Journal of Mathematics and System Science 3 (2013) 180-186

Robust Classification of Remote Sensing Data for Green Space Analysis

Dyah E. Herwindiati[1], Maman A. Djauhari[2] and Luan Jaupi[3]

1. Faculty of Information Technology, Tarumanagara University, Let.Jend. S. Parman No1, Jakarta 11440, Indonesia

2. Department of Mathematical Sciences, Faculty of Science, Universiti Teknologi Malaysia, Skudai, Johor Bahru 81310, Malaysia

3. Conservatoire National des Arts et Métiers, 292 Rue Saint Martin, Paris 75003, France

Received: Feburary 19, 2013 / Accepted: March 18, 2013 / Published: April 25, 2013

Abstract: All of the Landsat 7 data collected after 2003 contains missing pixels in the form of unsightly stripes across the images. To recover missing data of a Landsat image, different methods may be used. However, the gap filling process creates inconsistencies on pixel intensity values. The incongruous pixel numbers are anomolous observations and their classification in the reference specter is challenging. In an effort to contribute to this need, we propose a reliable robust approach to classify inconsistent pixels after the gap filling process. To estimate multivariate location-scale parameters a new robust DMVV (depth minimum vector variance estimator) is presented. The DMVV algorithm does not require any matrix inversion for its calculation, consequently its computational time is highly reduced. The results show that it has a high breakdown point and is very efficient for large data set. Landsat remote sensing data of Jakarta Province across years 2002 and 2010 are used as case study.

Key words: Depth function, minimum vector variance, covariance matrix, Mahalanobis distance.

1. Introduction

Landsat satellites data are frequently used to analyze land-use and land-cover changes, [1, 9-11]. A common approach of land cover change studies using Landsat data has been to use images to classify land into different categories, and to quantify changes in categories across different dates in time [4, 12, 16, 17, 19].

Since 2003, the, SLC (scan line corrector), 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 applied [22]. However, the gap filling process arises the problem of inconsistencies on pixel intensity values. The incongruous pixel numbers are anomolous observations and their classification in the reference specter is challenging because, it requires data processing methods, capable to classify inconsistent pixels and produce consistent land-cover monitoring. In an effort to contribute to this need, we have developed a reliable robust approach to classify discordant pixels after the gap filling process for land cover change studies using Landsat data.

The cornerstone of robust statistics is the robust estimation of multivariate location-scale parameters. Pioneering work in this area has been can be found in [5, 8, 15]. In statistical literature, we find several high breakdown estimators for multivariate mean and covariance matrix. A well known and largely used robust estimator is the MCD (minimum covariance determinant) [13]. Under regularity conditions, Hawkins in Ref. [6] proposed the FSA (feasible solution algorithm), which ensured an optimal solution for MCD. Afterwards, Rousseeuw and Van Driesen in Ref. [14] introduced an improved algorithm called the FMCD (fast minimum covariance determinant). The

Corresponding author: Dyah E. Herwindiati, Ph.D., associate professor, research fields: data mining. Email: herwindiati@untar.ac.id.

FMCD is a robust procedure with high breakdown point, but as indicated in Ref. [18], it might be inefficient for large data sets. To improve this aspect, Herwindiati et al. in Ref. [7] proposed the MVV (minimum vector variance), which is effective for huge data sets. MVV has the same breakdown point as FMCD but its computational aspects are by far advantageous.

This paper proposes a reliable MVV algorithm for robust supervised land-cover classification in remote sensing. To estimate multivariate location-scale parameters, a new robust depth function is presented. Its calculation does not require any covariance matrix inversion [2], which is a valuable asset when one deals with huge data sets. The supervised green space classification is done with a conventional two phase process: training sites and image cell classification. The sample areas for the training step are selected by human assessors. The outcomes of training sites are the spectral references of green space, i.e. the water catchment and vegetation areas. Then spectral reference values are used to classify the entire images from Landsat satellite. The area under investigation is Jakarta Province.

In order to make this paper self contained, in Section 2 we provide a background summary of remote sensing data and preprocessing for classification. In Section 3 we describe the methods and the algorithm used to classify land-cover into different categories. Then, monitoring results of green space areas of Jakarta Province across years 2002 and 2010 are presented in Section 4. The paper concludes with additional remarks and references.

2. Remote Sensing Materials

Landsat satellites have been providing multispectral images of Earth continuously since early 1970's. The purpose of the Landsat program is to provide world's scientists and application engineers with a continuing stream of remote sensing data for monitoring and managing earth's resources [20, 21]. A common approach using Landsat data has been to use images to classify land into differing categories, and to quantify changes in categories between different dates [12, 16, 19]. Land cover and land use changes are important indicators of human activities and climatic change. In Jakarta Province protected area public policies and their management by Governor's office are central to understand the recent land cover changes [24].

2.1 Study Area

The case of research is Jakarta multispectral imaging from Landsat 7 satellite. Jakarta is the capital of Indonesia that is spread over an area of around 700 km^2 with population up to 9.5 million in 2010. The supervised classification is done for change detection of Jakarta green space areas. The area under investigation is covered by coordinate 5° 19' 12" - 6° 23' 54"S latitude and 106° 22' 42" - 106° 58' 18"E longitude.

2.2 Data Sets

Tiff formatted images across years 2002 and 2010 are used as inputs. Data is captured by sensors having 7 bands involving the visible spectral, NIR, and MIR. The spatial resolution of bands n° 1-5, and n° 7 are 30 m^2, the resolution of the sixth band is 60 m^2.

On May 31, 2003, SLC of Landsat 7 ETM+ (Enhanced Thematic Mapper Plus), failed. Since that time all Landsat ETM+ images have wedge-shaped gaps. The impact of failure results in approximately 20% data loss. The gap filling is the preprocessing technique used to fill missing parts of remote sensed imagery. We do the gap filling procedure with the multi source. Fig. 1 reveals the Jakarta multispectral image with SLC in year 2010, and Fig. 2 shows the recovered image after the gap filling process.

3. Methods of Classification and Algorithm

3.1 Robust Minimum Vector Variance

Let $X_1, X_2, \ldots, X_n$ be a random sample from a *p*-variate distribution with location parameter μ and

Fig. 1 Multispectral Jakarta 2010 image with SLC.

Fig. 2 Multispectral Jakarta 2010 image after gap filling process.

covariance matrix Σ. Sample mean vector and sample covariance matrix are defined respectively by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad (1)$$

and

$$S = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^t \quad (2)$$

Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p \geq 0$ be eigenvalues of sample covariance matrix S. VV (The vector variance), of S is defined by

$$VV = \mathrm{Tr}\left(S^2\right) = \lambda_1^2 + \lambda_2^2 + \ldots + \lambda_p^2 \quad (3)$$

The advantage of VV consists in the fact that it measures multivariate dispersion even if the covariance matrix S is singular.

The MVV estimators of multivariate location-scale parameters are defined as the pair $\left(T_{MVV}, S_{MVV}\right)$ which minimize $\mathrm{Tr}\left(S_{MVV}^2\right)$ among all possible sample subsets H of size $h = \left[\frac{n+p+1}{2}\right]$, with

$$T_{MVV} = \frac{1}{h}\sum_{i\in H} X_i, \quad (4)$$

$$S_{MVV} = \frac{1}{h}\sum_{i\in H}\left(X_i - T_{MVV}\right)\left(X_i - T_{MVV}\right)^t \quad (5)$$

$$\mathrm{Tr}\left(S_{MVV}^2\right) = s_{11}^2 + s_{22}^2 + \ldots + s_{pp}^2 + 2\sum_{i=1}^{p}\sum_{j\neq i}^{p} s_{ij}^2 \quad (6)$$

MVV is an efficient robust estimator minimizing the square of parallelogram diagonal length. It was proposed in Ref. [7], in order to improve FMCD algorithm. By using Cholesky decomposition, we find that efficiency of MVV is of order $O\left(p^2\right)$ compared with FMCD which is of order $O\left(p^3\right)$ [7].

3.2 The Depth Function

We note d_i^2 the sample Mahalanobis distance define by

$$d_i^2 = \left(X_i - \bar{X}\right)^t S^{-1}\left(X_i - \bar{X}\right) \quad (7)$$

The sample version of Mahalanobis depth of X_i, noted MD_i is defines as;

$$MD_i = \frac{1}{1+\left(X_i - \bar{X}\right)^t S^{-1}\left(X_i - \bar{X}\right)} \quad (8)$$

from Eq. (7) and Eq. (8) we have the following equation.

$$MD_i = \frac{1}{1+d_i^2} \quad (9)$$

Part of MD_i denominator is Mahalanobis distance, which requires the inversion of sample covariance matrix S for its calculation. Aiming to reduce the complexity of FMCD and MVV algorithms, Djauhari and Umbara in Ref. [3], introduced a new depth function noted $\left|M_i\right|$ given by

$$\left|M_i\right| = \begin{vmatrix} 1 & \left(X_i - \bar{X}\right)^t \\ \left(X_i - \bar{X}\right) & S \end{vmatrix} \quad (10)$$

where M_i is a matrix of size $\left(p+1\right)\times\left(p+1\right)$ associated to sample $X_1, X_2, \ldots, X_n$. By using the property of partitioned matrix determinant, we have:

$$d_i^2 = 1 - \frac{|M_i|}{|S|} \quad (11)$$

From Eq. (10) and Eq. (11) we can write;

$$MD_i = \frac{|S|}{2|S| - |M_i|} \quad (12)$$

where $|S|$ and $|M_i|$ are respectively the determinants of S and M_i.

3.3 Algorithm

The supervised green space classification is done with a conventional two phase process: training sites and image cell classification. The sample areas for the training step are selected by human assessors. The outcomes of training sites are the spectral references of green space area, i.e. the water catchment and vegetation areas. To conduct the training process, DMVV estimator is proposed. DMVV is a modified version of MVV based on depth function given in Eq. (10). Its calculation does not require the inversion of covariance matrix, which is a valuable asset when one deals with large data sets. DMVV is a robust estimator that has the same breakdown point as MVV [7]. The algorithm to conduct the training phase is as follows:

Step (1): Collect images of the vegetation area in size $(a \times a)$ pixels based on red-green-blue multispectral visual and Normalized Difference Vegetation Index [23]. Let $\{X_1, X_2, ..., X_n\}$ be the training data set;

Step (2): Let $H_0 \subset \{X_1, X_2, ..., X_n\}$ such as card $\{H_0\} = h$ with $h = \left[\frac{n+p+1}{2}\right]$.

Step (3): Compute mean vector $\bar{X}_{H_O}$ and covariance matrix S_{H_O} of H_O.

Step (4): Compute

$$|M_i| = \begin{vmatrix} 1 & \left(X_i - \bar{X}_{H_O}\right)^t \\ \left(X_i - \bar{X}_{H_O}\right) & S_{H_O} \end{vmatrix}$$

for $i = 1, 2, ..., n$.

Step (5): Sort $\{|M_i| \,/\, i = 1, ..., n\}$ in decreasing order, $|M_{(1)}| \geq |M_{(2)}| \geq ... \geq |M_{(n)}|$.

Step (6): Define $H_W = \{X_{(1)}, X_{(2)}, ..., X_{(h)}\}$.

Step (7): From Eq. (1) and Eq. (2) calculate $\bar{X}_{H_W}$ and S_{H_W} respectively mean and covariance matrix of H_W.

Step (8): If $\mathrm{Tr}\left(S^2_{H_W}\right) = 0$ the process is stopped. Else, if $\mathrm{Tr}\left(S^2_{H_W}\right) \neq \mathrm{Tr}\left(S^2_{H_O}\right)$ repeat from Step 2 to Step7, until a stopping rule is satisfied: either according to number of iterations *k* or by the difference $\left|\mathrm{Tr}\left(S^2_{H_W,k}\right) - \mathrm{Tr}\left(S^2_{H_W,k+1}\right)\right| \leq \varepsilon$, where ε is a small constant.

Step (9): Let T_{VV} and S_{VV} be the location and covariance matrix calculated at Step 7. Based on T_{VV} and S_{VV} from Eq. (11) calculate robust squared distances $d^2_{VV,i}$ for *i* = 1, 2, …, *n*.

Step (10): Determine the range of each green space spectral area as $c_1 \leq d_{VV,i} \leq c_2$ where c_1 is the first quartile and c_2 is the third quartile of $d_{VV,i}$ for *i* = 1, 2, …, *n*.

Fig. 3 displays the scatter plot of $d_{VV,i}$ for green space spectral and Fig. 4 shows the scatter plot of $d_{VV,i}$ for green space reference spectral inside the interval $c_1 \leq d_{VV,i} \leq c_2$.

4. Results

4.1 Case Study

Tiff formatted images across years 2002 and 2010 are used as inputs. Data is captured by sensors having 7 bands involving the visible spectral, NIR, and MIR. The spatial resolution of 6 bands (n° 1-5, and n° 7) are 30 m^2, the resolution of the sixth band is 60 m^2.

The area under investigation is Jakarta Province covered by coordinate (5° 19' 12" - 6° 23' 54")S latitude and (106° 22' 42" - 106° 58' 18")E longitude. The classification step is done for Jakarta Province images by using the reference spectral from the the training step. Assume that $Y_1, Y_2, ..., Y_M$ are the pixels of whole Jakarta Province image. The distance $d^2_{vv,i}\left(Y_i, T_{VV}\right)$ $(i = 1, 2, ..., M)$ is calculate. Then each pixel is classified in one of three classes: water

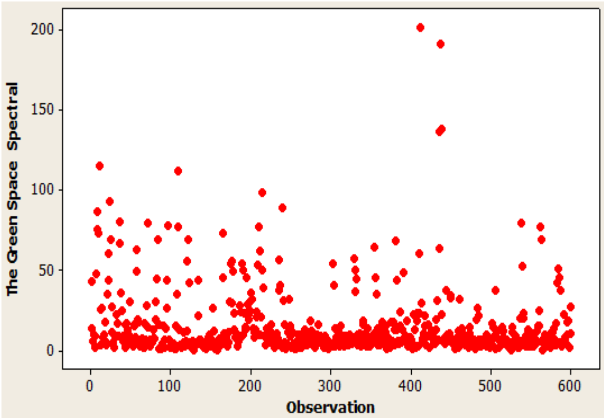

Fig. 3 Scatter plot of green space spectral.

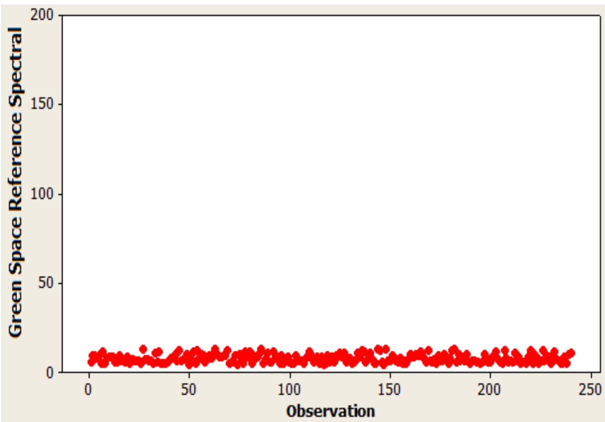

Fig. 4 Scatter plot of green space reference spectral.

catchment area, vegetation area and impervious area. The impervious area is defined as surface impenetrable by water including side walks, streets, highways, parking lots and rooftops. Observation Y_i is classified as impervious area if $d^2_{vv,i}(Y_i, T_{VV})$ is not in the interval $[c_1; c_2]$.

Figs. 5 and 6 display Jakarta pixels classification on the years 2002 and 2010, respectively. Vegetation area is labeled with green color, water catchment area is colored in yellow, and impervious area is presented with grey color.

On year 2002, the percentage of Jakarta green space was around 10.2569%. It was increased up to 11.24568% on year 2010. Table 1 shows percentages of green space areas on years 2002 and 2010.

Water catchment area on year 2010 is significantly greater than on 2002. The biggest increase has happened at Jakarta Halim Perdana Kusuma district. Fig. 7 shows land use changes. The blue color represents increased water catchment area and the red color represents decreasing one. Jakarta Halim Perdana Kusuma district is rounded by the white circle.

Table 1 Percentages of green space areas of Jakarta on years 2002 and 2010.

Year	Green space area		Total
	Water catchement area	Vegetation area	
2002	8.161%	2.096%	10.257%
2010	9.694%	1.552%	11.246%

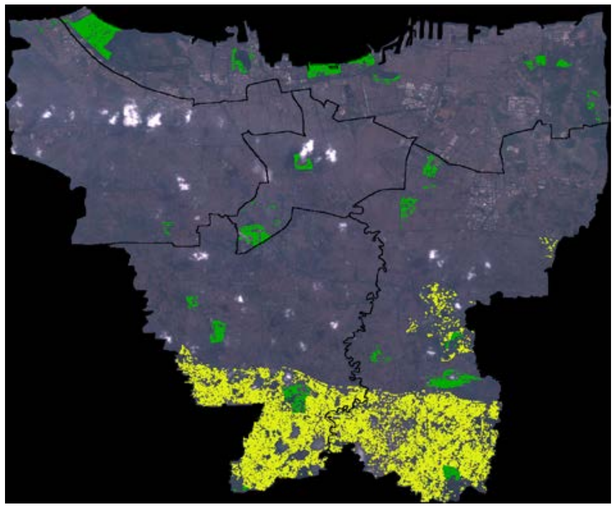

Fig. 5 Jakarta pixels classification on 2002. Vegetation area—green color; water catchment area—yellow color; impervious area—grey color.

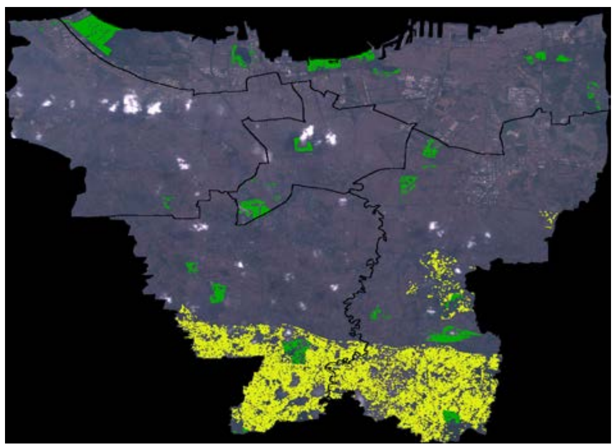

Fig. 6 Jakarta pixels classification on 2010. Vegetation area—green color; water catchment area—yellow color; impervious area—grey color.

4.2 Visualization of Area

Fig. 8 shows real visual of Jakarta Halim Perdana Kusuma borough after forestation and reforestation by Google Earth. The government of Special Capital

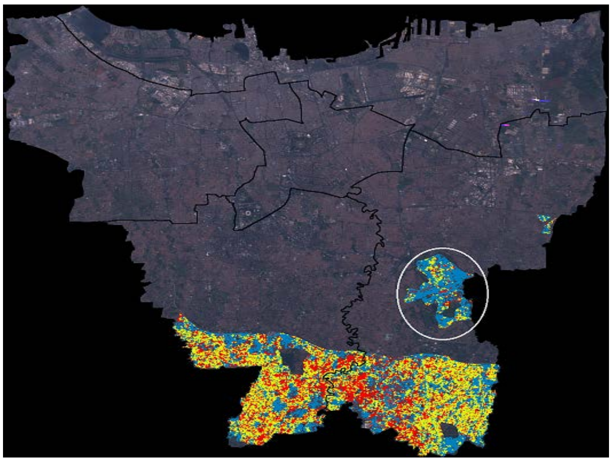

Fig. 7 Change vegetation area of Jakarta, during period 2002-2010: increased water catchment area—blue color; decreased water catchment area—red color.

Fig. 8 Jakarta Halim Perdana Kusuma district after forestation and reforestation by Google Earth on 2011.

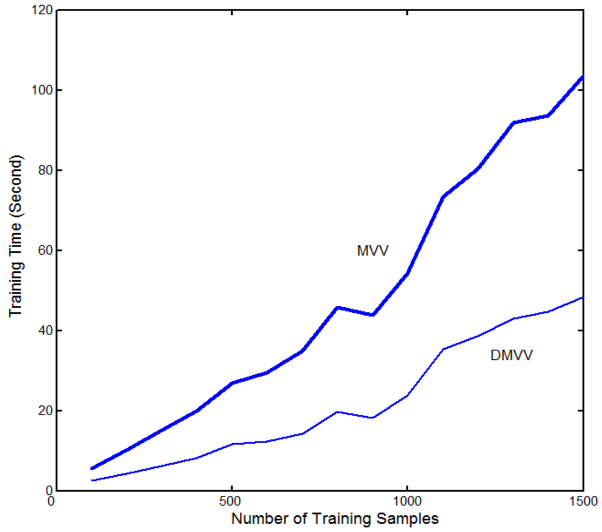

Fig. 9 Computation time in training process for MVV and DMVV estimators.

Region of Jakarta and its former Governors during the period from year 2000 till year 2008 made significant efforts to repair and develop Jakarta. The green land project budget was significantly increased and the Governor's office determined also special rules for the management and the implementation of protected area policies. For further information on special rules general program, we refer the reader to authentic official document ''The Special Rules Capital Regional District Jakarta Province Number 8 of 2007'', [24].

4.3 Comparisons of Computational Times in Training Process

The DMVV is an efficient estimator for classification of large remote sensing data. Fig. 9 shows graphical representation of times to estimate green space spectral reference in training phase for MVV and DMVV estimators. DMVV has significantly lower computation time than MVV. It is interesting to note that larger is the data set greater is the difference in calculation time between MVV and DMVV. Computations were operated by MATLAB 8.00 in an Intel® Core™ i7 CPU RAM 4.00 GB processor.

5. Remarks

The advantage of $|M_i|$ as a depth measure is that it does not require any matrix inversion in its computation. Its calculation only needs the computation of the determinant of a symmetric matrix. The modified minimum vector variance with depth function, DMVV is an efficient and effective robust estimator that should be considered for classification of large data sets. The empirical results provide strong evidence that DMVV is able to reduce significantly computational time in training step and it has a high breakdown point.

References

[1] H. Bagan, Y. Yamagata, Landsat analysis of urban growth: How Tokyo became the world's largest megacity during the last 40 years, Remote Sensing of Environment 127 (2012) 210-222.

[2] M.A. Djauhari, A robust estimation of location and scatter, Malaysia Journal of Mathematical Sciences 2 (1) (2008) 1-24.

[3] M.A. Djauhari, R.F. Umbara, A redefinition of mahalanobis depth function, Journal of Fundamental Sciences 3 (1) (2007) 150-157.

[4] S.N. Gillanders, N.C. Coops, M.A. Wulder, S. E. Gergel, Nelson, Multi-temporal remote sensing of landscape dynamics and pattern change: Describing natural and anthropogenic trends, Progress in Physical Geography 32 (2) (2008) 503—528.

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

[6] D.M. Hawkins, The feasible solution algorithm for the minimum covariance determinant estimator in multivariate data, Computational Statistics and Data Analysis 17 (1994) 197-210.

[7] D.E. Herwindiati, M.A. Djauhari, M. Mashuri, Robust Multivariate Outlier Labeling, J. Communication in Statistics—Simulation and Computation 36 (6) (2007) 1287-1294.

[8] P.J. Huber, Robust Statistics, Massachusetts, Wiley Series in Probability and Mathematical Statistics, John Wiley and Sons, New York, 1981.

[9] T. Lasanta, S. M. Vicente-Serrano, Complex land cover change processes in semiarid Mediterranean regions: An approach using Landsat images in northeast Spain, Remote Sensing of Environment 124 (9) (2012) 1-14.

[10] T.M. Lillesand, R.W. Kiefer, J.W. Chipman, Remote Sensing and Image Interpretation, Hoboken, John Wiley and Sons, New York, 2007.

[11] M. P. Lenney, C. E. Woodcock, J. B. Collins, H. Hamdi, The status of agricultural lands in Egypt: The use of multitemporal NDVI features derived from Landsat TM, Remote Sensing of Environment 56 (1996) 8-20.

[12] R. Romero-Calcerrada, G. L. W. Perry, The role of land abandonment in landscape dynamics in the SPA Encinares del río Alberche y Cofio, Central Spain, 1984-1999, Landscape and Urban Planning 66 (2004) 217—232.

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

[14] P.J. Rousseeuw, K. van Driessen, A fast algorithm for the minimum covariance determinant estimator, technometrics 41(1999) 212-223.

[15] P.J. Rousseeuw, A. M. Leroy, Robust Regression and Outlier Detection, John Wiley and Sons, New York, 1987.

[16] A. Shalaby, R. Tateishi, Remote sensing and GIS for mapping and monitoring land cover and land-use changes in the Northwestern coastal zone of Egypt, Applied Geography 27 (2007) 28—41.

[17] G. Shao, J. Wu, On the accuracy of landscape pattern analysis using remote sensing data, Landscape Ecology 23 (2008) 505—511.

[18] M. Werner, Identification of multivariate outliers in large data sets, Ph.D. Thesis, University of Colorado at Denver, 2003.

[19] F. Yuan, K. E. Sawaya, B. C. Loeffelholz, M. E. Bauer, Land cover classification and change analysis of the Twin Cities (Minnesota) metropolitan area by multitemporal Landsat remote sensing, Remote Sensing of Environment 98 (2005) 317—328.

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

[21] Natural Resources Canada, Fundamental of Remote Sensing, http://www.nrcan.gc.ca/sites/www.nrcan.gc.ca.../fundamentals_e.pdf

[22] USGS, Phase 2 gap-fill algorithm: SLC-off gap-filled products gap-fill algorithm Methodology, 2004, http://landsat.usgs.gov/documents/L7SLCGapFilledMethod.pdf.

[23] Normalized Difference Vegetation Index, NDVI http://earthobservatory.nasa.gov/Features/MeasuringVegetation/measuring_vegetation_2.php.

[24] The document ‘‘The Special Rules Capital Regional District Jakarta Province Number 8 of 2007”, might be found in the website of Jakarta Province: http://www.jakarta.go.id/web/produkhukum/download/346/PERDA_NO_8_TAHUN_2007_-_Tentang_Ketertiban_Umum.pdf.