Applied Mathematical Sciences, Vol. 7, 2013, no. 74, 3681 - 3694
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2013.34214

Comparative Analysis of Supervised and Unsupervised Classification on Multispectral Data

Asmala Ahmad

Department of Industrial Computing,
Faculty of Information and Communication Technology,
Universiti Teknikal Malaysia Melaka,
Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

Shaun Quegan

School of Mathematics and Statistics
University of Sheffield
Sheffield, United Kingdom

Abstract

The aim of this study is to compare two methods of image classification, i.e. ML (Maximum Likelihood), a supervised method, and ISODATA (Iterative Self-Organizing Data Analysis Technique), an unsupervised method. The former is knowledge-driven, while the latter is data-driven. The former needs a priori knowledge about the study area but the latter does not. In practice, the former can classify land covers with a higher accuracy and therefore is more widely used but there have been very few attempts to investigate this. Here we use both methods in our study area, Selangor, Malaysia and compare the outcomes by means of qualitative and quantitative analyses to have a better understanding of the underlying reasons that drive the performance of both methods.

Keywords: ML, ISODATA, Land Cover, Landsat

1. Previous Studies

In order to classify land covers, many researchers have showed greater interest in supervised rather than unsupervised methods [1], [3], [7]. Nevertheless, only a few researchers attempted to compare the performance of ML classification with other methods. They are such as Thompson et al. [8] and Low and Choi [6]. Thompson et al. [8] compared ML classification and ISODATA clustering methods for coasts and river corridors along the East coast of England from Compact Airborne Spectrographic Imager (CASI). Results are presented as classification maps, confusion matrices and feature space images. They showed that ML classification could produce excellent results in separating inland cover types while ISODATA clustering was considered as an acceptable alternative due to involving less user input and not dependence on a priori information on the study area. Low and Choi [6] performed a hybrid classification for landuse/cover mapping by using Landsat 7 ETM+ data over the Atlanta metropolitan area, the largest city of the state of Georgia, USA. The land use/cover classes within the study area are urban/industry, settlement, cleared land, crop land, forest and water. In their approach, ISODATA clustering was initially used, followed by a supervised fuzzy classification. The hybrid classification was compared with ISODATA clustering, ML classification and supervised fuzzy classification. The hybrid classification was found to be slightly better in classification accuracy than the ISODATA clustering, but the ML and supervised fuzzy classification produced much lower accuracies. Nevertheless, in these studies, no in-depth analysis of the methods was reported. Therefore, this study attempted to carry out land cover classification using ML and ISODATA methods and compare their performance qualitatively and quantitatively.

2. Methodology

Initially, ML classification and ISODATA clustering were applied to our study area which is in Klang, a district located in Selangor, Malaysia, which covers approximately 630 km^2 within longitude 101° 10’ E to 101°30’ E and latitude 2°99’ N to 3°15’ N. The satellite data come from band 1 (0.45 – 0.52 μm), band 2 (0.52 – 0.60 μm), band 3 (0.63 – 0.69 μm), band 4 (0.76 – 0.90 μm), band 5 (1.55 – 1.75 μm) and band 7 (2.08 – 2.35 μm) of Landsat-5 TM dated 11th February 1999, while the supporting data is a reference map from October 1991 of the study area produced by the Malaysian Centre for Remote Sensing using ground survey and high resolution satellite data. Prior to any data processing, masking of cloud and its shadow were carried out based on threshold approach [4]. The reference map (Figure 1(a)) together with Landsat bands 3, 4 and 5 data were used to identify major land cover classes. Assisted by the knowledge of the study area, these classes were water, coastal swamp forest, dryland forest, oil palm, rubber, industry, cleared land, urban, sediment plumes, coconut and bare land (Figure

1(b)). Coastal swamp forest covers most of Klang Island (i.e. at the south-west of the image) and coastal regions in the south-west of the scene. Most of the dryland forest can be recognised as a large straight-edged region in the north-east. Oil palm is the most important commercial crop and can be found in the centre towards the north-west, while rubber is unevenly distributed in the north and south-east of the scene. Oil palm plantations, mostly managed by a government agency called FELDA (Federal Land Development Authority, Malaysia) are far more abundant than rubber plantations due to higher demand and a better price in the global markets [5]. Urban areas fill the lower middle of the scene, from the coastal region and inland. Industry can be recognised as brighter patches near the urban areas, especially in the southwest and northeast. The relatively large urban and industry areas reflect the fact that Klang town and Klang port play an important role in stimulating the surrounding areas economically. Cleared land is spread all over the scene and indicated by line-like shapes and patches of no particular shape.

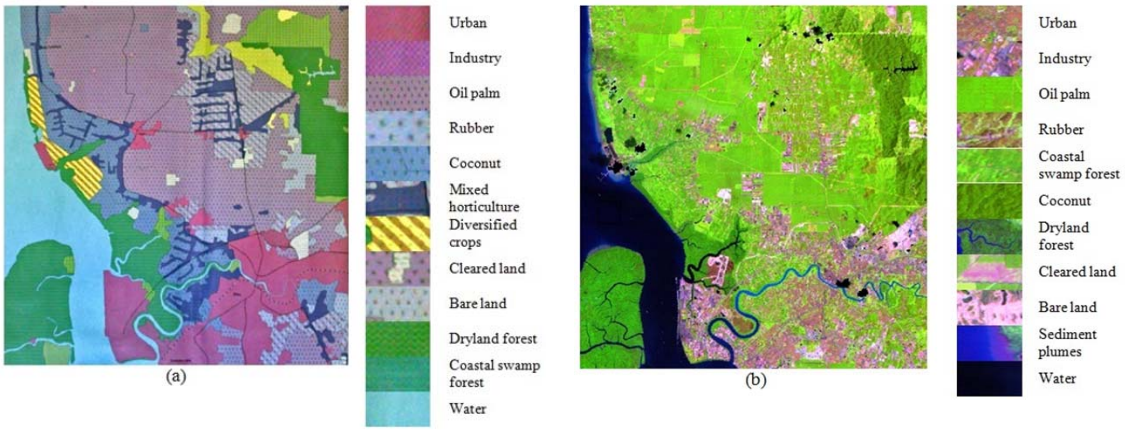


Fig. 1. The study area from (a) the reference map and (b) the Landsat-5 TM data with bands 5 4 and 3 assigned to the red, green and blue channels, with cloud and its shadow masked in black.

3. 2.1 ML Classification

In ML classification, regions of interest (ROIs) associated with the training pixels for 11 classes of land cover were determined based on the reference map. Training areas were established by choosing one or more polygons for each class. In order to select a good training area for a class, the important properties taken into consideration are its uniformity and how well they represent the same class throughout the whole image. Class separability of the chosen training pixels were determined by means of the JM distance [1]. For each class, these training pixels provide values from which to estimate the means and covariances of the spectral bands used. Accuracy assessment of the ML classification is determined by means of the confusion matrix (sometimes called error matrix), which compares, on a class-by-class basis, the relationship between reference data (ground truth) and the corresponding results of a classification [11]. The diagonal elements in a

confusion matrix represent the percentage of correctly assigned pixels and are also known as the producer accuracy. Producer accuracy is a measure of the accuracy of a particular classification scheme and shows the percentage of a particular ground class that is correctly classified. It is calculated by dividing each of the diagonal elements in a confusion matrix by the total of each column respectively. The minimum acceptable accuracy for a class is 90% [10].User Accuracy is a measure of how well the classification is performed. It indicates the percentage of probability that the class which a pixel is classified to on an image actually represents that class on the ground [10]. It is calculated by dividing each of the diagonal elements in a confusion matrix by the total of the row in which it occurs. A measure of overall behaviour of the ML classification can be determined by the overall accuracy, which is the total percentage of pixels correctly classified. The minimum acceptable overall accuracy is 85% [9]. The Kappa coefficient, κ is a second measure of classification accuracy which incorporates the off-diagonal elements as well as the diagonal terms to give a more robust assessment of accuracy than overall accuracy. The ML classification yielded an overall accuracy of 97.4% and kappa coefficient 0.97, indicating very high agreement with the ground truth.

2.2 ISODATA Clustering

For ISODATA, in order to determine the clustering that best matches the actual land cover, the ENVI ISODATA programming module was run several times with different numbers of clusters. After the clustering process ended, the clusters were manually labelled to the nearest match, based on the reference image. ISODATA clustering generates a cluster map with clusters assigned to arbitrary colours that need to be labelled according to the land cover class. In the labelling process, each cluster is matched to a class (or classes) from the reference image and given a specific colour so that at the end of the labelling process, classes (i.e. single or multiple) that exist in the cluster map can be easily recognised by their colours [2]. Finally, it was found that the 8 classes that can be classified by ISODATA were water, coastal swamp forest, dryland forest, oil palm, cleared land, bare land, urban and sediment plumes. Accuracy assessment of the cluster map by means of a confusion matrix, yielded an overall accuracy of 93.1%, with kappa coefficient 0.91, indicating quite good agreement with the ground truth pixels.

4. 3. Comparison of ML classification and ISODATA clustering

The results of both methods were compared in terms of visual, mean, standard deviation, classification accuracy, band correlation and decision boundary analysis.

5. 3.1 Visual Analysis

From Figure (b), it can be seen that ISODATA generates only eight classes, i.e. urban, industry, oil palm, dryland forest, coastal swamp forest, cleared land, water and sediment plumes, while ML is able to produce three more additional classes, i.e. coconut, rubber and bare land. Hence, some clusters in ISODATA consist of more than one class, e.g., some pixels from the oil palm cluster belong to rubber and coconut, while some pixels from the industry cluster belong to bare land. ISODATA has a more extensive oil palm area than ML because it contains pixels that belong to sediment plumes, dryland forest and cleared land classes. Similarly, the urban area in ISODATA is larger than that of ML, mainly because it comprises quite a large number of pixels from the cleared land and industry classes (Table 1). In overall, ML classifies most of the classes that exist in the study area with a good qualitative agreement with the reference map. This is due to the fact that ML is very much influenced by the use of training pixels for predefined classes, which are based on the reference map and user's knowledge. On the other hand, ISODATA performs the clustering task automatically, depending only on the statistical properties of the data per se.

Table 1: Classes determined by ML and ISODATA and their percentage area.

Class	Area (%)	
	ML	ISODATA
Cleared land	22.3	17.4
Urban	11.1	16.6
Oil palm	23.8	35.6
Water	11.5	11.7
Coastal swamp forest	6.5	8.0
Industry	13.8	2.0
Dryland forest	5.6	6.6
Sediment plumes	3.6	2.0

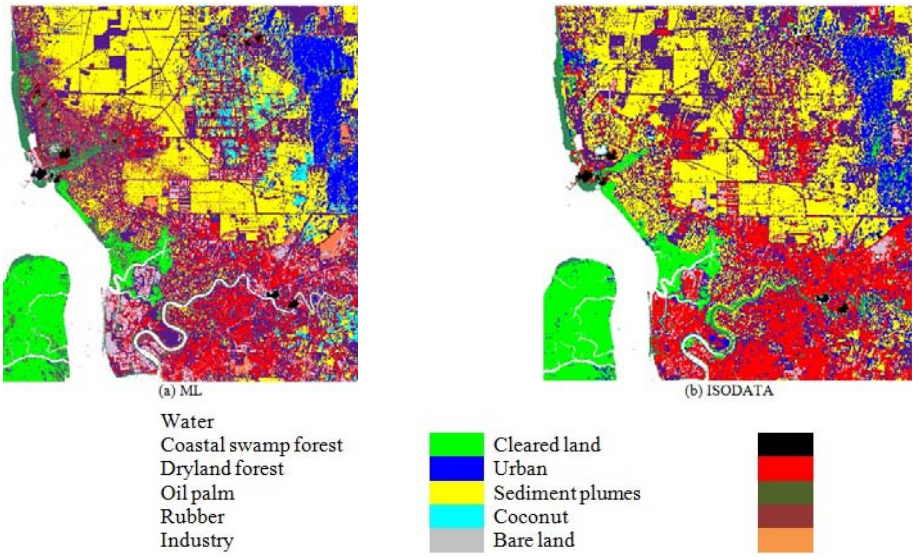


Fig. 1. Land cover classification using (a) ML and (b) ISODATA.

6. 3.2 Accuracy Analysis

Table 2 shows the producer and user accuracy for the classes generated using ISODATA and ML. In terms of individual classes, in descending order, the producer accuracy difference (ML – ISODATA) of the classes are cleared land (62%), sediment plumes (43%), industry (9%), water (8%), urban (4%), dryland forest (1%), coastal swamp forest (-0.06%) and oil palm (-5%) (Figure 3). ML is higher than ISODATA in the first six classes, with significant differences in cleared land and sediment plumes class (> 40%), while ISODATA has higher accuracy than ML in the last two classes, with relatively small differences (≤ 5%). Overall, it is clear that ML is better than ISODATA in terms of producer accuracy.

Table 2: Producer and user accuracy for the classes generated using ISODATA and ML.

Class	Producer Accuracy (%)		User Accuracy(%)	
	ISODATA	ML	ISODATA	ML
Coastal swamp forest (CSF)	99.80	99.74	98.59	99.99
Dryland forest (DLF)	98.09	99.25	96.91	99.93
Oil palm (OP)	97.57	92.36	93.75	99.64
Cleared land (CL)	31.84	93.84	45.23	82.90
Sediment plumes (SP)	52.68	95.91	56.82	96.78
Water (W)	91.83	99.89	99.74	100.00
Urban (U)	89.00	93.29	75.75	99.31
Industry (I)	90.29	99.71	64.23	82.90

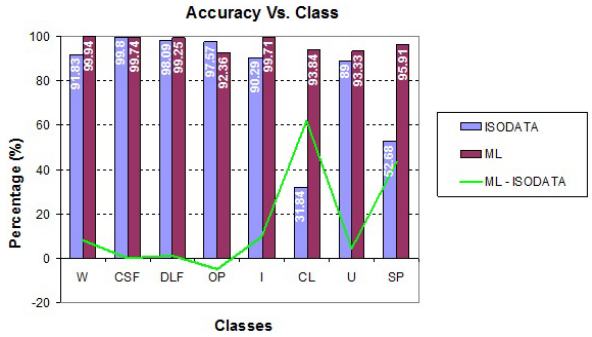


Fig. 2. The accuracy for individual classes of ML and ISODATA and the difference between them (ML – ISODATA).

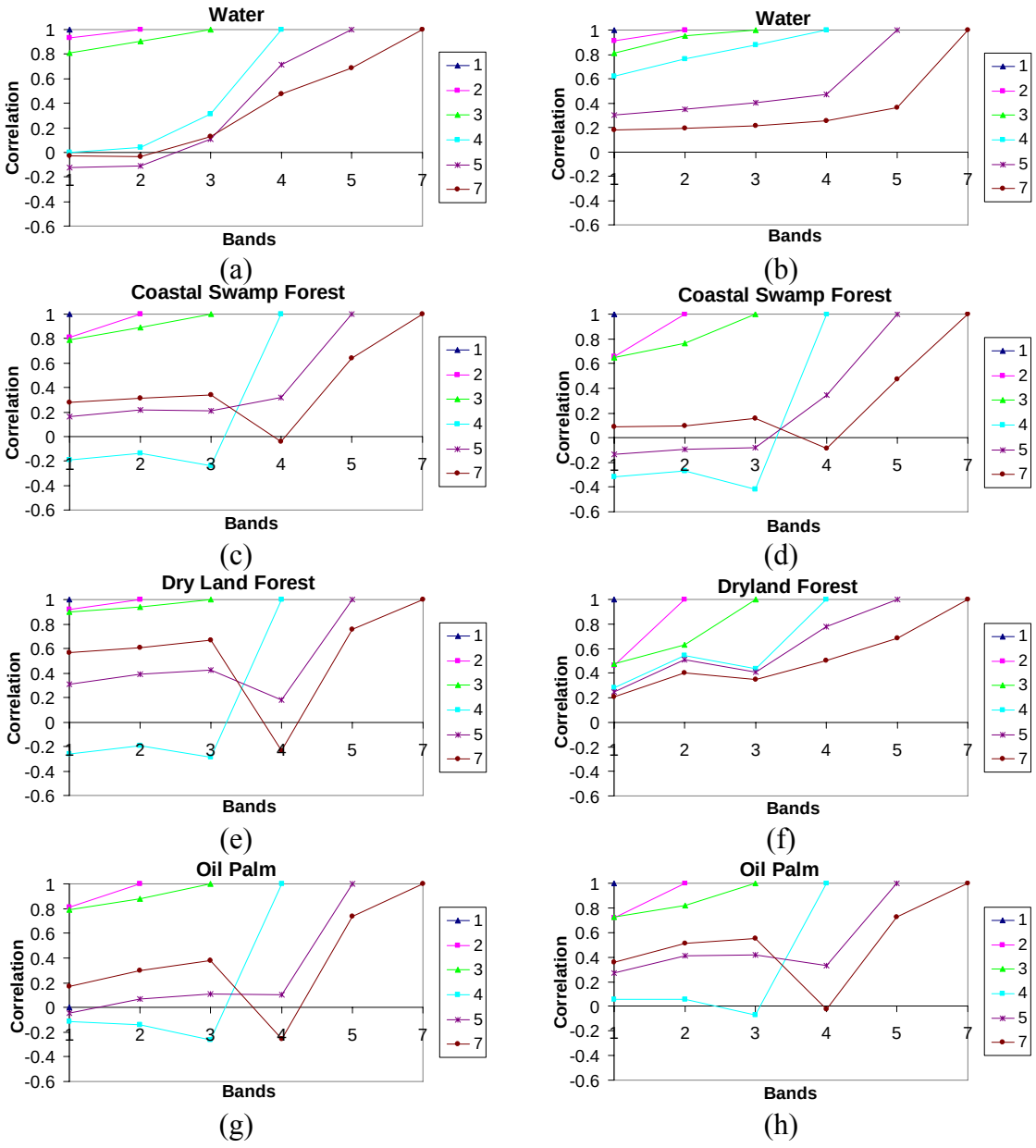
7. 3.3 Correlation Matrix Analysis

Classification uses the covariance of the bands. Nonetheless, covariance is not intuitive; more intuitive is correlation, $\rho_{k,l}$, i.e. covariance normalised by the product of the standard deviations of bands, k and l:

$$\rho_{k,l} = \frac{C_{k,l}}{\sigma_k \sigma_l} = \frac{E\left(\left(I_k - \mu_k\right)\left(I_l - \mu_l\right)\right)}{\sigma_k \sigma_l} \tag{0}$$

where $C_{k,l}$ is the covariance between bands k and l, σ_k and σ_l are the standard deviations of the measurements in bands k and l respectively, E is the expected value operator, and I_k and I_l and μ_k and μ_l are the intensities and means of bands k and l respectively. When using more than two bands, it is convenient to use a correlation matrix, where the element in row m and column n that correspond to band k and l is given by $\rho_{k,l}$. If $m = n$, then $\rho_{k,l} = 1$, so this will be the value of the diagonal elements of the matrix. Otherwise, if $m \neq n$, $\rho_{k,l}$ lies between -1 and 1. In order to analyse the correlation matrices, plots of correlation versus band pair for all classes from ISODATA and ML are shown in Figure 4. Each coloured curve represents correlation between a specific band (given by a specific colour) with all bands (on the x-axis). Landsat bands 1, 2 and 3 are located within a very close wavelength range of the visible spectrum, with their centre wavelengths differing only by about 0.1 μm. Measurements made from these bands normally exhibit similar responses and therefore are highly correlated. Poor correlations may result from mixed pixel problem (existence of more than one class in a pixel). Correlations between lower-numbered bands (i.e bands 1, 2 and 3) and higher-numbered bands (i.e. bands 4, 5, and 6) are much lower because involving non-adjacency wavelengths. This is because same classes may be measured differently from bands having wavelength regions far apart (i.e. visible and reflected infrared region). From Figure 4, for cleared land and sediment plumes, correlation in most band pairs is higher in ML than ISODATA, especially for bands 1, 2 and 3, which corresponds to the higher accuracy in these classes in ML than ISODATA. For certain classes, such as water (with very low reflectances), the superiority of ML over ISODATA is even clearer, as shown not only by the correlations from bands 1, 2 and 3, but also 4, 5 and 7 in ML that have higher correlations compared to ISODATA (Figure 4(a and b)). This is because the training pixels for water can be easily and precisely located compared to other land classes, therefore leading to a higher accuracy of water in ML compared to ISODATA. A high correlation is shown by industry (with very high reflectances) (Figure 4(m and n)) due to the strong relationships of variation between the brightness of pixels and mean brightness in all bands (1, 2, 3, 4, 5 and 7). These bands comprises of visible and reflected infrared regions that sense the strong solar reflectance from industry in a similar way.

ISODATA ML


Water
Correlation
Bands
(a)
Water
Correlation
Bands
(b)
Coastal Swamp Forest
Correlation
Bands
(c)
Coastal Swamp Forest
Correlation
Bands
(d)
Dry Land Forest
Correlation
Bands
(e)
Dryland Forest
Correlation
Bands
(f)
Oil Palm
Correlation
Bands
(g)
Oil Palm
Correlation
Bands
(h)
1
2
3
4
5
7

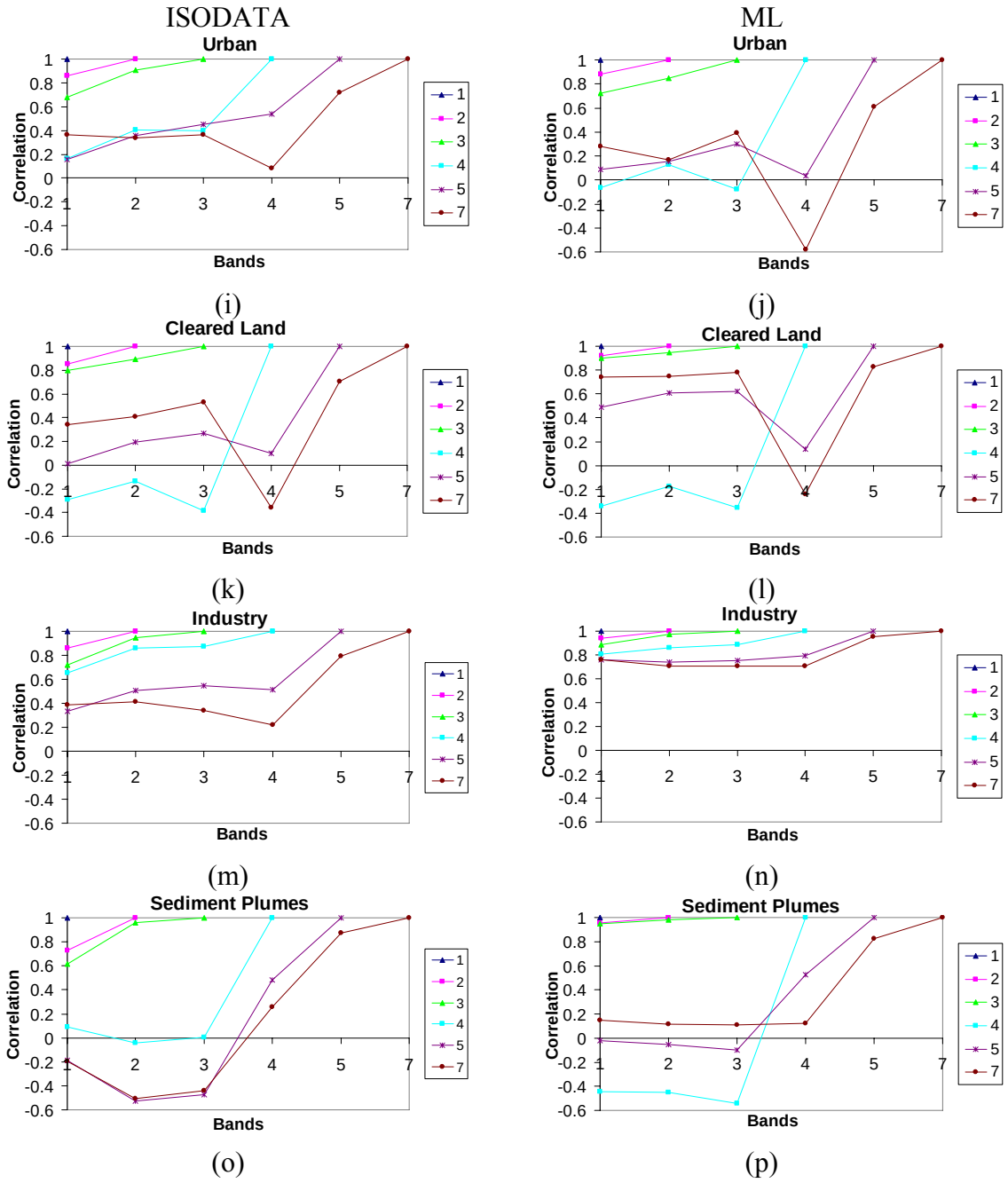
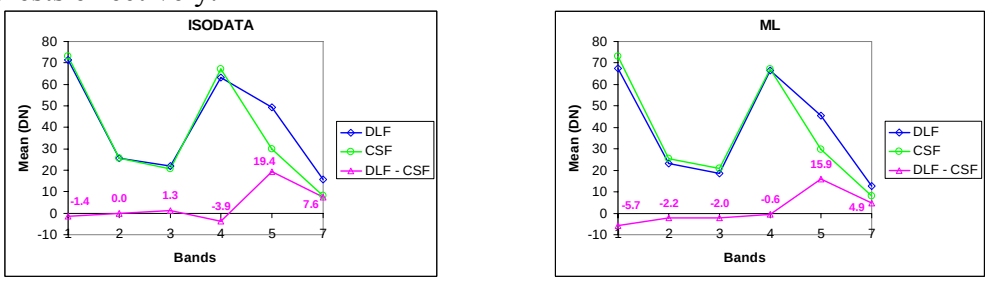


Fig. 1. Correlations between band pairs from ISODATA and ML.

8. 3.4 Mean and Standard Deviation Analysis

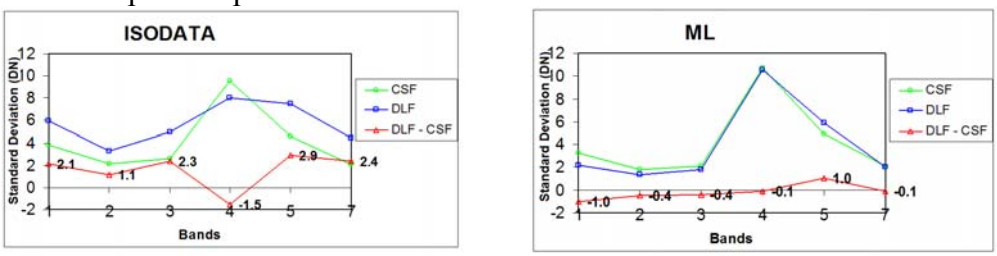
Here, we focus on the forest classes (i.e. dryland and coastal swamp forest) in order to analyse further ML and ISODATA. Despite of being very similar, both forests can still be separated quite effectively from each other using ML and ISODATA, as revealed in the previous analyses. Figure 5 shows the means of coastal swamp forest and dryland forest classes in ISODATA and ML, which are almost the same particularly in bands 1, 2 and 3. The higher difference (DLF-CSF) in band 5 and 7 indicates that these bands are essential for separating the forests effectively.



(a) (b)

Fig. 2. Means of coastal swamp forest (CSF) and dryland forest (DLF) classes in (a) ISODATA and (b) ML.

In term of standard deviation, both methods exhibit quite a different trend for both forests (Figure 6). It can be seen that for ISODATA, the standard deviation for dryland forest is bigger than coastal swamp forest in most of the bands except band 4. For ML however, the standard deviation of coastal swamp forest is bigger than dryland forest in most of the bands, except band 5. It is likely that the higher standard deviations caused by the present of incorrectly classified pixels in the coastal swamp forest class for ML and dryland forest class for ISODATA that can be associated with the lower producer accuracy of these classes (see Table 2). Apart from that, the range of the standard deviation difference is bigger in ISODATA (-1.5 to 2.9) than in ML (-0.4 to 1.0). This indicates the present of the incorrect dryland forest pixels in more severe in ISODATA than the incorrect coastal swamp forest pixels in ML.



(a) (b)

Fig. 3. Standard deviations of the coastal swamp forest and dryland forest classes in (a) ISODATA and (b) ML.

9. 3.5 Decision Boundary Analysis

We investigate further ISODATA and ML in terms of decision boundary. By assuming a given class i obeys a multivariate Gaussian distribution, the discriminant function can be expressed as:

$$g_i(\boldsymbol{\omega}) = \ln P(\boldsymbol{\omega} \mid i) = -\frac{1}{2}(\boldsymbol{\omega} - \boldsymbol{\mu}_i)^t C_i^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}_i) - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln(|C_i|) \quad (2)$$

Where $\boldsymbol{\omega}$ is feature vector, $\boldsymbol{\mu}_i$ is the mean for class i, N is the number of bands and $P(\boldsymbol{\omega} \mid i)$ is the likelihood function. The class regions are separated by decision boundaries, where, the decision boundary between class i and j occurs when:

$$g_i(\boldsymbol{\omega}) = g_j(\boldsymbol{\omega}) \quad (3)$$

By making use of Equation (2), this becomes:

$$-(\boldsymbol{\omega} - \boldsymbol{\mu}_i)^t C_1^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}_i) - \ln(|C_i|) + (\boldsymbol{\omega} - \boldsymbol{\mu}_j)^t C_j^{-1}(\boldsymbol{\omega} - \boldsymbol{\mu}_j) + \ln(|C_j|) = 0 \quad (4)$$

15 sets of decision boundaries are then generated using Equation (4) for all band pairs. Eight of them are are shown in Figure 7; ‘M1’ and ‘M2’ are the means for dryland forest and coastal swamp forest respectively, ‘Band k Vs. Band l’ denotes that the vertical axis is band k while horizontal axis is band l and ‘CSF’ and ‘DLF’ indicate coastal swamp forest and dryland forest respectively, i.e. to which class the boundary belongs to. However, due to the inconvinence of the boundary shape (or shapes) and to avoid confusion, the sign is not shown for pairs involving band 4. The decision boundaries formed by both methods have the form of conic sections. For ISODATA, pairs 2:1, 3:1 and 3:2 form an elliptic curve, while pairs 5:1, 7:1, 5:2, 7:2, 5:3, 7:3, 7:5, 5:4 and 7:4 are parabolic and pairs 4:1, 4:2 and 4:3 are hyperbolic, whereas for ML, pairs 2:1, 3:1, 7:1, 3:2 and 7:2 form an elliptic curve, while pairs 5:1, 5:2, 5:3, 7:3 and 7:5 form a parabolic curve and pairs 4:1, 4:2, 4:3, 5:4 and 7:4 form a hyperbolic curve. For ISODATA, most of the boundary is owned by coastal swamp forest due to the smaller standard deviation of coastal swamp forest than dryland forest most of the bands, while vice versa for ML. For most of the pairs that form elliptical boundary, the boundary size in ISODATA is smaller than ML, due to the bigger range of the standard deviation difference in ISODATA compared to ML. For ISODATA, M1 and M2 being located within the same boundary for pairs 2:1, 4:1, 3:2, 4:2 and 4:3, due to the very small differences between the means, particularly in bands 1, 2 and 3. For ML, in most bands (except band 4), the difference between the means is big

enough that M1 and M2 are located in the different side of the boundary. Hence, ML can separate between the forests better than ISODATA, due to its ability in positioning the means in the different side of the decision boundary. Nonetheless, in spite of the use of statistical means, ISODATA can discriminate between the forests quite efficiently, in which is evident from the shape and size of the corresponding decision boundaries.

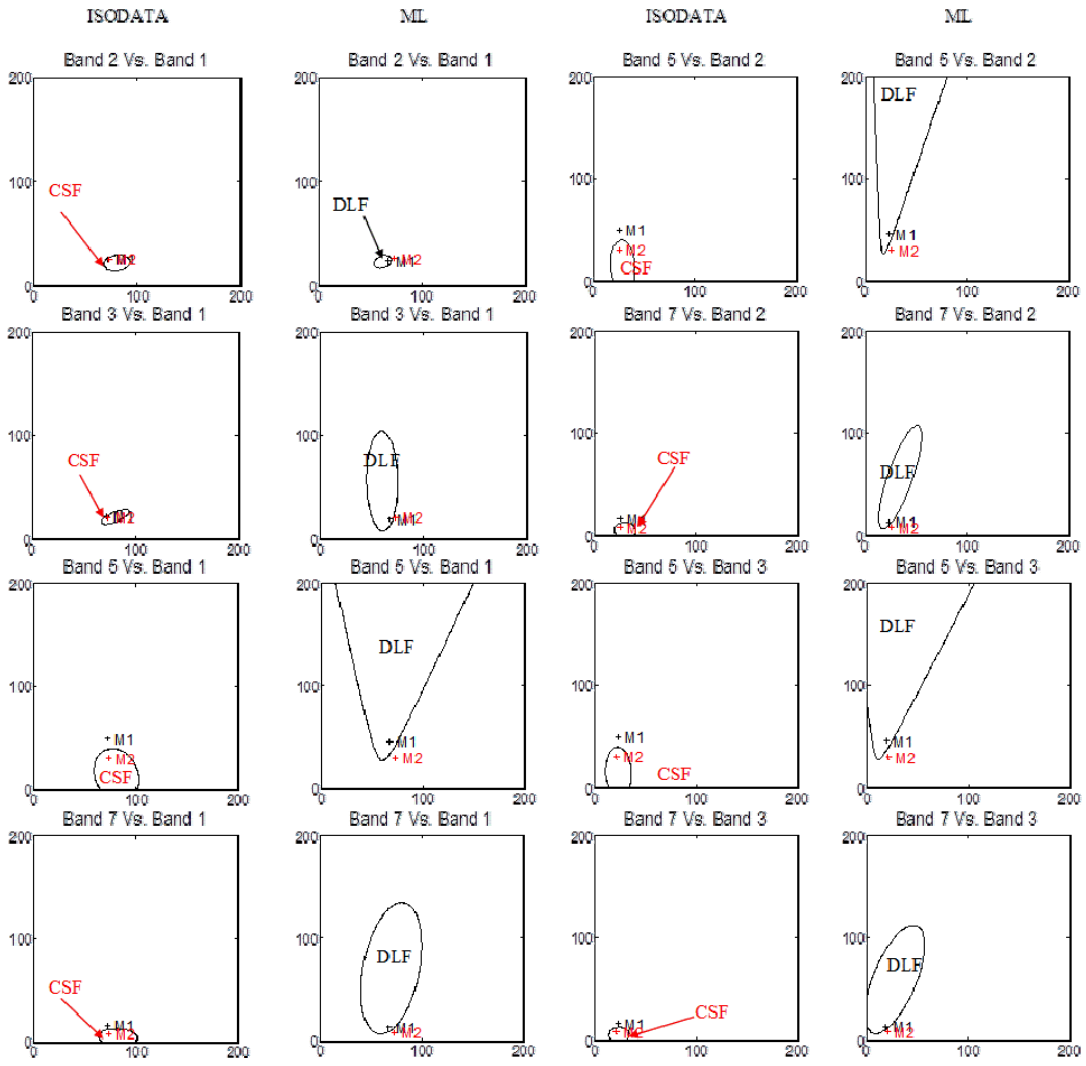


Fig. 1. Decision boundaries between coastal swamp forest and dryland forest for ISODATA clustering and ML classification. 'M1' and 'M2' are the means for dryland forest and coastal swamp forest respectively. 'Band k Vs. Band l' denotes that the vertical axis is band k while horizontal axis is band l. The boundary is owned either by coastal swamp forest ('CSF') or dryland forest ('DLF'), but this is not shown for pairs involving band 4.

4. Conclusions

In this study, a comparative analysis of ISODATA clustering and ML classification on multispectral Landsat satellite data has been carried out. ML classified the study area into 11 classes, which was chosen earlier, with accuracy 97% ($\kappa = 0.97$), while only eight can be clustered by ISODATA with accuracy 93% ($\kappa = 0.91$). Three classes can be classified by ML but not by ISODATA viz. coconut, rubber and bare land. ML classifies pixels based on known properties of each cover type, but the generated classes may not be statistically separable. ISODATA makes use of a natural grouping of the pixels to produce clusters that are statistically separable, but they may not be spectrally separable. ISODATA clustering is fast and straightforward but still able to separate quite well classes that are spectrally similar, e.g. coastal swamp forest and dryland forest. The band correlation of classes with high reflectance, e.g. industry, is higher for all band pairs in ML than for ISODATA because of the strong relationships of variation between the brightness of pixels and mean brightness in all bands. In decision boundary analysis, the separation between mean of the classes is better in ML compared to ISODATA; this is one of the main factors that leads to the higher classification accuracy in ML compared to ISODATA.

References

[1] A. Ahmad, Analysis of maximum likelihood classification technique on Landsat 5 TM satellite data of tropical land covers, *Proceedings of 2012 IEEE International Conference on Control System, Computing and Engineering (ICCSCE2012)*, (2012), 1 – 6.

[2] A. Ahmad, Analysis of Landsat 5 TM data of Malaysian land covers using ISODATA clustering technique, *Proceedings of the 2012 IEEE Asia-Pacific Conference on Applied Electromagnetic (APACE 2012)*, (2012) , 92 – 97.

[3] A. Ahmad, and S. Quegan, Analysis of maximum likelihood classification on multispectral data. *Applied Mathematical Sciences*, **6** (2012), 6425 – 6436.

[4] A. Ahmad, and S. Quegan, Cloud masking for remotely sensed data using spectral and principal components analysis, *Engineering, Technology & Applied Science Research (ETASR)*, **2** (2012), 221 – 225.

[5] A. Simeh and T.M.A.T. Ahmad, The case study on the Malaysian palm oil, *Regional Workshop On Commodity Export Diversification And Poverty Reduction In South And South-East Asia(Bangkok, 3-5 April, 2001),* UNCTAD, Bangkok, 2001.

[6] C.P. Low and J. Choi, A hybrid approach to urban land use/cover mapping using Landsat 7 Enhanced Thematic Mapper Plus (ETM+) images, *International Journal of Remote Sensing*, **25** (2004), 2687 – 2700.

[7] G.M. Foody, Fully fuzzy supervised classification of land cover from remotely sensed imagery with an artificial neural network, *Neural Computation & Applications*, **5 (**1997), 238 – 247.

[8] G. Thomson, R.M. Fuller and J.A. Eastwood, Supervised versus unsupervised methods for classification of coasts and river corridors from airborne remote sensing. *International Journal of Remote Sensing*, **19** (1998), 3423 – 3431.

[9] J.R. Thomlinson, P.V. Bolstad and W.B. Cohen, Coordinating methodologies for scaling landcover classifications from site-specific to global: steps toward validating global map products, *Remote Sensing of Environment*, **70** (1999), 16 – 28.

[10] M. Story and R. Congalton, Accuracy assessment: a user's perspective, *Photogrammetric Engineering and Remote Sensing,* **52** (1986), 397 – 399.

[11] T.M. Lillesand, R.W. Kiefer and J.W. Chipman, *Remote Sensing and Image Interpretation,* John Wiley & Sons, Hoboken, NJ, USA, 2004.

Received: April 12, 2013