IDENTIFYING THE NUMBER OF CLUSTERS AND OBTAINING OPTIMAL CLASSIFICATION RESULT

H. Zhu

Clark Labs, Clark University, 950 Main Street, Worcester, MA 01610 – hzhu@clarku.edu

WG VII/4

KEY WORDS: Remote sensing, Clustering, Classification, Cluster number identification, Classification optimization.

ABSTRACT:

Classification methods are often employed to derive land cover information from satellite images. Although a variety of classifiers have been developed, some primary issues remain to be further investigated. Among others, two of them are: a) the determination of the number of distinct clusters, and b) obtaining optimal classification result utilizing the available classifiers. To investigate the first issue, a skewness measurement and a separation-cohesion index (SCI) are used to describe some characteristics of a clustering result. By plotting the two indices against the number of clusters, kinks and slope changes may emerge in the curves. With consideration of the spatial resolution of the imagery and the context of an application, an optimal number of clusters can then be determined. For most cases, the clusters in remote sensing images are different from the land cover types of interest. We need either to merge some clusters in order to form a land cover type or to split a cluster into more than one land cover types of interest. Because the statistical model of a land cover type does not always follow a distribution pattern, it may contain multiple models, or has no noticeable patterns, a collection of classifiers are proposed to accommodate any scenarios. This study used three classifiers: the maximum likelihood method for parametric model based approaches, the Kohonen's self-organizing map (SOM) for neural networks, and a classification tree method. Through case studies, practical procedures are proposed: 1) identify the number of clusters within an application context; 2) associate clusters with land cover types; 3) classify images using the three classifiers, and assess their accuracies; 4) accept the result of classes from any one of the three classifiers, and process the remaining classes in the next iteration, and each class is then analyzed independently. The proposed procedures were shown to be efficient using case studies with three imagery data sets..

1. INTRODUCTION

Land cover and land use change monitoring and modeling are of importance for natural resource management. The trend of land cover change and rate of deforestation are important factors that may affect global climate and sustainable development. Satellite remote sensing has been widely used for collecting images for deriving land cover information. Image classification methods are often employed to derive such information from currently collected and archived satellite images. Although there are a variety of commercially available software packages with a spectrum of classifiers, including supervised and unsupervised, parametric and non-parametric, classification trees and neural networks, there are still certain issues to be further investigated. Among others, two of the issues are the determination of the number of distinct clusters in remotely sensed data and the proper classifiers for obtaining optimal classification results. There have been research efforts that focus on certain aspects of these issues (Fraley and A. E. Raftery, 1998; Li and Eastman, 2006; Zambon et al., 2006; and Zhu and Zhu 2007).

Before classifying images into classes, we often need to determine how many classes we want in a given application context and how many classes there are in the images. In satellite images, similar land cover types exhibit similar reflectance and form similar patterns. If we plot their distributions, using each measurement (an image band) as a dimension, each land cover type tends to form a cluster in the data space. Given the fact that each land cover type contains a large number of pixels, the central limit theorem can then be applied. That is to say, the distribution of pixel values in a multiple band feature space for each cluster is likely to obey the normal distribution. This forms the theoretical basis for identifying the number of distinct clusters using the skewness measurement.

The second component of this study focuses on how to achieve optimal classification results from some available classifiers. Classifier development has been an active research topic and many books have been dedicated for it (Tso and Mather, 2001; Richards and Jia, 2006; and Bishop, 2006). For most cases, the clusters in remote sensing images are different from the land cover types of interest. We need either to merge some clusters to form a land cover type or to split a cluster into different land cover types. In reality, due to the constraints from the sensor technology, mixed pixels effect, and spectral separability concerns, some clusters may obey a statistical model, and others do not. In order to classify imagery into land cover types, we need to take a collection of classifiers, each with different strength and be able to complement each other to obtain an optimal classification result.

2. INDICES FOR DETERMINING THE NUMBER OF CLUSTERS

Within remotely sensed imagery, determined by their physical characteristics, similar land cover types normally form clusters in the data space. Given that each cluster has a large number of pixels, according to the central limit theorem, the distribution of each cluster should be approximately normal. Based on these assumptions, we adopted two indices, a skewness and a separation cohesion index (SCI) (Mardia, 1970 and Zhu and Zhu, 2007), to determined the proper number of clusters. A skewness measure is one of the indices to evaluate the overall clustering validity. Skewness is the 3rd central moment: E{[X-

E(X)]3}, and is a measure of the asymmetry of a density function about its mean. A Gaussian density's skewness is 0; a negative skewness indicates that the density is skewed left (with a long tail on the right in the univariate case); a positive value indicates the opposite. A measure of P-variate skewness is given in (Mardia, 1970) as

$$sk = \tfrac{1}{N^2}\sum_{n=1}^{N}\sum_{j=1}^{N}((X_n-\mu)^T\Sigma^{-1}(X_j-\mu))^3 \qquad (1)$$

In practice, the covariance matrix may be very close to be singular in certain cases due. To avoid this problem, we use a weighted average of the absolute sknewness of each variable in each cluster, where the size of each cluster is used as its weight of *P* image bands:

$$\overline{sk} = \frac{1}{N}\sum_{k=1}^{K}\left(\frac{N_k}{P}\sum_{p=1}^{P}\left|sk_{k,p}\right|\right) \qquad (2)$$

where *sk*$_{k,p}$ is the skewness of the *p*th variable of the *k*th cluster

$$sk_{k,p} = \frac{\sum_{n=1}^{N_k}(X_{n,p}-\mu_{k,p})^3}{N_k\sigma_{k,p}^3} \qquad (3)$$

Here, $\sigma_{k,p}$ is the standard deviation of the *p*th variable of the *k*th cluster.

The second index to be used is separation-cohesion index (SCI). For each cluster, we calculate the distance between its centroid and the centroids of the other clusters. The ratio of the smallest distance (a measure of separation) and the cluster's standard deviation (a measure of cohesion) can be computed for each cluster. SCI is the weighted average of the ratio (Zhu and Zhu ,2007):

$$SCI = \frac{1}{N}\sum_{k=1}^{K}N_k \min_{\substack{j=1,\dots,K\\ j\neq k}}\left|\mu_k-\mu_j\right|/\sigma_k \qquad (4)$$

where

$$\sigma_k = \frac{1}{N_k}\sum_{n=1}^{N_k}(X_n-\mu_k)^T(X_n-\mu_k) \qquad (5)$$

We can plot $\overline{sk}$ and SCI against the number of clusters. By visually identifying the kinks and slope changes in the curves, together with context of an application, an optimal and practical number of clusters in the data can be determined.

3. USING A COLLECTION OF CLASSIFIERS FOR OBTAINING OPTIMAL RESULTS

Although there are many tools that can be used for image classification, no one classifier outperforms all others in all occasions. A consensus is that each classifier has its strength and may perform better within a given application, using a data set from a certain sensor at a specified geographic region. In this study, we select three widely used classifiers: the maximum likelihood, the Kohonen's self-organizing map (SOM) neural network, and the classification tree method. The IDRISI software package, which implements these classifiers, was used in this study. In addition to the three supervised classifiers, the k-means clustering method, an unsupervised approach, was used to assist with the determination of the number of clusters. A brief review of the three classifiers indicated the strengths of the maximum likelihood method was for parametric classification, the SOM's approach was in handling multiple models, and classification tree method was efficient in handling data with no distribution patterns. Detailed descriptions of the algorithms can be found from Bishop, 2006; Tso and Mather, 2001; and Richard and Jia, 2006.

3.1 Maximum likelihood classifier

The maximum likelihood classifier is one of the widely used classifiers. It uses Bayesian probability theorem and multivariate density (usually Gaussian) to evaluate the posterior probability of a multidimensional data point *x* belongs to a class *k* in the follow fashion:

$$p(c_k \mid x) = p(c_k)\,p(X \mid c_k)\left(\sum_{i=1}^{n}p(x \mid c_k)\cdot p(c_k)\right)^{-1} \qquad (6)$$

where *p(c$_k$|x)* is the posterior probability of pixel x being class c$_k$, *p(c$_k$)* is the prior probability of *x* being class *c$_k$*, and *p(x |c$_k$)* is the conditional probability of *x* given class c$_k$.

By assuming an equal prior probability for all classes and using the multivariate Gaussian density function to approximate the frequency distribution associated with each of the classes parameterized by its mean μ_k and covariance Σ_k (Swain and Davis, 1978), the class *k* which has the maximum value of:

$$p_k(X_n \mid \mu_k,\Sigma_k) = (2\pi)^{-\frac{P}{2}}(\Sigma_k)^{-\frac{1}{2}} \qquad (7)$$
$$\exp(-\tfrac{1}{2}(X_n-\mu_k)^T\Sigma_k^{-1}(X_n-\mu_k))$$

is assigned to that pixel.

3.2 Kohonen's self-organizing map (SOM)

The SOM neural network contains two layers, the input layer and the output layer. Its input layer represents the input feature vector and thus has neurons for each image band. The output layer is normally organized as a two-dimensional squared array of neurons. Each output layer neuron is connected to all neurons in the input layer by synaptic weights, and the weights are initialized with random weights. The organization procedure uses progressive adjustment of weights based on data characteristics and lateral interaction such that neurons with similar weights will tend to spatially cluster in the output neuron layer (Kohonen, 1990; Kohonen 2001; Li and Eastman, 2006). Specifically, let $X = \{x_1, x_2, \ldots, x_p\}$ be an p-dimensional feature vector input to the SOM, the Euclidean distances between an output layer neuron and an input feature vector can then be calculated, and the neuron in the output layer with the minimum distance to the input feature vector (known as the *winner*) is then determined as:

$$Winner = arg \min_j \left(\sqrt{\sum_{i=1}^{n} (x_i^t - w_{ji}^t)^2} \right) \qquad (8)$$

where x_i^t is the input to neuron i at iteration t, and w_{ji}^t is the synaptic weight from input neuron i to output neuron j at iteration t. The weights of the winning neuron and its neighbors within a radius γ are then altered (while those outside are left unaltered) according to a learning rate α^t as follows:

$$\begin{aligned} w_{ji}^{t+1} &= w_{ji}^t + \alpha^t \cdot (x_i^t - w_{ji}^t), \quad \forall d_{winner,j} \in \gamma^t \\ w_{ji}^{t+1} &= w_{ji}^t, \qquad \forall d_{winner,j} \notin \gamma^t \end{aligned} \qquad (9)$$

where α^t is the learning rate at iteration t and $d_{winner,j}$ is the distance between the winner and other neurons in the output layer. The learning rate is a value between 0 and 1 that decreases with time between its maximum and minimum values according to the following time-decay function:

$$\alpha^t = \alpha_{\max} \left(\frac{\alpha_{\min}}{\alpha_{\max}} \right)^{\frac{t}{t_{\max}}} \qquad (10)$$

An identical time-decay function is also used to reduce the radius (γ) from an initial size that can encompass all of the neurons in a single neighborhood to an ending one which includes only the winner neuron. This adjustment of weights thus proceeds from global order to local adjustments. As for a supervised classification, a majority voting technique is used to associate these neurons in the output layer with training data classes.

3.3 Classification tree

A classification tree approach successively splits the data to form increasingly homogenous subsets and resulting in a hierarchical tree of decision rules. In this study, a univariate binary tree structure is adopted. It is a supervised procedure, containing training and classification steps. For the training step, an entropy rule is chosen to guide the growth of the tree. In general, the rules attempt to locate a splitting point in one of the multiple input images in order to isolate the largest homogenous training samples from the remainder of the training samples. An entropy rule uses the entropy (Shannon, 1948) as a measure in identifying an optimal band from the input images and locating the best splitting point in that band for splitting a node:

$$Entropy = -\sum_{k=1}^{K} \left(\frac{N_k}{N} \bullet \log_2 \left(\frac{N_k}{N} \right) \right) \qquad (11)$$

where N_k is the number of pixels for cluster k in a group, and N is the total number of pixels in a group. The split should yield minimum entropy. Iteratively, the classification tree is grown by progressively splitting a node into two new nodes. A newly grown node may turn into a leaf when it only contains training pixels from one training class and that would stop further splitting. The second step is to classify the images, in which every pixel is classified into a class utilizing the hierarchical rules of the decision tree. Since it does not assume any distribution pattern in the training samples, it can deal with non-cohesive spectral characteristics of land covers.

4. CASE STUDY AND DISCUSSION

Case studies were performed to validate the methodologies for identifying the number of clusters and obtaining optimal classification results using three data sets.

4.1 Data sets

Three data sets were utilized: ASTER (advanced spaceborne thermal emission and reflection radiometer), SPOT, and simulated SPOT images. They were all from the region of Worcester Country, located in central Massachusetts of the United States. The first data set contained three bands of ASTER L1B data (received on June 24, 2006), corresponding to green, red, and near infrared wavelengths with a 15-meter spatial resolution, covering approximately 236 km2. The second data set was SPOT satellite imagery (received on July 10, 2003), containing four bands in blue, green, red, and near infrared wavelengths. The image contained 513*513 pixels with 20 meter resolution, covering some 105 km2. The third dataset was derived from the second one through a simulation process (see appendix for details), and it had four bands with known normal distribution for seven spectrally separable land cover types. A composite for each was given in Figure 1 (a), (b), and (c).

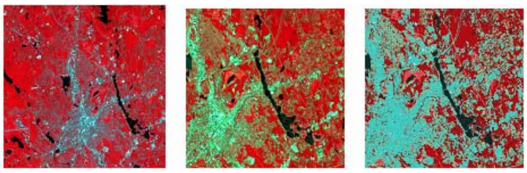

(a) (b) (c)

Figure 1. False color composites of the three data sets. (a)-ASTER, (b)-SPOT, and (c)-SPOT simulation.

The dominant land cover types of the study area were deciduous and coniferous forests, commercial and industrial build ups, low density residential areas, water bodies, some agricultural and recreational land uses. Field visits and Google Earth software were used to visually identify actual land cover types of clustered images. In the preparation of the SPOT simulation images, seven primary land cover classes were identified: *water body, coniferous forest, deciduous forest, concrete, asphalt, bare soil/rock,* and *grass*. In that process, more attention had been given to the land vegetation cover, while manmade objects and build up areas were grouped into *asphalt* and *concrete* categories, even with the awareness that *asphalt* land cover type had a wide data range. Each class was generated using a Gaussian model estimated from samples of visually identified classes, thus each pixel's class was known.

4.2 Identifying the number of clusters

The following three procedures were taken to identify the number of distinct clusters that exist in the three data sets. Among the three data sets, the SPOT simulation one was intended as a testing set since it had known number of clusters (seven) with known normal distribution.

4.2.1 Partitioning the imagery into clusters:

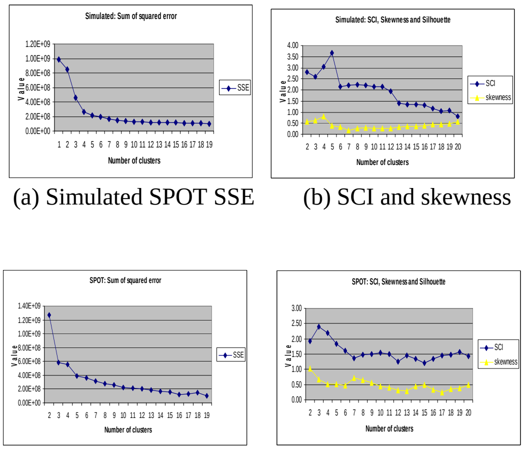
K-means unsupervised classification method was utilized to partition the imagery into 20 clusters using a random initial seeding option. Within the clustering process in k-means, its objective function is to minimize the total sum of squared errors to reach an optimal partition of the imagery (Bishop, 2006). After the k-means clustering process was completed with 20 clusters labeled, then iteratively, two closest clusters were merged and a new clustering image was derived using the updated cluster centroids. And the new clustered image contained one less than the number of clusters in the previously clustered image. Progressively, a series of clustered imaged were obtained. We stopped the cluster reduction process when the number of clusters was reduced to 2.

4.2.2 Deriving and plotting indices:

The sum of squared errors (SSE) to the centroids, the skewness, and the SCI indices against the number of clusters for the series of clustered images were derived and plotted in Figure 2.

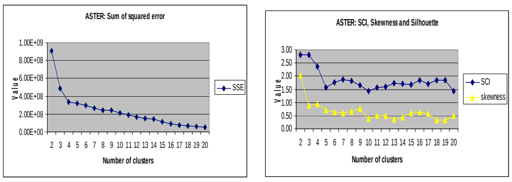
4.2.3 Identifying the optimal number of clusters in the images by interpreting the plots:

We first looked at the plots for the SPOT simulation data set, which contained seven known clusters in Figures 2 (a) and (b). The skewness index, consistent with the truth, yielded the smallest value at cluster number seven. When we examined the SSE and SCI indices, we found that both showed support at number seven: SSE had gone through a significant drop, and SCI was gradually decreasing before it dropped again after the cluster number got too large. We were aware that the SSE should monotonically decrease and the SCI did not have to be so due to the nature of k-means clustering algorithm and the algorithm for estimation of SCI.



(a) Simulated SPOT SSE (b) SCI and skewness

(c) SPOT SSE (d) SCI and skewness



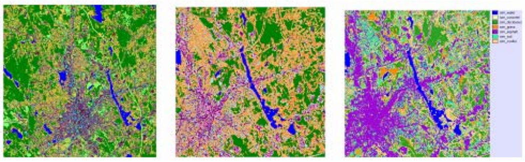
(e) ASTER SSE (f) SCI and skewness

Figure 2. Plot the sum of squared errors (SSE), sknewness, and SCI indices against the number of clusters for SPOT simulation, SPOT, and ASTER data sets.

As for Figures 2 (c) and (d), they were both derived from the real SPOT images. The SSE curve had experienced a noticeable drop and the SCI had a pit at seven clusters, although its skewness went up a little from six. Due to the complexity of a real data set, we may not be able to see support from all indices. In this case, at number six the skewness had reached a local minimum and SCI did so at number seven. Therefore, we could reasonably determine number seven as the optimal number of clusters presented in the imagery.

Finally, Figures 2 (e) and (f) indicate that there are likely 10 clusters. It was because at number ten, both SCI and skewness yielded local minima, and SSE had drop significantly after number four. From another point of view, domain knowledge suggested that a higher than SPOT spatial resolution (15 m vs. 20 m) should allow us to recognize more land cover types than that from SPOT, assuming that mixed pixels had similar effects.

With the identified number of clusters, the clustering images were then identified from the series. They were shown in Figure 3, of which, only the seven clusters in the SPOT simulation data were known.



(a) (b) (c)

Figure 3. Clustered images from ASTER, SPOT, and simulated SPOT images: (a) ASTER data with 10 clusters, (b) SPOT data with 7 clusters, and (c) SPOT simulation data with 7 known clusters.

4.3 Obtaining optimal classification result

In this investigation of achieving optimal classification result, the SPOT simulation image was only used to verify the parametric classifier, i.e. the maximum likelihood method. A test had been given and proved it worked as expected with all of the classes classified correctly. The rest of the study was focusing on using the real SPOT image data, and the ASTER data was only used for verification of the methodology.

4.3.1 Labeling clusters with land cover types:

After the number of clusters had been determined in the real SPOT data set, the clustered image was then examined and seven clusters could be labeled as *forest stands (including deciduous and coniferous), water bodies, croplands plus asphalt pavement, bare soil plus some asphalt, wetlands together with some asphalt, concrete surfaces, and grey asphalt* *surfaces.* We realized that *asphalt* had a very wide data range, from new pavement with very dark color, to old surfaces with very light color. That had caused misclassifications with *wetlands, bare soil, and harvested croplands* in the clustered image. Actually it constituted a generic case for testing the methodologies for achieving optimal classification results.

4.3.2 The preparation of the training and testing data sets:

The clustering result served as a guide for developing ground truth data, including training and testing data sets. Since any scenario may occur in real world application and without losing generality, the clustering result was taken as the ‘truth’ to develop training and testing data sets. Credit thus had been given to the theoretical base of identifying the number of clusters as explained in the previous section. And the mislabeled clusters that needed further merging and splitting processes, would be dealt with in the next iteration. In this study, we took about 28% of pixels randomly for the training, and the rest some 72% as testing data set.

4.3.3 Obtaining the optimal classification result:

The three classifiers, the maximum likelihood, the SOM, and the classification tree methods were all applied on the real SPOT images. An analysis started from analyzing the three error matrices for each classifier respectively (Congalton and Green, 1999). As it was implied previously, the analysis was not intended to provide evidence of which particular classifier performed better. It was because the creation of training and testing data in this case or any real cases may favor a particular classifier (the parametric maximum likelihood classifier in this case). We would rather focus on the accuracy of each class and to accept the classification result from any one of the three classifiers that better met a given accuracy threshold. Due to the size of the error matrix and the limitation of the paper length, the error matrix could not be inserted. According to those error matrices, classes for *water body* and *concrete* were accepted, from maximum likelihood and classification tree methods respectively. They had showed above 99% of consistency with the testing data set. In the first iteration, we could not obtain results for any other classes with acceptable classification accuracy. In the next iteration, the classes that did not meet a given accuracy were to be further analyzed. In this case, for example, the *forest* land cover needed to be split into *deciduous* and *conifers*; three different *asphalt* classes needed to be separated from their misclassification with *croplands, bare soil* and *wetlands.*

In the second round analysis, the already identified classes with acceptable accuracy (*water body* and *concrete*) were masked out from analysis, and only the remaining ones were analyzed. Each class was processed separately. That had made identifying the number of clusters straightforward. Most of them contained only two or three misclassified classes, such as *asphalt* from *wetlands*; or one correctly identified class needed to be further split into finer classes, such as splitting *forest* into *deciduous* and *conifers*. The second iteration completed all of classes.

The ASTER data set was used to validate the proposed methodology. There were two differences though, one was that the *deciduous* and *coniferous* forests were identified in the cluster identification phase, and they got classifier correctly in the first iteration. The other was that the *asphalt* land cover type was identified as several different land cover types. They could be aggregated into one single *asphalt pavement* class in a post processing, depending on the interest of an application.

We did not indicate which classifier obtained optimal classes at which iteration. Since their strengths were complementing each others weakness at any scenario, an optimal classification result was always available from one of the three classifiers for a given class within an iterative process.

5. CONCLUSION

We have presented procedures for identifying the number of clusters exist in remote sensing images and practical approaches leading to obtaining optimal classification result. The practical methodologies for supervised classification avoided the normally time consuming practice of editing training data to meet classification accuracy requirement and dilemmas of picking a classifier to accommodate all classes in a imagery data set. Future work is planned to investigate the effect of resolution change to the number of distinct clusters and responses of their statistical indices.

6. APPENDIX

The following procedure was used to generate the SPOT simulation imagery using the original SPOT imagery: 1) In the original imagery, seven clusters were identified and training samples were developed. They were *water body, conifer forest, deciduous forest, concrete, asphalt, bare soil/rock,* and *grass*. The samples were used to estimate the mean vector and variance and covariance matrix of each class. These parameters were treated as Gaussian density parameters to generate data later. 2) A supervised classification with seven classes of the SPOT image was undertaken. From this point onwards the class of a pixel was known and was fixed. 3) A principal components analysis (PCA) was applied to the original imagery, yielding four components. The Gaussian parameters estimated earlier were converted into the PCA data space. The 7 Gaussian models in the PCA space were used to regenerate the pixels to replace the original ones; each pixel’s class was preserved. Then, the pixels in the PCA space were converted back to the original data space. This procedure added white noise to the generated data and preserved the correlations between variables within each class.

7. REFERENCE

Bishop C. M.,2006. *Pattern Recognition and Machine Learning,* Springer.

Congalton, R. G. and K. Green,1999. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices¸ CRC/*Lewis Press.

Fraley, C. and A. E. Raftery,1998. "How Many Clusters? Which Clustering Method? Answers via Model-Based Cluster Analysis," *The Computer Journal*, vol. 41, pp. 578-588.

Kohonen, T.,1990. The Self-Organizing Map. *Proceedings of the IEEE,* 78: 1464-80.
Kohonen, T.,2001. *Self-Organizing Maps,* Third Edition. New York, Springer.

Li Z., J. R. Eastman,2006. The Nature and Classification of Unlabelled Neurons in the Use of Kohonen's Self-Organizing

Map for Supervised Classification, *Transactions in GIS* 10 (4), 599–613.

Mardia K. V.,1970. "Measures of Multivariate Skewness and Kurtosis with Applications," *Biometrika*, vol. 57, pp. 519-530.

Richards, J. A. and X. Jia, 2006. *Remote Sensing Digital Image Analysis, An Introduction,* Springer, Fourth Edition.

Shannon, C.E.,1948. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27: 379-423, 623–656.

Swain P. H. and S. M. Davis,1978. *Remote Sensing: The Quantitative Approach,* McGray-Hill.

Tso B., and P. M. Mather,2001. *Classification Methods for Remotely Sensed Data,* Taylor & Frances.

Zambon, M., R. Lawrence, A. Bunn, and S. Powell,2006. Effect of alternative splitting rules on image processing using classification tree analysis. *PERS,* 72(1): 25–30.

Zhu H. and H. Zhu,2007. Clustering Analysis using Data Range Aware Seeding and Agglomerative Expectation Maximization, *Proceedings of the Third International Conference on Signal-Image Technology & Internet-Based system (SITIS'2007),* Shanghai.