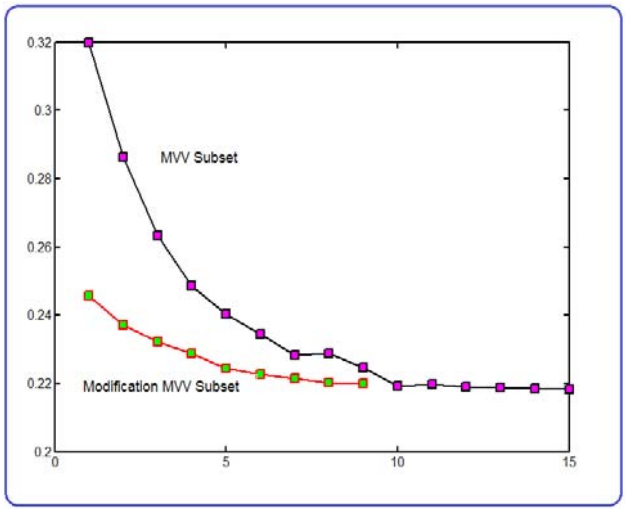

Figure 3. The Comparisson of C-Step

V. The Depth Function and Robust Estimator Computation

Covariance matrix plays an important role in multivariate data analysis. The inversion of covariance matrix is one of the problems encountered in robust estimation. Djauhari and Umbara [8] introduced a new depth function M_i; which is equivalent to Mahalanobis depth, but it is less complicated than Mahalanobis depth to replace the inversion process of covariance matrix.

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,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{\bar{X}}\right)\left(X_i - \bar{\bar{X}}\right)^t$$

$(i = 1, 2, \cdots, n)$

The new depth function M_i is defined as

$$M_i = \begin{bmatrix} 1 & \left(X_i - \bar{\bar{X}}\right)^t \\ \left(X_i - \bar{\bar{X}}\right) & S \end{bmatrix} \qquad (2)$$

M_i is a matrix of size $(p+1)\times(p+1)$ is associated with $X_1, X_2, \cdots, X_n$. The good characteristics of $|M_i|$ is that the measure does not need any matrix inversion in its computation.

The depth function M_i is applied to the robust algoritm of training process, the inversion process of covariance matrix is replaced with M_i. The detailed explanation of algorithm is available in Section VII.

VI. Case Study

The case study of research is rice fields located in Karawang, West Java. Data from MODIS (*Moderate Resolution Imaging Spectroradiometer*) satellite is used for mapping rice field.

Modis satellite is one of sattelites provided by EOS (Earth Observing System). The satellite rotates the surface of the earth once in one or two day(s). There are 36 spectral bands receiving the wave length. MODIS plays a significant role in validation development, global, interactive earth system model, prediction the global changing in an accurate way in supporting policy makers in creating right decision about the environment conservation, see Boccardo et.al [16]. The Indonesian Agency for the Assessment and Application of Technology (BPPT) have supported and provided the ground based hyper spectral data of rice plant for the research.

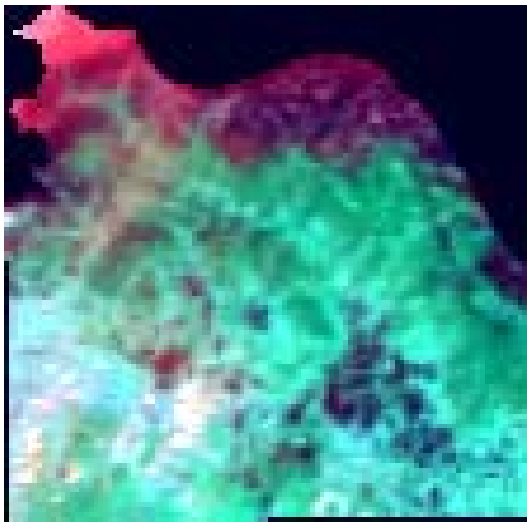

Figure 4. Karawang-West Java Using MODIS Satellite, in size (100x100) pixels

VII. Training and Mapping Process

Mapping of rice field is defined as classification as rice field and as non rice field. The rice field is categorized to vegetative rice field and reproductive rice field. The result of mapping field is highly influenced by the result of training process, therefore the use of powerful training algorithm is needed in providing acurate spectral reference in mapping process.

Two data sources given from two different measures; ground-based sensor and MODIS satellite sensor; are used in this research. The transformation of data sources are considered to have spectral reference of rice plant and mapping rice field. The transformation of source causes the potential problem in the training process.

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

A. Training Process

The objective of training step is to predict the range of reference spectral of rice plant consisting of two phases of plant; i.e vegetative and reproductive phase.

Training process is done before mapping process. The implementation of algorithm of training process is:

i. To reduce dimension reduction
ii. To select the preliminary initial of data subset using the Z- Score approach
iii. To estimate robust estimator using the depth function M_i
iv. To calculate the range spectral reference of two phases of rice plant (vegetative and reproductive phase).

The robust algoritm of training is described as follows,

1. Assume a data set of *p*-variate observations is $\{\vec{X}_1, \vec{X}_2, \cdots, \vec{X}_n\}$
2. Let $H_0 \subset \{1, 2, \cdots, n\}$ with $|H_0| = h$ and $h = \left[\frac{n+p+1}{2}\right]$
3. Compute the mean vector $\vec{\bar{X}}_{H_O}$ and covariance matrix S_{H_O} of H_O
4. 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 $i = 1, 2, \cdots, n$
5. Sort M_i in decreasing order, $M_{\pi(1)} \geq M_{\pi(2)} \geq \cdots \geq M_{\pi(n)}$
6. Define $H_W = \{\vec{X}_{\pi(1)}, \vec{X}_{\pi(2)}, \cdots, \vec{X}_{\pi(h)}\}$
7. Calculate the new mean vector and covariance matrix of H_W, that are $\vec{\bar{X}}_{H_W}$ and S_{H_W}
8. 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
9. 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 calculated from robust estimator $\vec{T}_{VV}$ and S_{VV} the distance is defined as,

$$d^2_{VV}\left(\vec{X}_i, \vec{T}_{VV}\right) = \left(\vec{X}_i - \vec{T}_{VV}\right)^t S^{-1}_{VV}\left(\vec{X}_i - \vec{T}_{VV}\right) \quad (3)$$

for all $i = 1, 2, \ldots, n$.

The last step of training process is to predict the range spectral reference of two phases of rice plant. The spectral references are useful as guidance of mapping process. The boxplot of robust distance is applied as tool to determine the spectral references. In this case, spectral reference of vegetative phase is $0.0034 \leq d_{VVG} \leq 11.998$ and $12.023 \leq d_{VVP} \leq 42.987$ is spectral reference of reproductive phase.

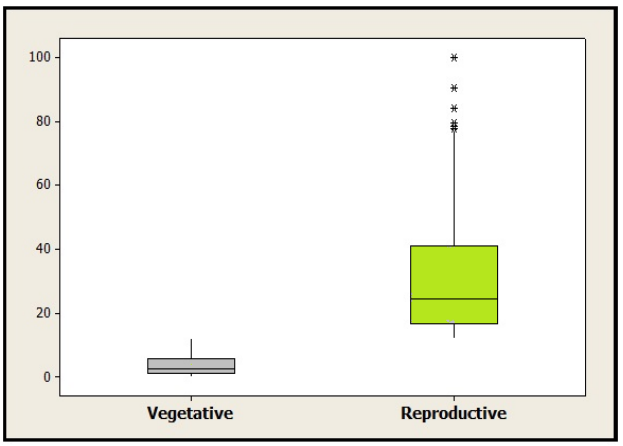

Figure 5. The Boxplot of Robust Distance

B. Mapping of Rice Field

The mapping of rice field using MODIS satellite image collected on March 25th is done for Karawang-West Java area with the spectral references in the training step.

Consider $\vec{z}_1, \vec{z}_2, \cdots \vec{z}_n$ are the pixels of whole imaging karawang having *p*-variate. The spectral references are used to classify of whole area. Assume, $\vec{T}_{VVG}$ and S_{VVG} are being the location and covariance matrix of vegetative phase, the distance $d_{RG}\left(\vec{z}_i, \vec{T}_{VVG}\right)$, where $i = 1, 2, \ldots n$ is computed to classify an each pixel . The pixels are belonging due to vegetative rice plant if the distances are in the vegetative spectral range $0.0034 \leq d_{RG} \leq 11.998$.

Figure 6 is the result of mapping of Karawang rice field. The vegetative phase of rice field is labeled with the soft green color, the productive phase of rice plant colored in the dark green. The result of mapping is given in Table 1.

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

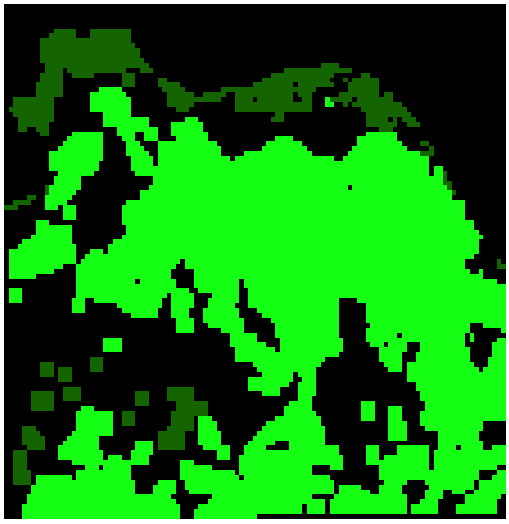

Figure 6. Mapping of Karawang Rice Field, in size (100x100) pixels

TABLE 1.
THE RESULT OF RICE FIELD MAPPING

	The Phase	
	Vegetative	Reproductive
Number of Pixel	4711	769
The Area Prediction in Hectare	11777.5	1992.5

VIII. CONCLUSION

This paper presents new alternative method for robust estimation in rice field mapping. This method is useful to enhance the mapping result, which consist of the modification of robust subset selection and replacement inversion process of covariance matrix with a new depth function M_i . The benefit of our method in training process are the spectral references of rice plant are well estimated, and the computational cost is become faster.

REFERENCES

[1] S. Mulyono, Student Member, Mohamad Ivan Fanany, T Basaruddin, "Genetic Algorithm Based New Sequence Of Principal Copmponent Regression (GA-NSPCR) For Feature Selection And Yield Prediction Using Hyperspectral Remote Sensing Data", International Geoscience And Remote Sensing Society (IGARSS) 2012, IEEE.

[2] B. Iglewicz, B. and D.C Hoaglin, "How to Detect and Handle with Outliers", American Society for Quality, Statistics Division, Vol 16, pp. 9-13, Milwaukee (1993).

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

[4] D.E. Herwindiati and S.M. Isa. "The Robust Principal Component Using Minimum Vector Variance". Proceedings of the World Congress on Engineering 2009 Vol I, pp 325-329, WCE 2009, July 1 - 3, 2009, London, U.K.

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

[6] D.M. Hawkins and D.J. Olive. (1999), Improved Feasible Solution Algorithm for High Breakdown Estimation, *Computational Statistics and Data Analysis*, Vol. 30, pp 1-11.

[7] I. T. Jolliffe, "Principal Component Analysis, 2nd Edn", New York: Springer-Verlag (2002).

[8] M.A. Djauhari M.A and R.F Umbara (2007). A Redefinition of Mahalanobis Depth Function, *Journal of Fundamental Sciences*, Vol 3. No.1, pp. 150-157

[9] M. Hubert, M., P.J. Rousseeuw and K. Vanden Branden. (2005). ROBCA: *a New Approach to Robust Principal Component Analysis*, *Technometrics*, Vol. 47, pp. 64-79

[10] N. Apriyana (2012). Kebijakan Pengendalian Konversi Lahan Pertanian Dalam Rangka Mempertahankan Ketahanan Pangan Nasional. [online] available: http://www.bappenas.go.id/blog/wp-content/uploads/2012/10/7_Policy-Paper-Pak-Nana-Apriana-Copy.pdf

[11] National Aeronautics And Space Administration. Specifications. http://modis.gsfc.nasa.gov/about/ specifications

[12] Natural Resources Canada: *Fundamental of Remote Sensing*, 28 January 2010, Available : http://www.ccrs.nrcan.gc.ca/index_e.php

[13] N. Billor, A.S. Hadi and P.V. Velleman. (2000). BACON: blocked adaptive computationally efficient outlier nominators, *Journal of Computational Statistics and Data Analysis*, Vol. 34, pp. 279 -298.

[14] P.J. Rousseeuw. (1985). Multivariate Estimation with High Breakdown Point, *in Mathematical Statistics and Applications*, Vol. B, eds, W. Grossman, G. Pflug G, J. Vincze and W. Wertz, Dordrecht: Reidel, pp. 283-297.

[15] P.J. Rousseeuw, and K. Van Driessen. (1999). A Fast Algorithm for The Minimum Covariance Determinant Estimator, *Journal of Technometrics*, Vol. 41, No. 3, pp. 212-223.

[16] P. Boccardo, E.B. Mondini, P. Claps and F. Perez. Image Ressolution Effect Vegetation Mapping From Landsat 7 ETM+ And Terra Modis Data. [online] available: http://www.idrologia.polito.it/~claps/Papers/Prin03_Boccardo.pdf

[17] R.J. Beckman, and R.D. Cook. (1983). Outlier …s, *Technometrics*, Vol 25, No 2, pp 119 – 149.

[18] T.M.Lillesand, R.W. Kiefer and J.W. Chipman. "Remote Sensing and Image Interpretation", Hoboken, NJ : John Wiley & Sons, (2007)

ISBN: 978-988-19252-9-9
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)